

**TRABAJO DESARROLLADO EN EL CENTRO DE  
NEUROCIENCIAS DE CUBA**

INSTITUCION AUTORIZADA: UNIVERSIDAD DE CIENCIAS  
MÉDICAS DE LA HABANA

**TITULO: CAUSALIDAD DE GRANGER SOBRE VARIEDADES  
ESPACIALES: APLICACIONES A LAS NEUROIMAGENES**

(Conectividad funcional en el cerebro humano evaluada a partir de imágenes cerebrales multifuncionales desde la perspectiva de la causalidad de Granger)

TRABAJO PARA OPTAR POR EL GRADO DE DOCTOR EN  
CIENCIAS

OPTANTE: Dr. C PEDRO ANTONIO VALDES SOSA

LA HABANA, CUBA

2011

## AGRADECIMIENTOS

Esta tesis es parte de la historia del Centro de Neurociencias desde su inyección. Honro con ella a nuestros maestros:

- Bjorn Holmgren y Ruth Urbá, neurofisiólogos y marxistas chilenos, guías de innumerables generaciones de Neurocientíficos cubanos. Ellos vinieron a Cuba por solicitud del senador Salvador Allende,
- Thalia Harmony, mi maestra, quien me enseñó no solo el amor a la ciencia sino también a la integridad política en bien de los pobres a través de una lucha de clases que nunca cesa.
- Erwin Roy John, voluntario menor de edad en la guerra contra los nazis, izquierdista americano, colaborador de la primer computadora cubana, quien me enseñó que los espacios multidimensionales se podían poner en bien de la salud mental humana.

Son parte de esa historia también mis compañeros, algunos presentes, otros desaparecidos, otros que han abandonado la lucha, todos que contribuyeron a lo que concibo como la ciencia cubana. No como una empresa individual para bien personal, sino una lucha perenne contra el desconocimiento, las enfermedades cerebrales, el encono del imperio y la indiferencia burocrática. Agradezco también a los líderes encabezados por el entrañable compañero Fidel, que nos enseñaron que la ciencia es el futuro de Cuba.

## DEDICATORIA

A mis padres Pedro y Eva, quienes me motivaron desde niño a ser comunista y científico. Nuestro hogar en Chicago como lugar de reunión del Movimiento 26 de Julio, y el consecuente regreso a la patria en 1961, me permitió ser parte para siempre de la epopeya revolucionaria cubana.

A mi hermano Mitchell, por descuidar títulos y honores para consagrarse a transformar la realidad en un contrapunteo dialectico desde que nacimos.

A mi esposa María Luisa, quien además de ser mi colaboradora científica y política más cercana, me ha inundado la vida con felicidad.

## SINTESIS

Se presenta el Capitulo “Granger Causality on Spatial Manifolds. Applications to neuroimaging” del “Handbook of Time Series Analysis” ( Wiley-VCH.2006) para optar por el grado de Doctor en Ciencias. Resume los aportes realizados por el autor en la determinación de la conectividad efectiva y funcional de estructuras cerebrales humanas, evaluada a partir de Neuroimágenes multimodales: el electroencefalograma (EEG), resonancia magnética funcional (fMRI), registros combinados de ambos (EEG/fMRI). Se generaliza el concepto de causalidad de Granger, anteriormente descrito solo para sistemas discretos, para sistemas definidos sobre variedades continuas—así posibilitando la modelación de la corteza cerebral. Ha despertado interés porque permite identificar in vivo los circuitos que se activan en distintos estados cerebrales. El capítulo viene acompañado de los 6 artículos, refrendados en la “Web of Science” con 259 citas, de las cuales es la consolidación y generalización. Uno de los artículos fue publicado en un número especial de la “Philosophical Transactions of the Royal Society”, editado por el propio optante. Se incluye además un artículo recientemente solicitado por la revista NeuroImage como síntesis de la polémica generada por el trabajo. Como se describe en detalle en las conclusiones generales el problema de la conectividad cerebral es clave en el conocimiento del cerebro normal y enfermo y se sustenta en los avances más recientes del estudio estadístico de las relaciones causales y la biofísica cerebral—al cual ha contribuido en forma importante el grupo que dirige el autor.

## SIGLAS

En este trabajo se usan las siglas de los términos en inglés para facilitar la lectura de los artículos que están en ese idioma. Los equivalentes en español están en la tercera columna.

<b>SIGLAS</b>	<b>INGLES</b>	<b>ESPAÑOL</b>
fMRI	Functional Magnetic Resonance Imaging	Resonancia magnética funcional
ELEG	Electroencephalography	Electroencefalografía
DCM	Dynamic Causal Modeling	Modelación Dinámica Causal
ICA	Independent Component Analysis	Análisis de componentes independientes
SPM	Statistical Parametric Mapping	Mapeo Paramétrico Estadístico
MEG	Magnetoencephalography	Magneto encefalografía
MAR	Multivariate Autoregressive Model	Modelo Autoregresivo Multivariado
BOLD	Blood Oxygen Level Dependent	Dependiente de Nivel de Oxígeno Sanguíneo
SMAR	Sparse Multivariate Autoregressive Model	Modelo Autoregresivo Multivariado "ralo"
FDR	False Discovery Rate	Tasa de Falsos Descubrimientos
FDA	Functional Data Analysis	Análisis de datos Funcionales

## Contents

INTRODUCCION	1
La Causalidad de Granger sobre Variedades Espaciales: Aplicaciones a las Neuroimágenes	13
Spatio-Temporal Autoregressive Models defined over brain manifolds	19
Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex	20
Decomposing EEG data into space-time-frequency components using parallel factor analysis	21
Introduction: multimodal neuroimaging of brain connectivity	22
Concurrent EEG/fMRI analysis by multiway Partial Least Squares	23
Estimating brain functional connectivity with sparse multivariate autoregression	24
Effective connectivity: influence, causality and biophysical modeling	25
Conclusiones	26
Recomendaciones	27

## INTRODUCCION

Titulo del trabajo:

### ***Casualidad de Granger sobre Variedades espaciales. Aplicaciones a las Neuroimagenes***

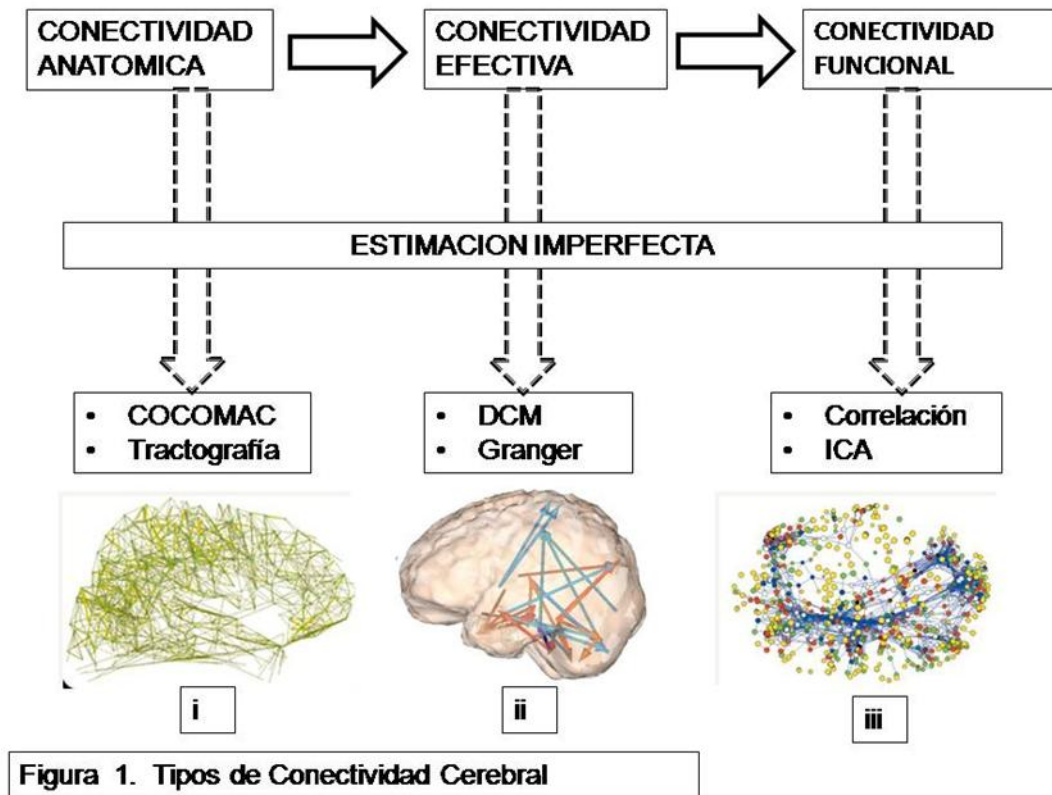
Autor: Pedro A. Valdes Sosa, Centro de Neurociencias de Cuba

El trabajo que se presenta par optar por el título de Dr. en Ciencias es el capitulo 18 del Libro **Handbook of time Series Analysis**, titulado "**Granger causality on spatial manifolds: appliactions to Neurimages**"; publicado por la Wiley-VCH en el año 2006.

Importancia del Tema de Conectividad Cerebral

Hasta hace poco la investigación con Neuroimagenes se centraba sobre la **Localización anatómica** de cambios estructurales y funcionales entre condiciones distintas experimentales o grupos. Sin embargo, fue tempranamente reconocido (solo basta recordar los trabajos de Luria) que muchas funciones cerebrales no son atributos de una estructura neural particular, sino que emergen de la integración de la actividad de masas neurales ampliamente distribuidas. Por tal motivo, resulta de esencial importancia el estudio de la **Conectividad Cerebral**. Debemos distinguir tres tipos fundamentales de **Conectividad cerebral: anatómica, efectiva y funcional**.

En la Fig. 1 se muestra la relación entre los tres tipos de conectividad.



La conectividad anatómica entre dos regiones cerebrales (A y B) se refiere a la existencia de proyecciones directas de axones de una a la otra. Cuando alguna de estas conexiones es activada en una función cerebral determinada se habla de conectividad efectiva, lo cual implica que la activación de la estructura A es causa directa de la activación de la estructura B. Finalmente, se habla de conectividad funcional cuando la actividad en una estructura A está correlacionada con la de otra estructura B, o sea, por mediación directa o a través de una tercera estructura C.

Un ejemplo sencillo que ilustra estas definiciones es el siguiente: existe conectividad anatómica entre la retina y el núcleo geniculado lateral y entre el núcleo geniculado lateral y la corteza estriada (visual).



Por tanto, si se establece que la activación de ciertas partes del geniculado lateral produce activación de alguna parte de la corteza estriada, estamos en presencia de conectividad efectiva. Sin embargo, establecer que la activación de la retina produce una activación de la corteza estriada, solo demuestra que hay una conectividad funcional entre ambas, ya que es mediada por el geniculado lateral.

En la práctica experimental la situación es más compleja, pues ninguno de estos tipos de conectividad se puede medir directamente, sino que son estimados de forma imperfecta, Por tanto, la estimación de cualquier tipo de conectividad, es de hecho, la solución a un Problema Inverso.

Pongamos algunos ejemplos:

- Un método que estima de forma imperfecta la **conectividad anatómica** es la Tractografía por MRI de difusión, en la porción inferior de la figura 1.i se muestra un grafo de las conexiones anatómicas estimadas para un sujeto. En este grafo los nodos son regiones corticales circunscritas y se dibuja un enlace entre dos nodos, solo si se ha determinado que existe un tracto que conecta esas dos regiones. El grafo no es dirigido ya que la Tractografía no puede determinar la direcciones de las proyecciones axonales. A priori, se conoce que habrá una proporción de enlaces estimados falsos positivos y negativos.
- Otro método para la estimación de la **conectividad anatómica** es el de realizar lesiones circunscritas en el cerebro de animales experimentales y estudiar la degeneración axonal resultante.

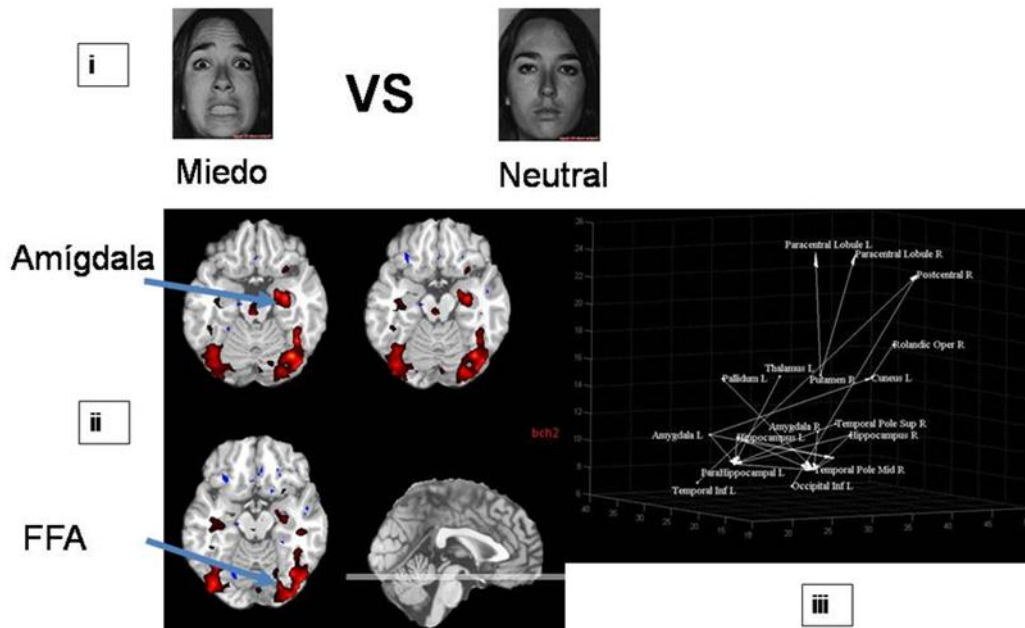
De hecho existe, para el cerebro de la especie Macaca Mulatta, una base de datos CoCoMac (Collations of Conectivity data on the Macaque Brain), con las conexiones deducidas mediante lesiones. Las mismas se clasifican en demostradas, probables y desconocidas, lo cual refleja las imperfecciones de la técnica (URL: <http://cocomac.org/home.asp>)

- Los métodos para la estimación de la conectividad funcional también producen un grafo no dirigido, donde los enlaces ahora se postulan cuando hay asociación entre las series temporales de los voxels de las Neuroimagenes funcionales tales como el fMRI o el EEG. Estas asociaciones se determinan con métodos que varían desde la correlación simple de las series de tiempo, hasta el uso de métodos mas sofisticados como el análisis de los componentes independientes ICA (Independent Component Analysis). En la figura 1.ii se muestra un grafo formado con las correlaciones de la actividad de fMRI.

- Para el estudio de la integración funcional cerebral resulta óptimo determinar la **Conectividad Efectiva**. Para su estimación existen dos enfoques complementarios que han suscitado mucho interés recientemente. Uno de ellos es el DCM (Dynamic Causal Modeling), basado en modelos biofísicos de la interacción entre zonas corticales. El otro es un método importado de la econometría, la causalidad de Granger, que mide la causalidad a través de la posibilidad de predecir una serie de tiempo por otra. Ambas pretenden producir grafos con enlaces dirigidos que postulan una relación causal entre las regiones cerebrales involucradas. Anexo G.

En forma similar a como el Mapeo Paramétrico Estadístico (SPM) se emplea para determinar la localización de cambios estructurales o funcionales, pueden también aplicarse los métodos de SPM a las conexiones donde son mostrados aquellos enlaces que sobrepasen un umbral estadístico.

Como ejemplo de aplicación de estos conceptos se muestra, en la Figura 2.ii, el grafo de conectividad efectiva correspondiente al experimento ya descrito de caras con expresiones emocionales y neutras. Este grafo muestra los enlaces que entre regiones cerebrales se activan de forma distinta para las dos condiciones experimentales. Anexo A.



**Figura 2 Conectividad efectiva del procesamiento de expresiones emocionales**

Aportes originales enumerados:

El capítulo que se presenta para optar por el Grado de Doctor en Ciencias, es la consolidación y generalización de 6 artículos científicos referenciados en la “Web of Science” y que se listan en la tabla I.

1. Introducción, por primera vez en la literatura, de un tratamiento de la causalidad de Granger (aporte matemático) sobre conjuntos de variables extendidas espacialmente.
2. Introducción, por primera vez en la literatura, del uso de método de regresión penalizada para la estimación de los coeficientes de un modelo auto-regresivo multivariado.
3. Integración de este concepto a las Neuroimágenes funcionales como un método de mapeo paramétrico estadístico (SPM) sobre conectividades y no sobre activaciones.
4. Demostración de la utilidad de estos métodos para el estudio del funcionamiento cerebral espontáneo y durante tareas cognitivas utilizando tanto la resonancia magnética funcional (fMRI) como el registro concurrente de fMRI y electroencefalograma (EEG).

Tabla I

	Articulo	Anexos	Factor de impacto de la revista	Total de citas
1	Valdes-Sosa PA, Sanchez-Bornot M, Lage-Castellanos A, Vega-Hernandez M, Bosch-Bayard , Melie-Garcia L, Canales-Rodriguez E, 2005. Estimating brain functional connectivity with sparse multivariate autoregression. Philosophical Transactions of the Royal Society B-Biological Sciences 360: 969-981	A	4.99	44
2	Miwakeichi F, Martinez-Montes E, Valdes-Sosa PA, Nishiyama N, Mizuhara H, Yamaguchia Y, 2004. Decomposing EEG data into space-time-frequency components using Parallel Factor Analysis. Neuroimage 22: 1035-1045.	B	4.86	76
3	Martinez-Montes E, Valdes-Sosa PA, Miwakeichi F, Goldman RI, Cohen MS, 2004. Concurrent EEG/fMRI analysis by multiway Partial Least Squares. Neuroimage 22: 1023-1034.	C	4.86	74
4	Valdes-Sosa PA, 2004. Spatio-temporal autoregressive models defined over brain manifolds. Neuroinformatics 2: 239-250	D	3	15
5	Freiwald WA, Valdes P, Bosch , Biscay R, imenez C, Rodriguez LM, Rodriguez V, Kreiter AK, Singer W, 1999. Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. Journal of Neuroscience Methods 94: 105-119.	E	1.36	46
6	Valdes-Sosa P, 1997. Quantitative electroencephalographic tomography. Electroencephalography and Clinical Neurophysiology 103: 19.	F	2.4	4

## Análisis integral del grupo de aportes

En estos momentos resulta de crucial importancia desarrollar métodos para la determinación in vivo de las redes neurales que subyacen el funcionamiento normal cerebral y puedan describir sus alteraciones en patologías neuropsiquiátricas.

Una línea de trabajo de la estadística contemporánea ha sido la determinación de relaciones de causalidad a partir de series cronológicas basado en la llamada "Causalidad de Granger". Sin embargo, el impulso inicial a este tipo de trabajo lo fue la econometría (trabajo por el cual obtuvo el premio Nobel en el año 2003 C. W. Granger).

Desde muy temprano nuestro grupo aplicó estos métodos a registros de la actividad eléctrica cerebral. Sin embargo, había dos escollos fundamentales a vencer para su real aplicación al estudio del cerebro:

1. Los métodos desarrollados hasta el momento solo permitían el análisis de un número muy limitado de series cronológicas, de 2 a 5 a lo sumo; lo cual hacía muy dudoso su uso en el estudio de las Neuroimágenes, en donde son frecuentes decenas o centenas de miles de series cronológicas, una por cada voxel medido.
2. No se tomaban en cuenta las relaciones espaciales existentes entre las series cronológicas, que en el caso del cerebro, se registran sobre una variedad que es un conjunto de contornos suaves sobre el cual se puede establecer un sistema de coordenadas geográficas.

El trabajo presentado resolvió los dos problemas mediante el uso de la regresión bayesiana en su formulación de regresión multivariada penalizada. A partir de este trabajo, en la literatura internacional, se ha iniciado una línea de publicaciones en los dos campos afines de neuroinformática y bioinformática.

Valoración del impacto de estos aportes:

#### Científico

El cuerpo de trabajo presentado ha logrado un impacto que puede medirse objetivamente por un total de 259 citas. El índice h de este trabajo es de 5 (eliminando las autocitas). Ha sido objeto de 4 conferencias invitadas en los Talleres de Conectividad Cerebral. Se creó una revista especializada de este tema del cual es editor el optante. Ha sido objeto de dos conferencias invitadas de los congresos mundiales de Mapeo Cerebral Humano (4,000 participantes).

Asimismo originó la invitación a editar un número especial de la revista “Philosophical Transactions of the Royal Society” en el cual apareció uno de los artículos de la tabla J

Debe mencionarse que ha surgido una metodología aparentemente distinta para medir la conectividad funcional: la Modelación Causal Dinámica de Karl Friston, miembro de la Royal Society y una de las figuras más prestigiosas del Mapeo Cerebral Humano.

El artículo 1 de la Tabla II describe una comparación experimental de la metodología de Granger y del DCM. Ello motivó un comentario muy crítico de métodos de Granger en el artículo 2 de la Tabla II, escrito por Friston.

A sugerencia del optante, la revista Neuroimage publicó los artículos 3-9 de la tabla II como parte de un “Comments and Commentary” que desde el año 2009 han generado 121 citas para un índice h de 6. Esta serie polémica se resume con un artículo conjunto del optante (como primer autor) y de Friston que acaba de ser publicado en que se sintetizan ambos enfoques. Anexo G.

Por el interés despertado por esta polémica, una propuesta del optante fue seleccionada como uno de los tres simposios centrales del congreso mundial de mapeo cerebral del 2011. El comité organizador informó que esta iniciativa logró el mayor puntaje de un concurso internacional de 37 iniciativas y al que asistieron 2000 oyentes.

Tabla II

	Articulo	Anexos	Total de citas
1	David, Olivier. (2009). fMRI connectivity, meaning and empiricism Comments on: Roebroeck et al. The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution. NeuroImage, 1-4. Elsevier Inc. doi: 10.1016/.neuroimage.2009.09.073	-	1
2	Roebroeck, Alard, Formisano, Elia, & Goebel, Rainer. (2009a). Reply to Friston and David fMRI : Model selection , causality and deconvolution. NeuroImage. Elsevier Inc. doi: 10.1016/.neuroimage.2009.10.077	-	4
3	Marinazzo, D., Liao, W., Chen, H., &Stramaglia, S. (2010). Nonlinear connectivity by Granger causality.NeuroImage. Elsevier Inc. doi: 10.1016/.neuroimage.2010.01.099	-	6
4	Friston, K. (2009b). Dynamic causal modeling and Granger causality Comments on: The identification of interacting networks	-	7



	in the brain using fMRI: Model selection, causality and deconvolution. NeuroImage, 2007-2009. Elsevier Inc. doi: 10.1016/.neuroimage.2009.09.031		
5	Daunizeau, ., David, O, & Stephan, K. E. (2009). Dynamic causal modelling: A critical review of the biophysical and statistical foundations. NeuroImage, 1-11. Elsevier Inc. doi: 10.1016/.neuroimage.2009.11.062	-	8
6	Bressler, S. L., & Seth, A. K. (2010). Wiener-Granger Causality: A well established methodology. NeuroImage, 7.doi: 10.1016/.neuroimage.2010.02.059	-	9
7	Roebroeck, Alard, Formisano, Elia, & Goebel, Rainer. (2009b). The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution. NeuroImage. Elsevier Inc. doi: 10.1016/.neuroimage	-	13
8	David, Olivier, Guillemain, I., Saitlet, S., Reyt, S., Deransart, C., Segebarth, C., et al. (2008). Identifying neural drivers with functional MRI: an electrophysiological validation. PLoS biology, 6(12), 2683-97. doi: 10.1371/ournal.pbio.0060315	-	34
9	Friston, K. (2009a). Causal modelling and brain connectivity in functional magnetic resonance imaging. PLoS biology, 7(2), e33. doi: 10.1371/ournal.pbio.1000033	-	39
10	Valdes-Sosa, P.A., Roebroeck A., Daunizeau J., Friston K., (2011) Effective connectivity; influence, causality and biophysical modelling, Neuroimage	G	

### Formación de Personal

El trabajo enumerado es parte de la línea de trabajo de Neuroinformática que se ejecuta en el Centro de Neurociencias de Cuba. Esta especialidad fue creada en Cuba por el optante, al cual se le reconoce como uno de los contribuyentes a su creación

La tesis tutoradas por el optante más directamente relacionados con el tema del trabajo son:

- Doctor en Ciencias Estadísticas y Postdoctorado de Fumikazu Miwakeichi (entonces en el Instituto de Estadística Matemática y el Instituto RIKEN de Japón).
- Doctores en Ciencias Físicas Eduardo Martínez Mont Lester Melie García.
- Dr. En Ciencias Matemáticas Jose Miguel Sánchez Bornot.
- Entrenamiento de los Investigadores Agustín Lage Castellanos y Mayrím Hernández Vega.

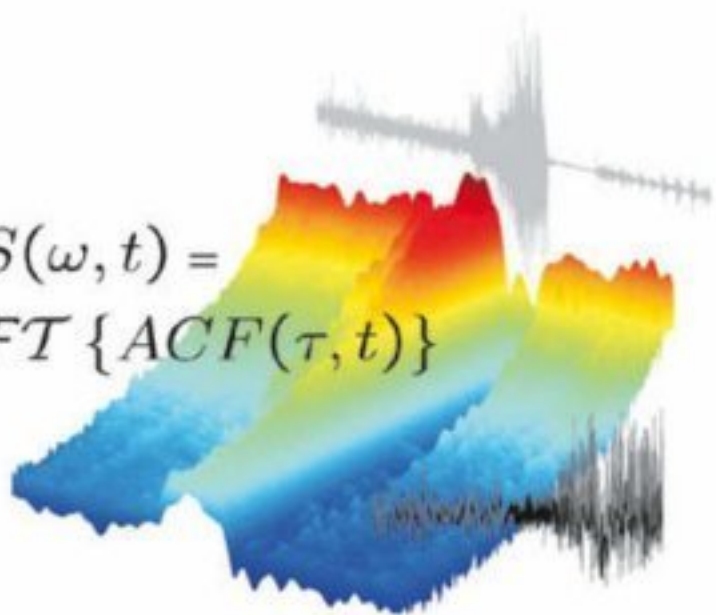
El trabajo ha recibido 2 premios de la Academia de Ciencias de Cuba, los cuales recibieron la distinción especial del Ministro del CITMA.

Edited by B. Schelter,  
M. Winterhalder and J. Timmer

WILEY-VCH

# Handbook of Time Series Analysis

Recent Theoretical Developments and Applications

$$S(\omega, t) = \mathcal{FT} \{ACF(\tau, t)\}$$


Traducción al Español de la Introducción de:

## **La Causalidad de Granger sobre Variedades Espaciales: Aplicaciones a las Neuroimágenes**

Pedro A. Valdés-Sosa, Jose Miguel Bornot-Sánchez, Mayrim Vega-Hernández

Lester Melie-García, Agustín Lage-Castellanos, and Erick Canales-Rodríguez

Departamento de Neuroinformática

Centro de Neurociencias de Cuba, Ciudad Habana, Cuba

**Palabras claves:** Causalidad espacial de Granger, EEG, fMRI, Conectividad cerebral efectiva.

La elaboración de métodos para inferir la conectividad efectiva y funcional de las diferentes regiones del cerebro es actualmente un tema importante para la esfera de las neuroimágenes [21]. La tarea es determinar los patrones cambiantes de las influencias causales que las diferentes estructuras neuronales ejercen entre sí. Esta tarea se debe llevar a cabo a través del análisis de datos de las imágenes dinámicas del cerebro. Este tipo de datos incluye la distribución de fuentes de EEG / MEG, los registros ópticos [65]\* y resonancia magnética funcional [36], que son, desde un punto de vista estadístico, conjuntos de datos espaciotemporales [48] [73]- es decir, series de tiempo mostrada de una variedad subyacente continua  $\Omega$  de puntos

espaciales. El uso de los modelos multivariados autoregresivos (en particular los lineales) para las series de tiempo vectoriales, ha demostrado ser una herramienta esencial e informativa para las ciencias aplicadas. Dentro de este

marco de trabajo, Granger [33] formuló una definición de la causalidad entre las series de tiempo, que ha sido aplicada extensamente en muchos campos, sobre todo en el de las neurociencias [3] [56]

Sin embargo, es importante señalar que el trabajo en este campo se ha limitado a *series* de tiempo vectoriales, en los que la dimensión  $p$  es muy pequeña [64] [7]-incluso si, como es habitual en las aplicaciones reales, el número  $N$  de muestras de tiempo recogidas es grande. Como señaló Granger, su definición de causalidad sería válida sólo si todas las variables relevantes se incluyeran en el análisis, una tarea muy compleja, lo cual es comprendido perfectamente ya que ellos estudian el cerebro, que es el sistema complejo por excelencia. Por ello, hemos centrado nuestra atención en los modelos autoregresivos multivariados (MAR, por sus siglas en inglés) definidos sobre variedades espaciales (un ejemplo particular de lo cual es el cerebro) y en como manejar, el problema de series de tiempo muestreadas densamente (de alta dimensión, altamente correlacionadas) que surgen a partir de la discretización en voxels de un continuo espacial subyacente [68].

Como un ejemplo concreto, el cual se utilizará en todo el trabajo, se considera la unión de las series de tiempo concurrentes del EEG y el fMRI para analizar

el origen de los ritmos cerebrales en reposo [31] [52] [55]. El paradigma de la adquisición se describe más detalladamente en la Sección 8. Patrones estructurados de las correlaciones se han encontrado entre los componentes espectrales variables en el tiempo en diferentes bandas de EEG y en la señal dependiente del nivel de oxígeno en sangre (BOLD, por sus siglas en inglés) localizadas en diferentes voxels. Estos patrones revelan la presencia de sistemas anatómicos de amplia distribución aparentemente involucrados en la generación de estas oscilaciones (véanse las figuras 1-5). Aquí  $N = 108$ , el número de series de tiempo del EEG es sólo 16, pero el número de series de tiempo del fMRI es 12.640! El modelo MAR frecuentemente utilizado no puede ser ajustado para esta cantidad de datos.

El enfoque utilizado en este trabajo sigue la estrategia de Análisis de Datos Funcionales [61]. Las cantidades de interés en los MAR espaciales (coeficientes autoregresivos) se estiman sujetos a limitaciones que tienen sentido anatómico y fisiológico. Estas, no sólo permiten la inferencia de datos de la densidad de la muestra, sino que también encajan muy bien con los métodos abreviados de cálculo que hacen factible los procedimientos propuestos. En los modelos MAR clásicos, la causalidad de Granger de un conjunto de series de tiempo sobre otro conjunto es calculada a través de medidas de influencia [24] [25]. En el caso lineal, estas medidas de influencia son por lo general tests multivariados de la hipótesis, donde ciertos coeficientes de regresión son iguales a cero. En el MAR espacial (SMAR) aplicado en nuestros experimentos extendemos este concepto al de un campo de influencia.

Para las Neuroimágenes funcionales, estos son mapas topográficos de la influencia de un sitio del cerebro (voxel) sobre el resto. Por ejemplo, en el experimento de EEG- fMRI realizado en paralelo que acabamos de mencionar, sería de interés conocer que influencia podría tener un sitio en la corteza visual (Figura 1) sobre el resto del cerebro.

En este tipo de situación los tests multivariados clásicos son difíciles de realizar o no funcionan. Por lo que se propone aplicar el enfoque univariado masivo que es el concepto del Mapeo Paramétrico Estadístico (SPM, Statistical Parametric Mapping) [74]. El SPM calcula en esencial un estadígrafo (uni o multivariado) en cada voxel de una imagen del cerebro y luego determina cuales son las regiones significativamente activadas por medio de procedimientos que controlan el error de tipo I. Esto último se logra ya sea a través del uso de la Teoría de los Campos Aleatorios [74], a través de los métodos de remuestreo [11], o del uso de la tasa de Descubrimiento Falso (FDR, por sus siglas en inglés) [13].

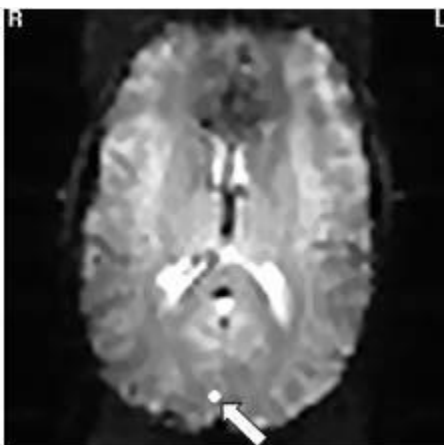


Figura 1: Imagen de resonancia magnética como un ejemplo de una superficie del cerebro. EPI- imagen de resonancia magnética del cerebro de un sujeto a partir de [30]. La sección de RM está a un nivel que traspasa la corteza visual primaria o estriada (VC). La flecha marca el voxel en la corteza visual primaria para los que la respuesta BOLD durante el ritmo alfa muestra la correlación más alta con el poder en esa banda.

Proponemos evaluar una extensión espacial de la causalidad de Granger a través de un SPM de los campos de influencia. En efecto, estamos interesados en la detección de regiones significativas en el conjunto producto cartesiano  $\Omega \times \Omega$ . En esta situación una alternativa al uso de técnicas multivariadas ordinarias de regresión es lograr la solución de un gigantesco problema de regresión multivariado e intentar las pruebas asociadas de los coeficientes de regresión. Para ser capaz de hacer esto, se deberá trabajar con la regresión sobre la base de la penalización en el espíritu del Análisis de Datos Funcionales (FDA, por sus siglas en inglés) [61]. Este enfoque reduce drásticamente el número de conexiones "efectivas" a determinar. Este fue el enfoque adoptado al [69] introducir una variante del Análisis de Datos Funcionales de los modelos MAR que impuso suavidad espacial sobre el campo de influencia. La reducción masiva de datos se logró por medio de la descomposición en valores singulares y este trabajo demostró la factibilidad de trabajar en la situación  $p > N$ . Un artículo posterior [70] también utiliza la regresión penalizada, en este caso se introducen los Modelos Multivariados



Autoregresivos Ralos (Sparse). Este último se puede estimar a través de un proceso de dos etapas que implican a) regresión penalizada y b) desechar las conexiones poco probables por medio de la tasa local de descubrimiento falso desarrollada por Efron. Se realizaron amplias simulaciones en redes corticales ideakizadas con una topología de mundo pequeño y una dinámica estable. Esto muestra que la eficiencia en la detección de conexiones del procedimiento propuesto es bastante alta. Por otra parte, la rareza o la independencia condicional no tiene que ser especificada a priori, sino que se descubre automáticamente mediante un proceso iterativo. En resumen, utilizamos el hecho de que el cerebro está conectado ralmente como parte de la solución, en vez de tratarlo como un problema de especificación. Este capítulo une los dos enfoques- el de la suavidad y la rareza espacial en un marco mucho más general.

# Granger Causality on Spatial Manifolds: applications to Neuroimaging

Pedro A. Valdés-Sosa, Jose Miguel Bornot-Sánchez, Mayrim Vega-Hernández,  
Lester Melie-García, Agustin Lage-Castellanos, and Erick Canales-Rodríguez  
Department of Neuroinformatics

Cuban Neuroscience Center, Ciudad Habana, Cuba

**Keywords:** Spatial Granger Causality, EEG, fMRI, effective brain connectivity

## 1 Introduction

Devising methods for inferring the effective and functional connectivity of different brain regions is currently a major concern in Neuroimaging [21]. The task is to determine the changing patterns of causal influences that different neural structures exert on each other. This is to be done by the analysis of dynamical brain imaging data. This type of data include EEG/MEG source distributions, optical recordings [65] and fMRI [36] which are, from the statistical point of view, spatiotemporal data sets [48][73] – that is time series sampled from an underlying continuous manifold  $\Omega$  of spatial points. Multivariate autoregressive models (in particular linear ones) for vector time series have proven to be

an essential and informative tool for the applied sciences. Within this framework Granger [33] formulated a definition of causality between time series that has been pursued extensively in many fields and especially in the neurosciences [3][56].

It is striking though, that work in this field has been limited to vector valued time series in which the dimension  $p$  is very small [64][7] –even if, as usual in real applications, the number  $N$  of time samples gathered is large. As Granger himself pointed out, his definition of causality would be valid only if all relevant variables would be included in the analysis, a formidable task that is readily appreciated by neuroscientists since they study the brain, which is the complex system by excellence. We have therefore directed our attention to multivariate autoregressive models (MAR) defined over spatial manifolds (a particular example of which is the brain) and to deal with the issue of densely sampled (high dimensional, highly correlated) time series that arise from a discretization of an underlying spatial continuum into voxels [68].

As an concrete example (whish will be used throughout the paper), consider the concurrent EEG and fMRI time series gathered in order to analyze the origin of resting brain rhythms [31][52][55]. The acquisition paradigm is described more fully in Section 8. Structured patterns of correlations have been found between time-varying spectral components in different EEG bands and the BOLD signal at different voxels. These reveal widely distributed anatomical systems apparently involved in the generation of these oscillations (see Figures 1-5). Here  $N = 108$ , the number of EEG time series is only 16, but the number

of fMRI time series is 12,640! Usual MAR model can not be fit to this amount of data .

The approach explored in this paper follows the strategy of Functional Data Analysis [61]. Quantities of interest in the spatial MAR (autoregressive coefficients) are estimated subject to constraints that make anatomical and physiological sense. They not only allow inference for densely sampled data, but also dovetail nicely with computational shortcuts that make the proposed procedures feasible. In classical MAR models, Granger causality of one set of time series on another set is quantified by means of influence measures [24][25]. In the linear case, these influence measures are usually multivariate tests that certain regression coefficients are zero. In our spatial MAR (sMAR) we extend this concept to that of an influence field. For functional Neuroimages, these are topographic maps of the influence of one brain site (voxel) on rest of the brain. For example in the concurrent EEG-fMRI experiment just mentioned one is interested to know what influence a site in the visual cortex (Figure 1) might have on all the rest of the brain.

For this type of situation classical multivariate testing is difficult or fails. We propose rather to apply the massive univariate approach that is at the heart of Statistical Parametric Mapping (SPM) [74]. SPM essentially calculates a (uni or multivariate) statistic at each voxel of a brain image and then determines significantly activated regions by means of procedures that control the type I error. The latter is achieved either by the use of Random Field Theory [74], resampling methods [11], or the use of the False Discovery rate (FDR) [13].

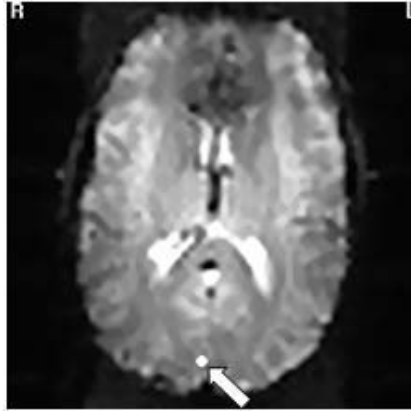


Figure 1: MRI image as an example of a brain manifold. EPI MRI image of the brain of a subject from [30]. The MRI section is at a level that passes through the striate or primary visual cortex (VC). The arrow marks the voxel in VC for which the BOLD response during alpha rhythm shows the highest correlation with the power in that band.

We propose to evaluate a spatial extension of Granger causality by a SPM of influence fields. In effect, we are interested in detecting significant regions in the Cartesian product set  $\Omega \times \Omega$ . An alternative to using ordinary multivariate regression techniques for this situation is to attempt a huge multivariate regression problem and associated testing of the regression coefficients. To be able to do so we shall work with regression based on penalization in the spirit of Functional Data Analysis (FDA) [61]. This approach drastically reduce the number of "effective" connections to be determined. This was the approach taken in [69] by introducing a FDA variant of MAR modeling that imposed spatial smoothness on the influence field. Massive data reduction was achieved by means of the singular value decomposition and this paper showed the feasibility of working in the  $p > N$  situation. A subsequent paper [70] also used penalized regres-

sion, in this case introducing Sparse Multivariate Autoregressive models. The latter can be estimated in a two stage process involving a) penalized regression and b) pruning of unlikely connections by means of the local false discovery rate developed by Efron. Extensive simulations were performed with idealized cortical networks having small world topologies and stable dynamics. These show that the detection efficiency of connections of the proposed procedure is quite high. Furthermore, the sparsity or conditional independence did not have to be specified a priori but is disclosed automatically by an iterative process. In short, we use the fact that the brain is sparsely connected as part of the solution, as opposed to treating it as a specification problem. This chapter unifies the two approaches—spatial smoothness and sparseness in a much more general framework.

## 2 The continuous spatial Multivariate Autoregressive model and its discretization

We shall be dealing with the following spatial Multivariate Autoregressive (sMAR) model defined in discrete time:

$$y(s, t) = \sum_{k=1}^r \iiint_{\Omega} a_k(s, u) y(u, t - k) du + e(s, t) \quad (1)$$

where  $y(s, t)$  is the variable of interest (for example, in our case, either functional Magnetic Resonance Image BOLD, and optical image, EEG, or MEG).

It is a stochastic process which is indexed by the continuous spatial position variable  $s$  and time  $t = 1, \dots, N$ . We posit an innovation process that is also a function of space and time. Note that the integration is over the volumetric set  $\Omega$ . Of central interest here are the functions  $a_k(s, u)$  that specify the influence of site  $u$  on site  $s$  at after  $k$  time delays. This is actually a function  $a_k : \Omega \times \Omega \rightarrow \mathfrak{R}$  which will specify the influences produced by small neighborhoods of each point  $s$  of the manifold  $\Delta(s) \subset \Omega$ , which will be  $a_k(s, u) \Delta(s)$ . We now introduce 3 definitions of spatial influence measures:

- A *point influence measure*  $I_{s \rightarrow u}$  is the simple test  $H_0 : a(s, u) = 0$  for given  $s, u \in \Omega$ .
- An *influence field*  $I_{s \rightarrow \Omega}$  is a multiple test  $H_0 : a(s, u) = 0$  for a given  $s \in \Omega$  and all  $u \in \Omega$ .
- An *influence space*  $I_{s \rightarrow \Omega}$  is a multiple test  $H_0 : a(s, u) = 0$  for all  $s, u \in \Omega$ .

These concepts are illustrated in Figure 2.

Of these, point influence measures have been studied to date and recently we have addressed those for fields. The exploration of the entire influence space will be touched upon in the final section.

Now suppose that we sample the  $y(s, t)$  centering our discretization at voxels  $s = \{s_1, \dots, s_i, \dots, s_p | s_i \in \Omega\}$ . In this case, the data at time  $t$  will be represented by a vector:

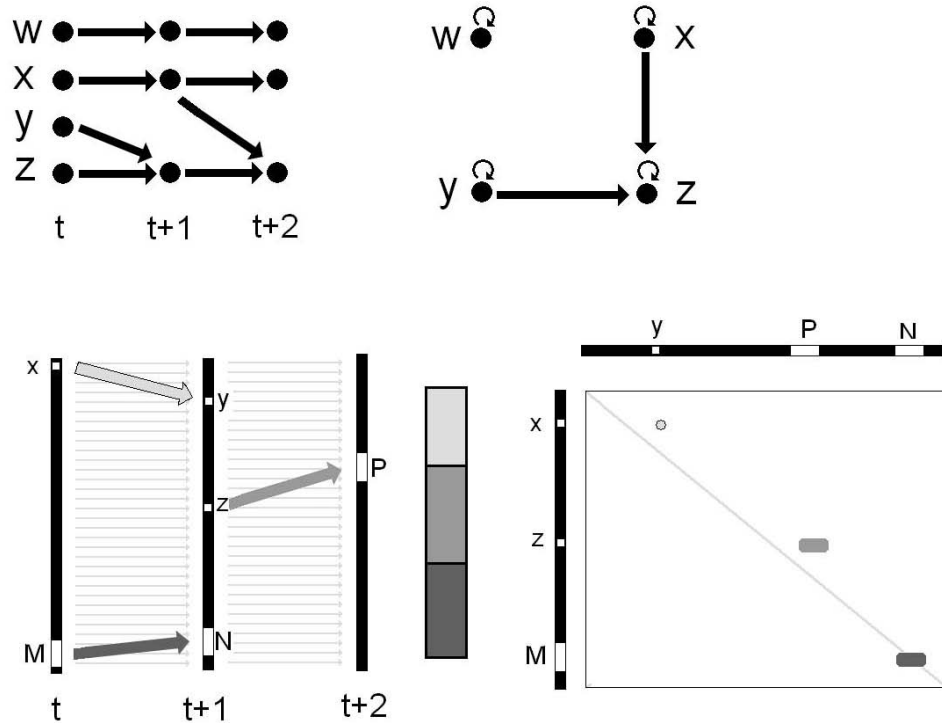


Figure 2: Classical and spatial influence measures. On the left are the set of nodes and how activity is propagated by a linear autoregressive model for successive time instants. Arrows indicate nonzero autoregressive coefficients at different time lags. On the right are the corresponding causality graphs indicating nonzero point influence measures  $I_{x \rightarrow y}$ . Top: causality analysis of a time series graph with only four nodes. In this hypothetical example only two time lags are relevant. Note that each node depends on its own past through a order two autoregressive model. Here we say  $y$  influences  $z$  at lag 1 and  $x$  influences  $z$  at lag 2. Bottom: spatial extension of the concept of influence measure. The manifold  $\Omega$  in this case is a line segment. Also here only two time lags are relevant. Here each point also depends on its past through an order two autoregressive model. Additionally, we also have nonzero point influence measures of  $x$  on  $y$  with lag 1, point  $z$  influences the whole of set  $P$  at lag 2, and set  $M$  influences set  $N$  at lag 1



$$\mathbf{y}_t = \begin{bmatrix} y_{1;t} \\ \vdots \\ y_{i;t} \\ \vdots \\ y_{p;t} \end{bmatrix}_{p \times 1}$$

, where  $i = 1, \dots, p$  indexes the voxels with  $y_{i,t} = \iint_{\Delta(s_i)} y(u, t) du$ . We shall assume that the neighborhood of the  $s_i$  is sufficiently large to avoid spatial aliasing problems. The discretized version of 1 leads to the Multivariate Autoregressive Model (MAR) for the  $\mathbf{y}_t$  :

$$\mathbf{y}_t = \sum_{k=1}^r \mathbf{A}_k \mathbf{y}_{t-k} + \mathbf{e}_t \quad (2)$$

where the continuous function  $a_k(s, s')$  transforms to a matrix  $\mathbf{A}_k$  with dimensions  $p \times p$  and with elements  $a_{i,j}^k = \int \cdots \int_{\Delta(s_i) \times \Delta(u_i)} a_k(s'_i, u'_j) ds' du'$ . In what follows we assume  $\mathbf{e}_t \sim N(0, \boldsymbol{\Sigma})$ , but of course this assumption may be relaxed. Note that the larger the number of sampling points the better the representation so we deal with a case in which ideally  $p \rightarrow \infty$

$$\text{Define } \mathbf{B} = [\mathbf{A}_1, \dots, \mathbf{A}_r]^T, \mathbf{Z} = [\mathbf{y}_{r+1}, \dots, \mathbf{y}_N]^T, \text{ and } \mathbf{X} = \begin{bmatrix} \mathbf{y}_r^T & \cdots & \mathbf{y}_1^T \\ \cdot & & \cdot \\ \cdot & \cdots & \cdot \\ \cdot & & \cdot \\ \mathbf{y}_{N-1}^T & \cdots & \mathbf{y}_{N-r}^T \end{bmatrix}$$

with dimensions  $p \times r$ ,  $N - r \times p$ , and  $N - r \times p$  respectively. we can now

recast the original sMAR 1 as a multivariate regression model:

$$\mathbf{Z} = \mathbf{X}\mathbf{B} + \mathbf{E} \tag{3}$$

where  $E = [\mathbf{e}_{r+1}, \dots, \mathbf{e}_N]^T$ . Some additional notation will be useful. We shall denote the vectorized version of  $\mathbf{B}$ ,  $\boldsymbol{\beta} = \text{vec}(\mathbf{B})$  formed by stacking the columns of  $\mathbf{B}$ ,  $\boldsymbol{\beta}^i$ . Note that  $\boldsymbol{\beta}^i$  measures the influence of a voxel  $i$  on the rest of the brain for all time lags and, in turn, comprises the vectors of autoregressive

coefficients for each time lag:  $\boldsymbol{\beta}^i = \begin{bmatrix} \beta_1^i \\ \dots \\ \beta_r^i \end{bmatrix}$ . Thus the linear effect of voxel  $i$  at lag  $k$  on voxel  $j$  is measured by the coefficient  $\beta_{j,k}^i$ .

### 3 Testing for spatial Granger Causality

As noted before, MAR modeling has been widely applied in the neurosciences [1] [47] [36] for the analysis of causality. Though some doubt that causal analysis is possible at all [40], early work with Structural Equation Modeling [53] did face up to the issue of inferring directional influences and was firmly grounded in modern statistical techniques [58] via graphical models. These initial studies [53] in Neuroimaging were based on non dynamical PET data and ignored temporal information. The concept of Granger Causality [33][41][26]) does make use of temporal information in order to establish a measure of directed influence. Granger Causality  $I_{x \rightarrow y}$  of the time series  $x$  on  $y$  is demonstrated when one can

reject the null hypothesis of  $y$  not being predicted by the past of  $x$  [2][37][3]. Recent work[4] have combined the notion of Granger Causality analysis with that of causality analysis via graphical models [59] . In this view, a system modeled by a MAR is a network in which each node is a time series. These ideas generalize to the more general linear sMAR in Equation 2 introduced above, by noting that the coefficients  $a_{i,j}^k$  measure the influence that time series  $j$  exerts on time series  $i$  after  $k$  time instants. Knowing that  $a_{i,j}^k$  is non-zero is equivalent to establishing effective connectivity [21] and tests for this hypothesis have been proposed as influence measures [33]; [47][27][37][69][15]. From the graphical points of view the question is: does an edge exists between the corresponding nodes? The maximum likelihood (ML) estimation of equation 2, or equivalently equation 3 can be obtained by standard methods [35][48]:

$$\hat{\mathbf{B}} = \arg \min_{\mathbf{B}} \|\mathbf{Z} - \mathbf{XB}\|^2 \quad (4)$$

where for any matrix  $\mathbf{X}$ ,  $\|\mathbf{X}\|^2 = tr(\mathbf{X}^T \mathbf{X})$ , is the Frobenius norm. This results in the well known explicit solution, the OLS estimator:

$$\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Z} \quad (5)$$

It should be noted that the unrestricted ML estimator of the regression coefficients does not depend on the spatial covariance matrix of the innovations [35]. One can therefore carry out separate regression analyses for each node. In other words, it is possible to estimate separately each column

$\beta^i$  of  $\mathbf{B}$ :

$$\hat{\beta}^i = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{z}^i \quad (6)$$

for  $i = 1 \dots p$ , where  $\mathbf{z}^i$  is the  $i$ -th column of  $\mathbf{Z}$ . Consider that we obtain the usual  $t$  statistic for each regression coefficient:

$$t_{k,j}^i = \frac{\hat{\beta}_{k,j}^i}{SE(\hat{\beta}_{k,j}^i)} \quad (7)$$

where  $SE$  is the usual standard error of the regression coefficient. Then we can use SPM type procedures to detect which voxels are influenced by voxel  $i$  at lag  $k$ . This suggests the one possible specific definition of *influence field*:

$$I_{k,i \rightarrow \Omega} = \{t_{k,j}^i\}_{1 \leq j \leq p} \quad (8)$$

If, as is usual, we wish to collapse over the lags, then we use instead of the ordinary  $t$  statistic we can use the Hotelling's  $T^2$  statistic. Unfortunately there is a problem with this approach when dealing with Neuroimaging data: the total number of parameters to be estimated for model 2 is  $g = r \cdot p^2 + \frac{(p^2 + p)}{2}$ , which becomes rapidly large for increasing  $p$ , a situation for which usual time-series methods break down since the OLS estimator will not exist. In the next section we shall review some attempts to cope with this problem by dimensionality reduction in order to apply classical causality analysis. In the following section we shall explain our approach to address the full problem via variable penalization.

## 4 Dimension reduction approaches to sMAR models

### 4.1 ROI based causality analysis

One common approach is to pre-select a small group of sets of voxels or regions of interest (ROI) on the basis of prior knowledge (for example known anatomical structures) and to obtain an average time series over these volumes. In other words the original manifold  $\Omega$  is partitioned into sub-manifolds and the following holds:

$$\begin{aligned}\Omega &= \bigcup_{g=1}^G \Omega_g \\ y_{g,t}^{ROI} &= \iint_{\Omega_g} y(s,t) ds\end{aligned}\tag{9}$$

Causality analysis may then be assayed by the methods described above since now  $N > G$ . Recent example of this type of linear Granger causality analysis for fMRI time series is [29][28]. As an example, a ROI analysis of the concurrent EEG-fMRI times series is shown in Figure 2 (taken from). The fMRI time-series are of length  $N = 109$  for six ROI in the brain identified by previously looking at the correlation with the EEG alpha atom: visual cortex, thalamus, left and right insulae and left and right somatosensory areas. The resulting causality diagram clearly shows that electrophysiological activity is driving the BOLD response in different brain structures, which is to be expected since the BOLD response measured in fMRI experiments is secondary hemodynamic response to neural activity. Thalamus and cortex have reciprocal relations and with other

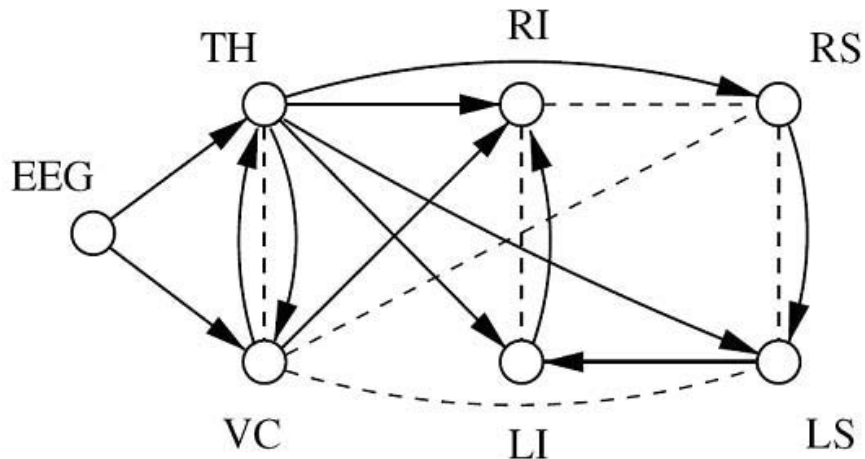


Figure 3: ROI Granger causality graphical model for concurrent EEG-fMRI recording during alpha rhythm. The MRI from Figure 1 has been divided into regions of interest (ROI) and a MAR model fitted to identify significant influences. The EEG node corresponds to the EEG PARAFAC  $\alpha$  component power time series as shown in Figures 3-4. The rest of the nodes are fMRI time

structures. These results in general are in agreement with previous studies of this material showing the utility of this type of analysis. However, the ROI strategy has the potential problem of the appearance of spurious influences induced by the brain structures not included in the analysis. An additional problem is that it is not always clear how to establish the partition (9).

## 4.2 Latent Variable based causality analysis

A different approach for dimensionality reduction is the use of latent variable analysis (LVA). Essentially this involves creating linear or nonlinear combinations of the original time series in an attempt to find series are in some sense

the actual underlying "physiological components":

$$y_{c,t}^{LVA} = f(\mathbf{y}_t) \tag{10}$$

in which  $f$  is the transformation from the original time series to the desired components for  $c = 1, \dots, C$ . This approach has a long history in neuroscience, different methods used being PCA [62][62][20]), ICA [46].

We now give a recent example of LVA which extracted by means of multilinear techniques and applied to the EEG-fMRI data described in [55]. The multichannel EEG evolutionary spectrum  $S(f, t, d)$  is obtained from a channel by channel wavelet transform, where  $\omega$  is frequency,  $d$  is the derivation (channel) and  $t$  is time, Parallel Factor Analysis (PARAFAC) [52][55] decomposes three-way data array  $S$  into the sum of "atoms":

$$S(d, \omega, t) = \sum_k a_k(d) b_k(\omega) c_k(t) + es(\omega, t, d) \tag{11}$$

where the  $k$ -th atom is the trilinear product of loading vectors representing spatial ( $a_k$ ), spectral ( $b_k$ ), and temporal ( $c_k$ ) "signatures". This decomposition is shown schematically in Figure 4. Two atoms were found  $\alpha$  and  $\theta$ , identified on the basis of the frequency signature (Figure 5a) peaking at the known frequency of these well known EEG rhythms. The spatial distribution of these components both in the EEG and the fMRI were occipital and frontal for the  $\alpha$  and  $\theta$  atoms respectively (Color Plate 1). Perusal of the time signatures of these atoms shows a strong influence of imposing either a resting condition or a mental

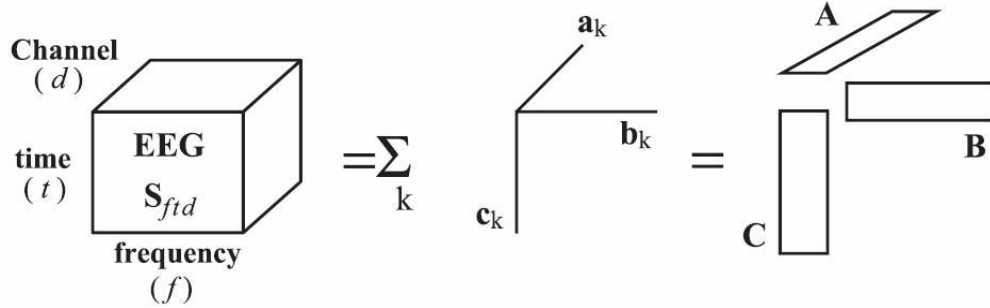


Figure 4: Schematic representation of the PARAFAC model. The multichannel EEG evolutionary spectrum  $S(d, \omega, t)$  is decomposed into the sum of “atoms” where the  $k$ -th atom is the trilinear product of loading vectors representing spatial ( $a_k$ ), spectral ( $b_k$ ), and temporal ( $c_k$ ) “signatures”

task on the subject (Figure 5b). Moreover, since only two time series were involved, classical methods for measuring influences were applied easily yielding the causality analysis shown in Figure 6. It is to be noted that assessment of the model order for all fMRI time series models presented in this paper indicated that only a first order model ( $r = 1$ ) is required .

While consistent with known hypothesis about the brain, this type of analysis only uses the instantaneous covariances to fit the model since time lags are not usually included in the analysis. A more promising approach are methods developed for geostatistics [48][51][73] in which time series methods are combine with component extraction. The latter techniques, to our knowledge, have not been applied in neuroscience. In any case extraction of components avoids the issue of analyzing directly spatial Granger causality, a point to which we shall now turn ou attention.



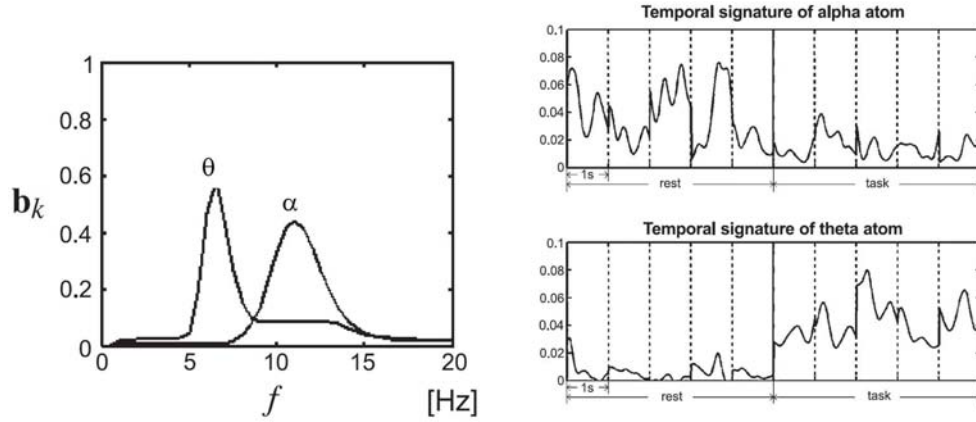


Figure 5: Spectral and temporal signatures of the EEG PARAFAC atoms. On the left the Spectral signatures  $b_k(f)$  of the two atoms corresponding to frequency peaks in the traditional  $\theta$  and  $\alpha$  bands. The horizontal axis is frequency  $\omega$  in Hz and the vertical axis is the normalized amplitude. right temporal signatures,  $c_k(t)$ , of the  $\theta$  and  $\alpha$  atoms.

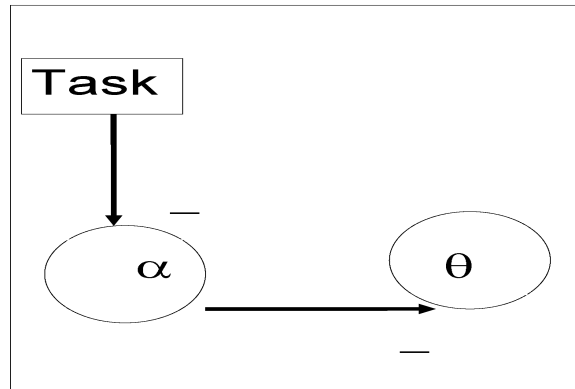


Figure 6: Influence measure analysis of the EEG-fMRI atoms. The external variable imposition of a mental task was found to directly influence (negatively) the activity of the  $\alpha$  atom, which in turn influenced negatively the  $\theta$  atom ( $I_{task \rightarrow \alpha}, I_{\alpha \rightarrow \theta} > 0$ ).

## 5 Penalized sMAR

### 5.1 General model

This section introduces a Bayesian sMAR that generalizes those proposed in [69, Valdes-NI][70, Valdes-PTRS]. Consider once more the sMAR model:

$$\mathbf{Z} = \mathbf{X}\mathbf{B} + \mathbf{E} \quad (12)$$

We now posit that the elements of  $\boldsymbol{\beta}$  are sampled from an a priori that is the product of several generalized multivariate normal densities:

$$\pi(\boldsymbol{\beta}; (P_1, \boldsymbol{\Sigma}_1), \dots, (P_M, \boldsymbol{\Sigma}_M)) = C \cdot \prod_{m=1}^M \exp(-P_m(\boldsymbol{\Sigma}_m^{-1} \boldsymbol{\beta})) \quad (13)$$

where  $C$  is a normalizing constant, the  $\boldsymbol{\Sigma}_m$  are a priori covariance matrices for the  $\boldsymbol{\beta}$ . The MAP estimate that follows from the likelihood of 12 and the prior 13 is:

$$\hat{\mathbf{B}} = \arg \min_{\mathbf{B}} \|\mathbf{Z} - \mathbf{X}\mathbf{B}\|_{\boldsymbol{\Sigma}}^2 + \sum_{m=1}^M P_m(\boldsymbol{\Sigma}_m^{-1} \boldsymbol{\beta}) \quad (14)$$

where for any matrix  $\mathbf{X}$   $\|\mathbf{X}\|_{\boldsymbol{\Sigma}}^2 = \text{tr}(\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})$ . Finally,  $P_m(\mathbf{w})$  for any vector  $\mathbf{w}$  is defined as:  $P_m(\mathbf{w}) = \sum_{l=1}^{\text{length}(\mathbf{x})} p_m(|w_l|)$ , and the functions  $p_m(\theta)$  are defined for  $\theta > 0$  are appropriate penalty functions with the properties specified in [17]. Some examples are given in Table 1 as well as illustrated in Figure 8.

Thus, our model consists of  $M$  regularization constraints, each comprising

a:

1. Covariance matrix used to enforce a priori spatial constraints on the autoregressive coefficients; and a
2. Penalization function to enforce constraints on the magnitude of the variables and therefore carry out variable selection.

Name	Abbreviation
LASSO	L1
Smoothly Clipped Absolute Deviation	SCAD
<i>Hard thresholding</i>	HT
Ridge	L2
Mixture of generalized <i>Gaussians</i>	MIX
Normal-gamma	NG
Normal-exponential-gamma	NEG

Table 1. Examples of penalty functions

Name	notation	inverse of matrix
Spherical	$\mathbf{I}_n$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & 0 & 1 \end{bmatrix}$
1D gradient	$\mathbf{D}_n^1$	$\begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ & & \dots & & \\ 0 & \dots & 0 & 1 & -1 \\ 0 & \dots & 0 & 0 & 1 \end{bmatrix}$
2 D gradient	$\mathbf{D}_{nm}^2$	$\begin{bmatrix} \mathbf{I}_n \otimes \mathbf{D}_m^1 \\ \mathbf{D}_m^1 \otimes \mathbf{I}_n \end{bmatrix}$
2 D laplacian	$\mathbf{L}_{nm}^2$	$\mathbf{D}_n^1 \oplus \mathbf{D}_m^1$
3D gradient	$\mathbf{D}_{nmp}^3$	$\begin{bmatrix} \mathbf{I}_n \otimes \mathbf{I}_m \otimes \mathbf{D}_p^1 \\ \mathbf{I}_n \otimes \mathbf{D}_m^1 \otimes \mathbf{I}_p \\ \mathbf{D}_n^1 \otimes \mathbf{I}_m \otimes \mathbf{I}_p \end{bmatrix}$
3 D-laplacian	$\lambda \mathbf{L}_{n,m,p}^3$	$\lambda \mathbf{D}_m^1 \oplus \mathbf{D}_n^1 \oplus \mathbf{D}_p^1$

Table 2. Examples of a priori covariance  $\Sigma_m$  matrices defined in terms of

their inverses. These definitions are valid over rectangular domains in dimensions from 1-3. For irregular domains (areas in an image where there is gray matter for example) these matrices are masked a 0-1 indicator function for the selected voxels.

The penalization  $p_m$  functions that we have explored are summarized in Ta-

ble 1 with their abbreviations. These abbreviations, together with the those for the covariance matrices  $\Sigma_1$ , allow the introduction of a notation for a particular sMAR model based on the penalty function used. Thus  $(L1, \mathbf{I}_{rp^2})$  is an sMAR model with a penalty that comprises only one term, the use of the  $l1$  penalty and a spherical covariance matrix. It should be noted that the proposed MAP 14 includes as particular cases many currently used regularization schemes frequently applied in isolation, some new combinations proposed in the literature, as well as totally new proposals. Unfortunately in the penalized case it is not possible in general to carry out separate regressions for each  $\beta^i$ . For the sake of simplicity, and to retain the possibility of independent estimation for each influence field, we have been assuming that  $\Sigma$  is diagonal, that is we assume that the innovations are spatially independent. In the final section we shall discuss avenues to avoid this restriction.

## 5.2 Achieving sparsity via variable selection

In a previous paper we proposed that attention be restricted to networks with *sparse connectivity*. That this is a reasonable assumption that is justified by studies of the numerical characteristics of network connectivity in anatomical brain databases [38][49][66])

Sparsity of causal explanations may be achieved by variable selection. Researchers into causality [63][60] have explored the oldest of variable selection techniques for regression—stepwise selection for the identification of causal graphs. This is the basis of popular algorithms such as PC embodied in programmes

such as TETRAD. These techniques have been used in graphical time-series models [8]. Unfortunately these techniques do not work well for  $p \gg N$ . A considerable improvement may be achieved by stochastic search variable selection (SSVS) of George and McCulloch [23][22], which relies on Markov chain–Monte Carlo (MCMC) exploration of possible sparse networks [9][45]. These approaches, however, are computationally very intensive and not practical for implementing a pipeline for Neuroimaging analysis.

An alternative to MCMC like methods is variable selection via penalized regression models [17][18]) which unifies nearly all variable selection techniques into an easy-to implement iterative application of minimum norm or ridge regression. These techniques have been shown to be useful for the identification of the topology of huge networks [50][54]. Penalized regression models were introduced for the first time for the study of brain connectivity used in [70][69]. Consider the variant of the general model 14 with only one component ( $M = 1$ ) and a spherical covariance matrix. Some of the possible models are:

- $(L2, \mathbf{I}_{rp^2})$  is the usual ridge regression model [39] or quadratic regularization,  $\lambda$  being the regularization parameter which determines the amount of penalization enforced. Due to the possibility of efficient computation this is a widely applied form of regularization, recently applied for example to analyze microarray data [72].
- $(L1, \mathbf{I}_{rp^2})$  is, as mentioned above, the LASSO [14].
- $(HT, \mathbf{I}_{rp^2})$  is the Hard Thresholding of regression coefficients only ap-

plicable in the  $p < N$  case.

- $(SCAD, \mathbf{I}_{rp^2})$  [17] is a form of regression designed to avoid bias for larger coefficients.
- $(MIX, \mathbf{I}_{rp^2})$  uses the penalty function  $-\ln(p_0 f_{p0}(\beta) + (1 - p_0) f_{p1}(\beta))$  where the mixture density are univariate generalized gaussians. This is a regression model designed to produce sparsity and implements a non MCMC variant of the "spike and slab" models for variable selection, the best known being the SSVS method of George & McCulloch [23] .

We introduce in this chapter a further generalization of the variable selection penalties previously used. As pointed out in [34] it has been shown that most of the mixture priors previously discussed are particular instances of scale mixtures of normal distributions [71] that embody a high prior probability of the regression coefficients in the proximity of zero. These authors proposed a natural class of prior distribution that bridges the gap between classical normal-Jeffreys priors, passing throughout ridge regression down to the double exponential distribution used in the LASSO. Some particular mixture distribution of interest are shown in Table 3. We single out for mention the following regression models used for the first time to study brain connectivity.:

- $(NG, \mathbf{I}_{rp^2})$ , uses as a penalty the minus log of the normal-gamma (NG) distribution is often called as variance-gamma distribution, has the marginal distribution:  $p(\beta_j) = \frac{1}{\sqrt{\pi 2^{\lambda-1/2} \gamma^{\lambda+1/2} \Gamma(\lambda)}} |\beta_j| K_{\lambda-1/2} (|\beta_j| / \lambda)$ , where  $K_v(a)$  is the modified Bessel function of the third kind.

- $(NEG, \mathbf{I}_{rp^2})$  is based on the normal-exponential-gamma (NEG) can be expressed as  $p(\beta_j) = \frac{\lambda 2^\lambda}{\sqrt{\pi} \gamma} \Gamma(\lambda + 1/2) \exp\left(\frac{\beta_j^2}{4\gamma^2}\right) D_{-2(\lambda+1/2)}(|\beta_j|/\lambda)$ , where  $D_v(a)$  is the parabolic cylinder function, the parameters  $\gamma$  and  $\lambda$  control the scale and the heaviness of the tail respectively.

distribution	density
Normal-Jeffreys:	$g(\theta) \propto 1/\theta$
t distribution	$g(\theta) = IG\left(\frac{\lambda}{2}, \frac{\gamma^2 \lambda}{2}\right)$ $\lambda, \gamma > 0$
Mean-zero double exponential	$g(\theta) = Ga\left(\theta \lambda, \frac{1}{2\gamma^2}\right)$ $\lambda = 1$
Normal-gamma (NG)	$g(\theta) = Ga\left(\theta \lambda, \frac{1}{2\gamma^2}\right)$ $\lambda > 0, \gamma < \infty$
Normal-exponential-gamma (NEG)	$g(\theta) = \frac{\lambda}{\gamma^2} (1 + \theta/\gamma^2)^{-(\lambda+1)}$ $\lambda > 0, \gamma < \infty$

Table 3. Mixing distribution of interest represented in the scale mixture form, where  $IG(a, b)$  and  $Ga(a, b)$  are the inverse gamma and the gamma with shape  $a$  and natural parameter  $b$ .

### 5.3 Achieving spatial smoothness

The other constraint that makes sense is that of *spatial smoothness* of influence fields. Consider Figure 8(left) which depicts the influence of a given brain structure on three others: two that are close to each other in the same hemisphere and another that is further away in another hemisphere. It is a priori



more likely that the influences *from* the given voxel on the two closer voxels be more similar than the influence on the distant voxel. This can be quantified by requiring  $\sum_{k=1}^r \iiint_{\Omega} \left| \frac{\partial a_k(s,u)}{\partial s} \right|^2 du$  be small, the the distribution of influences to *targets* be smooth. Alternatively, one may require that the distribution of *sources* influences to a single target as in Figure 8(right) be smooth by imposing that  $\sum_{k=1}^r \iiint_{\Omega} \left| \frac{\partial a_k(s,u)}{\partial u} \right|^2 du$  be small. These definitions are actually for the  $L2$  penalization (and therefore specify Gaussian fields as a priori distributions). The discrete version of this is set up by specifying the matrix operators defined in Table 3. Additionally, one may modify the quadratic norm by applying the different penalties described in Table 1. One may also conceive combinations of the two conditions– smoothness of target or of source influences all these conditions following from the choice of appropriate roughness penalty or, equivalently, the a priori covariance matrix. Imposing smoothness on the influence fields involves imposing conditions on each column of  $\mathbf{B}$  ( $\beta^i$ ) separately. It would be possible to impose similar conditions on the rows of  $\mathbf{B}$ , that is on the map of sources of a given target, but this is not computationally feasible at the moment for large  $p$ .

We shall now mention some one component sMAR models that impose different types of smoothness:

- $(L1, \mathbf{L}_{rp^2})$  this is the data "Fusion" model mentioned in [67], now applied to sMAR.
- $(L2, \mathbf{L}_{rp^2})$  is a spline regression model in which the spatial laplacian of the estimated coefficients are to be minimized. Popularized for the solution

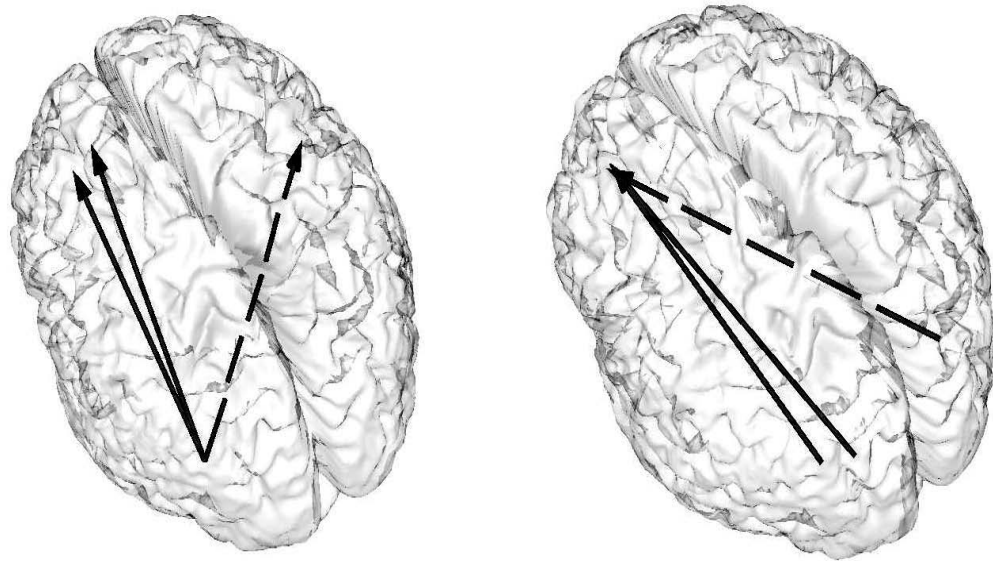


Figure 7: Spatial Constraints

of EEG inverse problems as "LORETA" [57] , this model was used for the first time to study fMRI time-series connectivity in one of our previous paper [69].

We wish to emphasize that penalizing with roughness penalties is equivalent to penalizing a spatial Fourier transform of the coefficients to be estimated.

#### 5.4 Achieving sparseness *and* smoothness

There is no reason to restrict the number of penalty/smoothness constraints imposed. In fact, recent work in statistical learning has advanced the use of models which are easily recognized in the framework of our general model. For example:

- $(L1, \mathbf{I}_{rp^2})$   $(L2, \mathbf{I}_{rp^2})$  can be recognized as the recently introduced "Elastic Net" [75] regression technique applied to sMAR. The elastic net has been shown to improve on the variable selection properties of the LASSO when  $p \gg N$ . Simulations have shown that when there are sets of correlated variables LASSO picks just one variable from each set. In contrast, the elastic net picks all of the members of the set giving them similar weights. When applied to sMAR this would produce a "patchy" influence field. One would hope that these patches correspond to coherent sets of neurons that act together in influencing other brain structures.
- $(L1, \mathbf{I}_{rp^2})$   $(L1, \mathbf{D}_{rp^2})$  can be recognized as the recently introduced "LASSO-Fusion" [67] regression technique applied to sMAR. It is claimed that this

also selects patches of related variables and outperforms the LASSO when  $p \gg N$ .

Both these procedures were previously developed in the context of particular algorithms: quadratic programming and LARS for LASSO-Fusion and the elastic net respectively. However, we have that it is possible even for huge problems(see next section) to work with any number of combinations of penalties/covariance matrices. We have therefore tried out the following new models:

- $(L2, \mathbf{I}_{rp^2}) (L2, \mathbf{D}_{rp^2})$  which we call "Ridge-Fusion" in analogy to LASSO-Fusion.
- $(L1, \mathbf{I}_{rp^2}) (L1, \mathbf{L}_{rp^2}) (L2, \mathbf{I}_{rp^2}) (L2, \mathbf{L}_{rp^2})$  which can be seen either as: a) a combination of the LASSO-Fusion and Ridge-Fusion or, alternatively as b) a combination of the Elastic NET applied with LORETA both for the L1 and L2 norm.

From our previous comment at the end of the last section it is obvious that these attempts to combine norms are equivalent to penalizing/selecting variables from the original coefficient domain as well as from the spatial frequency domain.

## 6 Estimation via the MM algorithm

For implementation of algorithms for the estimation of the model 14 , advantage was taken of the recent demonstration [17][18][42] that estimation of any of

many penalized regression for the influence field of voxel  $i$  can be carried out by iterative application of ridge regression:

$$\hat{\boldsymbol{\beta}}_{k+1}^i = (\mathbf{X}^T \mathbf{X} + \mathbf{D}(\hat{\boldsymbol{\beta}}_{k+1}^i))^{-1} \mathbf{X}^T \mathbf{z}_i \quad (15)$$

where  $k = 1, \dots, N_{iter}$ , with  $N_{iter}$  the number of iterations and  $\mathbf{D}(\hat{\boldsymbol{\beta}}_{k+1}^i)$ , a diagonal matrix is defined by

$$\mathbf{D}(\boldsymbol{\beta}^i) = \sum_{m=1}^M \text{diag}(p'_m(|w_l^i|)/|w_l^i|) \quad (16)$$

for  $l = 1, \dots, rp^2$ , where  $\mathbf{w} = \Sigma_m^{-1} \boldsymbol{\beta}^i$  and  $p'_\lambda$  is the derivative of the penalty function being evaluated. the derivatives  $p'_m$  for different penalty functions are provided in Table 4

Type	Derivatives
L1	$p'_\lambda(\theta) = \lambda_{L1} \theta$
SCAD	$p'_\lambda(\theta) = \lambda_{SCAD} \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)}{a-1} I(\theta > \lambda) \right\}$ for some $a > 2$
HT	$p'_\lambda(\theta) = -2(\theta - \lambda_{HT})$
L2	$p'_\lambda(\theta) = 2\lambda_{L2} \theta$
MIX	$p'_\lambda(\theta) = -\lambda_{Mix} \left[ \frac{p_0 f'_{p_0}(\theta) + p_1 f'_{p_1}(\theta)}{p_0 f_{p_0}(\theta) + p_1 f_{p_1}(\theta)} \right]$ where $f_p(\theta) = \frac{p^{1-\frac{1}{p}}}{2\sigma_p \Gamma(\frac{1}{p})} \exp\left(-\frac{1}{p} \frac{ x-x_0 ^p}{\sigma_p}\right)$ and $\Gamma(\cdot)$ denotes the Gamma function
NG	$p'_\lambda(\theta) = \frac{1}{\gamma_{NG}} \frac{K_{\lambda-3/2}\left(\frac{\theta}{\gamma_{NG}}\right)}{K_{\lambda-1/2}\left(\frac{\theta}{\gamma_{NG}}\right)}$ where $K_v(z)$ is the modified Bessel function of the third kind
NEG	$p'_\lambda(\theta) = \frac{\lambda_{NG}+1/2}{\gamma_{NG}} \frac{D_{-2(\lambda+1)}\left(\frac{\theta}{\gamma_{NG}}\right)}{D_{-2(\lambda+1/2)}\left(\frac{\theta}{\gamma_{NG}}\right)}$ where $D_v(z)$ is the parabolic cylinder function

Table 4.  $p'_\lambda(\theta)$ , derivatives of penalty functions for  $\theta > 0$

The reason that this algorithm works may be inferred from Figure 8. At each step of the iterative process, the regression coefficients of each node with all others are weighted according to their current size and the penalty function chosen. Many coefficients are successively down-weighted and ultimately set to zero—effectively carrying out variable selection in the case of the LASSO, HT, SCAD, MIX, and NG penalization. It must be emphasized that the number of variables set to zero in any of the methods described will depend on the value of the regularization parameter, with higher values selecting fewer variables. In

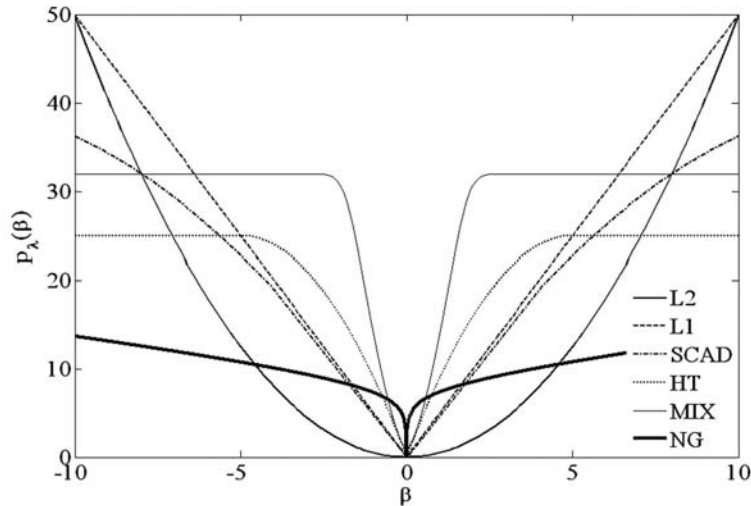


Figure 8: Plot of the penalization functions used to implement sparse and spatially constrained regression techniques. The meaning of the abbreviations is summarized in Table 1.

this paper, the value of the tuning parameters was selected to minimize the generalized crossvalidation criterion (GCV).

The specific implementation of penalized regression used in this work is that of the maximization– minorization (MM) algorithm [42][43][44] which exploits an optimization technique that extends the central idea of EM algorithms and Variational Bayes techniques to situations not necessarily involving missing data or even maximum likelihood estimation. The MM algorithm retains virtues of the Newton-Raphson algorithm. It is numerically stable and is never forced to delete a covariate permanently in the process of iteration. The general convergence results known for MM algorithms imply among other things that the newly proposed algorithm converges correctly to the maximizer of the perturbed penalized likelihood whenever this maximizer is the unique local max-

imum. The selected model based on the maximized penalized likelihood satisfies  $p_m(|w_l^i|) = 0$  for certain  $\mathbf{w} = \Sigma_m^{-1} \beta^i$ , which components accordingly are not included in this final model, and so model estimation is performed at the same time as model selection. The tuning parameters  $\lambda_M$  may be chosen by a data-driven approach such as cross-validation or generalized cross-validation[32]. An important point is that Hunter and Li showed that simple use of iterations 15 with the matrix  $\mathbf{D}$  may permanently delete variables permanently from consideration being included in further iterations.

Hunter and Li [44] showed that a perturbed version of  $p_m(\theta)$ , may be used to define a new objective function that is similar to the original but does not lead to permanent variable deletion. To this end, they define:

$$p_{m,\epsilon}(\theta) = p_m(\theta) - \epsilon \int_0^{|\theta|} \frac{p_\lambda}{\epsilon + t} dt \quad (17)$$

which in practice is equivalent to using the following matrix  $\mathbf{D}_\epsilon$  instead of  $\mathbf{D}$ :

$$\mathbf{D}_\epsilon(\beta^i) = \sum_{m=1}^M \text{diag}(p'_m(|w_l^i|) / (|w_l^i| + \epsilon)) \quad (18)$$

Note that in the computations the original set of variables to be estimated  $\beta$  is by definition augmented with spatial transforms (defined by the matrix operators laid out in Table 2). Suppose that we have defined a model with covariance matrices  $\Sigma_1, \dots, \Sigma_M$ . Then we can use the following computational



"trick", defining  $\mathbf{S} = [\boldsymbol{\Sigma}_1^{-T}, \dots, \boldsymbol{\Sigma}_M^{-T}]^T$  and  $\mathbf{T} = \frac{1}{M} [\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_M]$  we have

$$\mathbf{q} = \mathbf{S} \boldsymbol{\beta} \quad (19)$$

one may carry out penalized regression on this new set of variables by defining  $\mathbf{X}_M = \mathbf{X} \mathbf{T}$  and solving the new (larger) problem, where the definition of  $\mathbf{Q}$  is self evident:

$$\hat{\mathbf{Q}} = \arg \min_{\mathbf{B}} \|(\mathbf{Z} - \mathbf{X}_M \mathbf{Q})\|_{\boldsymbol{\Sigma}}^2 + \sum_{m=1}^M P_m(\mathbf{q}) \quad (20)$$

Back transformation to the desired solution is obtained by  $\hat{\mathbf{B}} = \mathbf{T} \hat{\mathbf{Q}}$ . We have found this algorithm to work well in practice

## 7 Evaluation of simulated data.

The procedures described in the two previous sections have been thoroughly tested with simulated data. For simulations an "ideal cortex" was modeled by a small world network defined over a two dimensional grid on the surface of a torus (Figure 9). This structure has periodic boundary conditions in the plane.

In simulations described in detail in [70], the existence of a connection was generated with a binomial probability that decreased with distance. The network mean connectivity was: 6.23, the scaled clustering: 0.87, the scaled length: 0.19. This type of small-world network has a high probability of connections between geographical neighbors and a small proportion of larger range connections. The network mean connectivity was: 6.23; the scaled clustering: 0.87;

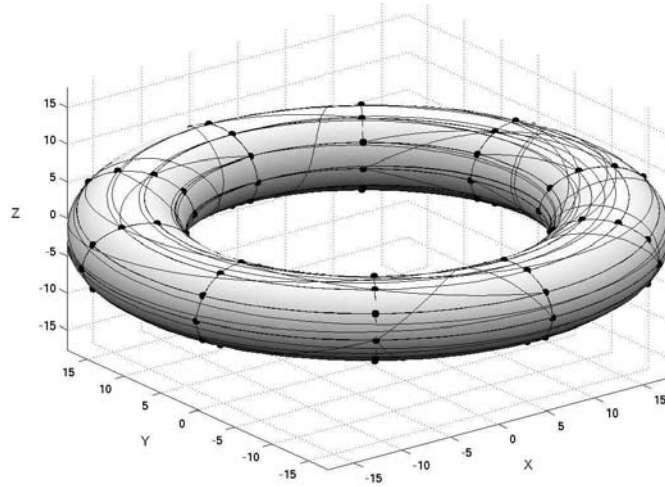


Figure 9: Ideal "cortex" us for simulations was modeled by a small world network defined over a two dimensional grid on the surface of a torus. This structure has periodic boundary conditions in the plane. Different combinations of strengths for were used for defining the autoregressive matrices used to create simulated fMRI time series.

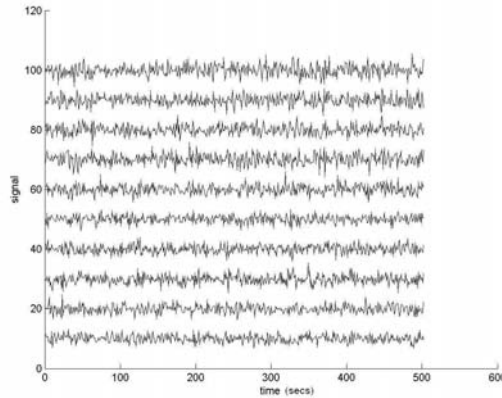


Figure 10: Simulated fMRI time-series generated by a first order multivariate autoregressive model.

the scaled length: 0.19. The autoregressive matrix being sampled from equation 2. The innovations were sampled from a Gaussian distribution with a different prescribed covariance matrixes, including non diagonal ones. A simulated fMRI is shown in Figure 10. The effect of different observed lengths of time-series (N) on the detection of connections was studied. The behavior of different procedures was compared by measuring the area under the ROC curve (AUC). We found that while performance deteriorated with an increasing  $\frac{p}{N}$  ratio there was still significant detection rates with this ratio near 10. The performance of the methods also deteriorated with increasing spatial innovation correlation. This latter observation underscores the need for estimating also the covariance matrix  $\Sigma$ . Doing this with computational efficiency is still work in progress.

A number of further simulations were carried out in similar conditions as those reported before to explore the usefulness of multiple penalty/covariance matrix combinations. The  $\frac{p}{N}$  ratio was now set at 2. From Table 4 it is evident that, except for one exception, imposing simultaneously sparseness and smoothness outperforms either criteria alone.

Method	<b>I</b>	<b>L</b>	<b>I + L</b>
<i>L2</i>	0.6825	0.7026	0.7438
<i>L1</i>	0.6157	0.7102	0.7657
<i>L1 + L2</i>	0.5766	0.6222	0.6257
<i>NG</i>	0.6722	0.6963	0.7434

Table 5. The numerical results of simulations testing of the ROC for the different studied methods are presented.

## 8 Influence fields for real data

To be able to apply these techniques to actual data it is necessary to have a decision procedure as to which variables to finally retain. We have found that although the methods described above do enforce considerable selection of variables, there is still a "grey zone" of variables with small values, for which the decision has to be taken as whether to include or not.

We have therefore combined methods for penalized regression with procedures for the control of the false discovery rates (FDR) [10][11][12] in situations where a large number of null hypothesis is expected to be true. The situation  $p \gg n$  this case becomes strength instead of a weakness, because it allows the non-parametric estimation of the distribution of the null hypotheses to control false discoveries. To carry out this type of decision procedure it is preferable to work with the influence measures defined by the  $t$  statistics 7. For this we must estimate the standard errors of the  $\hat{\beta}$ . We have explored two procedures for this estimation. One is the "sandwich" formulas as described in [6][44][16]. However, we have found the estimation of the standard errors by means of the bootstrap more robust than with the sandwich estimator.

In [70] it was shown that efficient detection of connections possible simulated neural networks. The method was additionally shown to give plausible results with real fMRI data and is capable of being scaled to analyze very large data sets. In that publication the variable-selection method combined with FDR was illustrated by the identification of the neural circuitry related to emotional processing as measured by BOLD.

As a final, real-world example, we describe in some more detail the concurrent EEG-fMRI experiment that has been used as an example throughout this paper. This is a problem of sufficient size to test the practicality of the procedures proposed since  $p$  the number of voxels is 16,240 and  $N$  is only 108. The EEG was sampled at 200 Hz from an array of 16 bipolar pairs, (Fp2-F8, F8-T4, T4-T6, T6-O2, O2-P4, P4-C4, C4-F4, F4-Fp2; Fp1-F7, F7-T3, T3-T5, T5-O1, O1-P3, P3-C3, C3-F3, F3-Fp1), with an additional channel for the EKG and scan trigger. The fMRI time series was measured in six slice planes (4 mm, skip 1mm) parallel to the AC-PC line, with the second from the bottom slice through AC-PC. More details about this data set can be found in [30]. In the work presented here we report a typical subject from a set of five simultaneous EEG/fMRI recordings from three different subjects.

For the fMRI, we examined the influence field with a source at that voxel that had the largest (negative) correlation with the EEG PARAFAC component for  $\alpha$  rhythm. This latter component is the one obtained in the section above on LVA methods and shown topographically in Color Plate 1(left) . The selected voxel is marked in Figure 1 (arrow).

The influence fields for the selected voxel obtained by using different models is shown in Color Plate 2. The penalties are labeled on the left and the covariances on the top. It is to be noted that the use of the spherical covariance matrix produces quite "rough" influence fields. When combined with the  $L1$  penalty only a scattering of points is selected, at most the same as  $N$  that is 108— a known property of the LASSO. The  $(L2, \mathbf{L}_{r,p^2})$  solution ("Ridge-Fusion") pro-

duces a more pleasing (but perhaps excessively smooth) map that is in very good correspondence with previous studies with simple correlations as well as with PARAFAC. All the most realistic seeming solutions are those that combine the spherical covariance matrix as well as the laplacian roughness penalty. In fact, the solution that combines the spherical and laplacian covariance matrices and also the  $L1$  and  $L2$  norm seems to be subjectively the best solution. This impression is born out by comparison of the GCV values for all models. GCV not only serves to fit the tuning parameters but also provides a yardstick for comparing models. In this particular case, related to the models fit and displayed in Color Plate 2) there is a progressive decrease of GCV from top to bottom and from left to right, indicating that the simpler models do not provide adequate modeling flexibility and providing some empirical support for the usefulness of model 13.

## 9 Possible extensions and conclusions

Work with the SMAR model (13) is proceeding in several directions. Obviously this approach can be extended for nonlinear autoregressions. This can be done by

- Including bilinear (or higher order) terms in the  $\mathbf{X}$  matrix [5]; or by
- Defining a kernel weighting in the state space for the autoregressive coefficients as in [19].

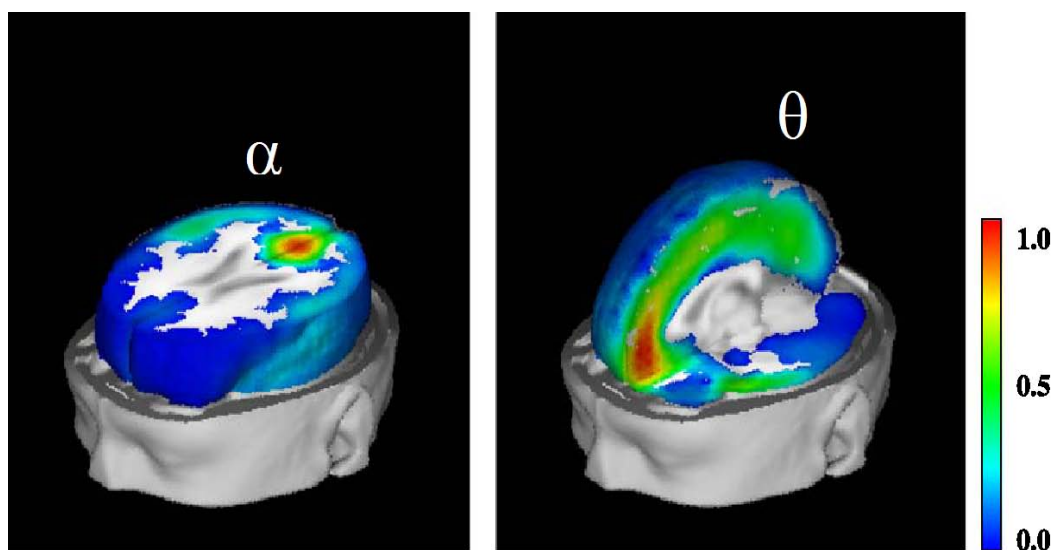


Figure 11: Color Plate 1 Spatial distribution of the  $\alpha$  and  $\theta$  atoms as determined by both PARAFAC of the EEG and Multilinear Partial Least Squares of concurrent EEG-fMRI recordings. Inverse solutions obtained from the spatial  $\alpha_k$  signatures. Note the occipital and frontal distributions of the spatial signature for the  $\alpha$  and  $\theta$  atoms respectively.

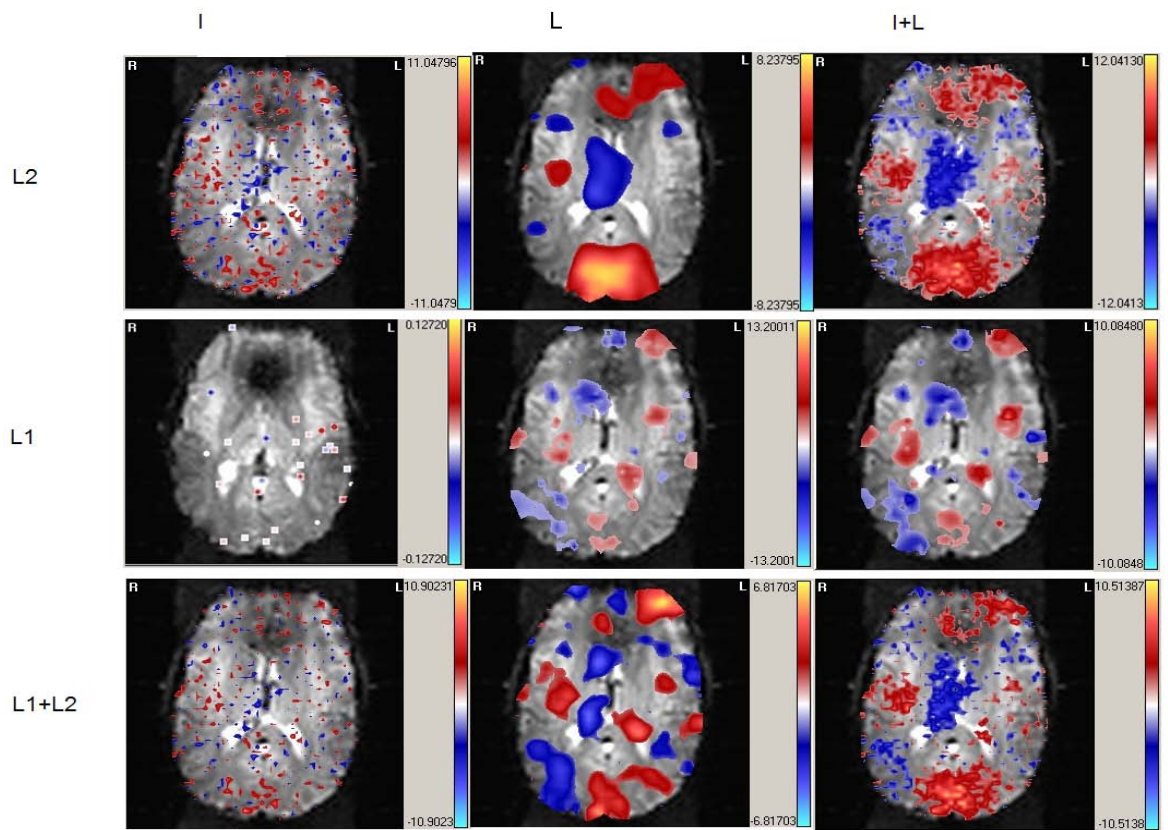


Figure 12: Color Plate 2. Results of fitting the sMAR with multiple penalties/covariance matrixes. The a priori covariance matrix assumed are stated on the top (spherical, laplacian, and a combination of both). The type of penalization is stated on the left (L2 norm, L1 norm, and a combination of both known as the elastic net). Each sub figure is the influence field of the voxel marked in Figure 1 with an arrow on the rest of the voxels corresponding to the slice immediately below .



On the other hand, a kernel method at different times would accommodate nonstationary time series as in [37].

Extensions to the frequency domain of sMAR causality analysis are quite straight forward. Either the sandwich formula or the bootstrap can be used to provide estimates of any linear combination of influence fields and therefore to the temporal Fourier transform of the influence fields over the different delays.

A vexing problem is the estimation of the covariance matrix  $\Sigma$ . We are currently attempting to this by including a zero lag autoregressive matrix  $\mathbf{A}_0$  in the formulation of the model.

In conclusion, we have introduced a spatial multivariate autoregressive model based on a Bayesian formulation that combines several components of different types of penalizations as well as spatial a priori covariance matrices. These are shown by simulations and work with real data to be practical, even for huge data sets, and that give plausible results. The methods continue to bring into the framework of Statistical Parametric Mapping the analysis of effective connectivity via the analysis of Granger Causality.

## 10 Acknowledgments

We wish to thank Maria Luisa Bringas for her untiring help in the preparation of this paper. Also I would like to thank Mark Cohen and Robin Goldman for their continuing support and intellectual exchange.

## 11 Figure Legends

Figure 1 MRI image as an example of a brain manifold. EPI MRI image of the brain of a subject from [30]. The MRI section is at a level that passes through the striate or primary visual cortex (VC). The arrow marks the voxel in VC for which the BOLD response during alpha rhythm shows the highest correlation with the power in that band.

Figure 2 Classical and spatial influence measures. On the left are the set of nodes and how activity is propagated by a linear autoregressive model for successive time instants. Arrows indicate nonzero autoregressive coefficients at different time lags. On the right are the corresponding causality graphs indicating nonzero point influence measures  $I_{x \rightarrow y}$ . a) Causality analysis of a time series graph with only four nodes. In this hypothetical example only two time lags are relevant. Note that each node depends on its own past through a order two autoregressive model. Here we say  $y$  influences  $z$  at lag 1 and  $x$  influences  $z$  at lag 2. b) Spatial extension of the concept of influence measure. The manifold  $\Omega$  in this case is a line segment. Also here only two time lags are relevant. Here each point also depends on its past through an order two autoregressive model. Additionally, we also have nonzero point influence measures of  $x$  on  $y$  with lag 1, point  $z$  influences the whole of set  $P$  at lag 2, and set  $M$  influences set  $N$  at lag 1.

Figure 3 ROI Granger causality graphical model for concurrent EEG-fMRI recording during alpha rhythm. The MRI from Figure 1 has been divided into regions of interest (ROI) and a MAR model fitted to identify significant influ-

ences. The EEG node corresponds to the EEG PARAFAC  $\alpha$  component power time series as shown in Figures 3-4. The rest of the nodes are fMRI time series obtained by averaging activity over the following ROI: TH (thalamus), VC (Visual cortex), RI (right insula), LI (left insula), RS (right somatosensory cortex), and LS (left somatosensory cortex).

Figure 4 Schematic representation of the PARAFAC model. The multi-channel EEG evolutionary spectrum  $S(d, \omega, t)$  is decomposed into the sum of “atoms” where the  $k$ -th atom is the trilinear product of loading vectors representing spatial ( $a_k$ ), spectral ( $b_k$ ), and temporal ( $c_k$ ) “signatures”.

Figure 5 Spectral and temporal signatures of the EEG PARAFAC atoms. On the left the Spectral signatures  $b_k(f)$  of the two atoms corresponding to frequency peaks in the traditional  $\theta$  and  $\alpha$  bands. The horizontal axis is frequency  $\omega$  in Hz and the vertical axis is the normalized amplitude. right temporal signatures,  $c_k(t)$ , of the  $\theta$  and  $\alpha$  atoms.

Figure 6 Influence measure analysis of the EEG-fMRI atoms. The external variable imposition of a mental task was found to directly influence (negatively) the activity of the  $\alpha$  atom, which in turn influenced negatively the  $\theta$  atom ( $I_{task \rightarrow \alpha}, I_{\alpha \rightarrow \theta} > 0$ ).

Figure 7 Penalization functions

Plot of the penalization functions used to implement sparse and spatially constrained regression techniques. The meaning of the abbreviations is summarized in Table 1.

Figure 8 Spatial Constraints

Figure 9 Ideal "cortex" us for simulations was modeled by a small world network defined over a two dimensional grid on the surface of a torus. This structure has periodic boundary conditions in the plane. Different combinations of strengths for were used for defining the autoregressive matrices used to create simulated fMRI time series.

Figure 10 Simulated fMRI time-series generated by a first order multivariate autoregressive model.

Color Plate 1 Spatial distribution of the  $\alpha$  and  $\theta$  atoms as determined by both PARAFAC of the EEG and Multilinear Partial Least Squares of concurrent EEG-fMRI recordings. Inverses solutions obtained from the spatial  $\alpha_k$  signatures. Note the occipital and frontal distributions of the spatial signature for the  $\alpha$  and  $\theta$  atoms respectively.

Color Plate 2. Results of fitting the sMAR with multiple penalties/covariance matrixes. The a priori covariance matrix assumed are stated on the top (spherical, laplacian, and a combination of both). The type of penalization is stated on the left (L2 norm, L1 norm, and a combination of both known as the elastic net). Each sub figure is the influence field of the voxel marked in Figure 1 with an arrow on the rest of the voxels corresponding to the slice immediately below

## References

- [1] L. A. Baccala, M. A. L. Nicolelis, C. H. Yu, and M. Oshiro, *Structural-analysis of neural circuits using the theory of directed-graphs*, Computers and Biomedical Research **24** (1991), 7–28.
- [2] C. Bernasconi and P. Konig, *On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings*, Biological Cybernetics **81** (1999), 199–210.
- [3] S. L. Bressler, M. Z. Ding, and W. M. Yang, *Investigation of cooperative cortical dynamics by multivariate autoregressive modeling of event-related local field potentials*, Neurocomputing **26-7** (1999), 625–631.
- [4] D. R. Brillinger, H. L. Bryant, and J. P. Segundo, *Identification of synaptic interactions*, Biological Cybernetics **22** (1976), 213–228.
- [5] C. Buchel and K. Friston, *Interactions among neuronal systems assessed with functional neuroimaging*, Revue Neurologique **157** (2001), 807–815.
- [6] R. J. Carroll, S. Wang, D. G. Simpson, A. J. Stromberg, and D. Ruppert, *The sandwich (robust covariance matrix) estimator*, Technical Report. Preprint available at <http://stat.tamu.edu/ftp/pub/rjcarroll/sandwich.pdf>. (1998).
- [7] R. Dahlhaus, *Fitting time series models to nonstationary processes*, Annals of Statistics **25** (1997), 1–37.

- [8] S. Demiralp and K. D. Hoover, *Searching for the causal structure of a vector autoregression*, Oxford Bulletin of Economics and Statistics **65** (2003), 745–767.
- [9] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. A. Yao, and M. West, *Sparse graphical models for exploring gene expression data*, Journal of Multivariate Analysis **90** (2004), 196–212.
- [10] B. Efron, *Robbins, empirical bayes and microarrays*, Annals of Statistics **31** (2003), 366–378.
- [11] ———, *Large-scale simultaneous hypothesis testing: The choice of a null hypothesis*, Journal of the American Statistical Association **99** (2004), 96–104.
- [12] ———, *Bayesians, frequentists, and physicists*, <http://www-stat.stanford.edu/brad/papers/physics.pdf> (2005).
- [13] ———, *Selection and estimation for large-scale simultaneous inference*, <http://www-stat.stanford.edu/people/faculty/efron/papers.html> (2005).
- [14] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Annals of Statistics **32** (2004), 407–451.
- [15] M. Eichler, *Graphical time series modelling in brain imaging*, Philosophical Transaction of the Royal Society of London, B **index issue** (2005).

- [16] P. H. C. Eilers, I. D. Currie, and M. Durban, *Fast and compact smoothing on large multidimensional grids*, Computational Statistics and Data Analysis **50** (2006), 61–76.
- [17] J. Q. Fan and R. Z. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association **96** (2001), 1348–1360.
- [18] J. Q. Fan and H. Peng, *Nonconcave penalized likelihood with a diverging number of parameters*, Annals of Statistics **32** (2004), 928–961.
- [19] W. A. Freiwald, P.A. Valdes, J. Bosch, R. Biscay, J. C. Jimenez, L. M. Rodriguez, V. Rodriguez, A. K. Kreiter, and W. Singer, *Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex*, Journal of Neuroscience Methods **94** (1999), 105–119.
- [20] K. Friston, J. Phillips, D. Chawla, and C. Buchel, *Revealing interactions among brain systems with nonlinear pca*, Human Brain Mapping **8** (1999), 92–97.
- [21] K. J. Friston, *Functional and effective connectivity in neuroimaging: a synthesis*, Human Brain Mapping **2** (1994), 56–78.
- [22] E. I. George, *The variable selection problem*, Journal of the American Statistical Association **95** (2000), 1304–1308.
- [23] E. I. George and R. E. McCulloch, *Approaches for bayesian variable selection*, Statistica Sinica **7** (1997), 339–373.

- [24] J. F. Geweke, *Measurement of linear-dependence and feedback between multiple time-series*, Journal of the American Statistical Association **77** (1982), 304–313.
- [25] ———, *Measures of conditional linear-dependence and feedback between time-series*, Journal of the American Statistical Association **79** (1984), 907–915.
- [26] ———, *Measures of conditional linear-dependence and feedback between time-series*, Journal of the American Statistical Association **79** (1984), 907–915.
- [27] R. Goebel, A. Roebroeck, D. S. Kim, and E. Formisano, *Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping*, Magn Reson. Imaging **21** (2003), 1251–1261.
- [28] ———, *Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping*, Magnetic Resonance Imaging **21** (2003), 1251–1261.
- [29] R. Goebel, T. D. Waberski, H. Simon, E. Peters, F. Klostermann, G. Curio, and H. Buchner, *Different origins of low- and high-frequency components (600 hz) of human somatosensory evoked potentials*, Clinical Neurophysiology **115** (2004), 927–937.
- [30] R. I. Goldman, J. M. Stern, J. Engel, and M. S. Cohen, *Simultaneous eeg and fmri of the alpha rhythm*, Neuroreport **13** (2002), 2487–2492.



- [31] R.I. Goldman, J. M. Stern, J. Engel, and M. S. Cohen, *Acquiring simultaneous eeg and functional mri*, *Clinical Neurophysiology* **111** (2000), 1974–1980.
- [32] G. H. Golub, M. Heath, and G. Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, *Technometrics* **21** (1979), 215–223.
- [33] C. W. J. Granger, *Investigating causal relations by econometric models and cross-spectral methods*, *Econometrica* **37** (1969), 414–and.
- [34] Brown P.J. Griffin J.E., *Alternative prior distributions for variable selection with very many more variables than observations*, Tech. report, Department of Saticistic, University of Warwick, Coventry, CV4 7AL, U.K., 2005.
- [35] J. D. Hamilton, *Time series analysis*, Princeton University Press: Princeton, New Jersey, 1999.
- [36] L. Harrison, W. D. Penny, and K. Friston, *Multivariate autoregressive modeling of fmri time series*, *Neuroimage* **19** (2003), 1477–1491.
- [37] W. Hesse, E. Moller, M. Arnold, and B. Schack, *The use of time-variant eeg granger causality for inspecting directed interdependencies of neural assemblies*, *Journal of Neuroscience Methods* **124** (2003), 27–44.
- [38] C. Hilgetag, R. Kotter, and K. E. Stephan, *Computational methods for the analysis of brain connectivity*, *ascoli operator: network typesetting ed.*, ch. 14-Hilgetag, 2002.

- [39] A. E. Hoerl and R. W. Kennard, *Ridge regression - biased estimation for nonorthogonal problems*, *Technometrics* **12** (1970), 55–and.
- [40] B. Horwitz, *The elusive concept of brain connectivity*, *Neuroimage* **19** (2003), 466–470.
- [41] Y. Hosoya, *The decomposition and measurement of the interdependency between 2nd-order stationary-processes*, *Probability Theory and Related Fields* **88** (1991), 429–444.
- [42] D. R. Hunter, *Mm algorithms for generalized bradley-terry models*, *Annals of Statistics* **32** (2004), 384–406.
- [43] D. R. Hunter and K. Lange, *A tutorial on mm algorithms*, *American Statistician* **58** (2004), 30–37.
- [44] D. R. Hunter and R. Li, *Variable selection using MM algorithms*, *Annals of Statistics* **33** (2005), no. 4, 1617–1642.
- [45] B. Jones and M. West, *Covariance decomposition in multivariate analysis*, <http://ftp.isds.duke.edu/WorkingPapers/04-15.pdf> .) (2005).
- [46] T. P. Jung, S. Makeig, M. J. McKeown, A. J. Bell, T. W. Lee, and T. J. Sejnowski, *Imaging brain dynamics using independent component analysis*, *Proceedings of the Ieee* **89** (2001), 1107–1122.
- [47] M. Kaminski, M. Z. Ding, W. A. Truccolo, and S. L. Bressler, *Evaluating causal relations in neural systems: Granger causality, directed transfer*

- function and statistical assessment of significance*, Biological Cybernetics **85** (2001), 145–157.
- [48] K.V. Mardia J.T. Kent and J.M. Bibby, *Multivariate analysis*, Academic Press, London, San Diego, New York, Boston, Sydney, Tokyo, Toronto, 1979.
- [49] R. Kotter, K. E. Stephan, N. Palomero-Gallagher, S. Geyer, A. Schleicher, and K. Zilles, *Multimodal characterisation of cortical areas by multivariate analyses of receptor binding and connectivity data*, Anatomy and Embryology **204** (2001), 333–350.
- [50] Ch. Leng, Y. Lin, and G. Wahba, *A note on the lasso and related procedures in model selection*, <http://www.stat.wisc.edu/~wahba/ftp1/tr1091rxx.pdf> (2005).
- [51] K. V. Mardia, C. Goodall, E. Redfern, and F. J. Alonso, *The kriged kalman filter - rejoinder*, Test **7** (1998), 277–285.
- [52] Eduardo Martinez-Montes, Pedro A. Valdes-Sosa, Fumikazu Miwakeichi, Robin I. Goldman, and Mark S. Cohen, *Concurrent eeg/fmri analysis by multiway partial least squares*, Neuroimage **22** (2004), 1023–1034.
- [53] A. R. McIntosh and F. Gonzalez-Lima, *Structural equation modeling and its applications to network analysis in functional brain imaging*, Human Brain Mapping **2** (1994), 2–22.

- [54] N. Meinshausen and P. Bühlmann, *Consistent neighbourhood selection for sparse high-dimensional graphs with the lasso*, <http://stat.ethz.ch/research/> (2004).
- [55] F. Miwakeichi, E. Martinez-Montes, P. A. Valdes, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi, *Decomposing eeg data into space-time-frequency components using parallel factor analysis*, *Neuroimage* **22** (2004), 1035–1045.
- [56] A. Neumaier and T. Schneider, *Estimation of parameters and eigenmodes of multivariate autoregressive models*, *Acm Transactions on Mathematical Software* **27** (2001), 27–57.
- [57] R. D. Pascual-Marqui, M. Esslen, K. Kochi, and D. Lehmann, *Functional imaging with low-resolution brain electromagnetic tomography (loreta): A review*, *Methods and Findings in Experimental and Clinical Pharmacology* **24** (2002), 91–95.
- [58] J. Pearl, *Graphs, causality, and structural equation models*, *Sociological Methods and Research* **27** (1998), 226–284.
- [59] ———, *Graphs, causality, and structural equation models*, *Sociological Methods and Research* **27** (1998), 226–284.
- [60] ———, *Causality*, Cambridge University Press, 2000.
- [61] J. O. Ramsay and B. W. Silverman, *Functional data analysis*, Springer, 1997.

- [62] D. S. Ruchkin, E. R. John, and J. Villegas, *Analysis of average evoked potentials making use of least mean square techniques*, Annals of the New York Academy of Sciences **115** (1964), 799–and.
- [63] R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson, *The tetrad project: Constraint based aids to causal model specification*, Multivariate Behavioral Research **33** (1998), 65–117.
- [64] T. Schneider and A. Neumaier, *Algorithm 808: Arfit - a matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models*, Acm Transactions on Mathematical Software **27** (2001), 58–65.
- [65] R. K. Shields, S. Madhavan, K. R. Cole, J. D. Brostad, J. L. DeMeulenaere, C. D. Eggers, and P. H. Otten, *Proprioceptive coordination of movement sequences in humans*, Clinical Neurophysiology **116** (2005), 87–92.
- [66] O. Sporns, D. R. Chialvo, M. Kaiser, and C. C. Hilgetag, *Organization, development and function of complex brain networks*, Trends in Cognitive Sciences **8** (2004), 418–425.
- [67] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, *Sparsity and smoothness via the fused lasso*, Journal of the Royal Statistical Society Series B-Statistical Methodology **67** (2005), 91–108.
- [68] P.A. Valdes, *Quantitative electroencephalographic tomography*, Electroencephalography and Clinical Neurophysiology **103** (1997), 19.

- [69] P. A. Valdes-Sosa, *Spatio-temporal autoregressive models defined over brain manifolds*, Neuroinformatics **2** (2004), 239–250.
- [70] P. A. Valdes-Sosa, J. M. Sanchez-Bornot, A. Lage-Castellanos, M. Vega-Hernandez, J. Bosch-Bayard, L. Melie-Garcia, and E. Canales-Rodriguez, *Estimating brain functional connectivity with sparse multivariate autoregression*, Philosophical Transactions of the Royal Society B-Biological Sciences **360** (2005), 969–981.
- [71] M. West, *On scale mixtures of normal-distributions*, Biometrika **74** (1987), 646–648.
- [72] ———, *Bayesian factor regression models in the “large  $p$ , small  $n$ ” paradigm*, Working Papers of the Institute of Statistics and Decision Science, Duke University (2002).
- [73] C. K. Wikle and N. Cressie, *A dimension-reduced approach to space-time kalman filtering*, Biometrika **86** (1999), 815–829.
- [74] K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, and A. C. Evans, *A unified statistical approach for determining significant signals in images of cerebral activation*, Human Brain Mapping **4** (1996), 58–73.
- [75] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J.R. Statisc. Soc. B **67** (2005), 301–320.

## **Spatio-Temporal Autoregressive Models defined over brain manifolds**

# **Spatio-Temporal Autoregressive Models defined over Brain Manifolds**

Pedro A. Valdes-Sosa<sup>1\*</sup>

<sup>1</sup>Cuban Neuroscience Center, Havana Cuba

\*Corresponding author.

Address: Ave 25 #15202 esquina 158, Cubanacán, Ciudad Habana, Cuba

## ***Keywords:***

Granger Causality, Neuroimages, Bayesian Multivariate Autoregressive  
Model, Spatio-temporal model



## ***Abstract***

Multivariate Autoregressive time series models (MAR) are an increasingly used tool to explore functional connectivity in Neuroimaging. They provide the framework for analyzing the Granger Causality of a given brain region on others. In this paper we shall limit our attention to linear MAR models, in which a set of matrices of autoregressive coefficients  $\mathbf{A}_k$  ( $k=1,\dots,p$ ) describe the dependence of present values of the image on lagged values of its past. Methods for estimating the  $\mathbf{A}_k$  and determining which elements are zero are well known and are the basis for directed measures of influence. However, to date, MAR models are limited in the number of time series they can handle, forcing the a priori selection of a (small) number of voxels or regions of interest for analysis. This ignores the full Spatio-Temporal nature of functional brain data which are in fact collections of time series sampled over an underlying continuous spatial manifold--the brain. A fully Spatio-Temporal MAR model (ST-MAR) is developed within the framework of Functional Data Analysis. For spatial data each row of a matrix  $\mathbf{A}_k$  is the *influence field* of a given voxel. A Bayesian ST-MAR model is specified in which the influence fields for all voxels are required to vary smoothly over space. This requirement is enforced by penalizing the spatial roughness of the influence fields. This roughness is calculated with a discrete version of the spatial Laplacian operator. A massive reduction in dimensionality of computations is achieved via the Singular Value Decomposition, making an interactive exploration of the model feasible. Use of the model is illustrated with an fMRI

time series that was gathered concurrently with EEG in order to analyze the origin of resting brain rhythms.

## ***Introduction***

Devising methods for inferring the effective and functional connectivity of different brain regions is currently a major concern in neuroimaging (Friston, 1994; Buchel & Friston, 2001; Lee, Harrison, & Mechelli, 2003; Buchel & Friston, 2000). The task in hand is to determine the changing patterns of causal influences that different brain structures exert on each other by means of the analysis of dynamical brain imaging data. This type of data include EEG/MEG source distributions (Valdés, Riera, & Casanova, 2000), optical recordings (Schiessl et al., 2000) and fMRI (Harrison, Penny, & Friston, 2003) and are, from the statistical point of view, spatiotemporal data sets (Mardia, Goodall, Redfern, & Alonso, 1998; Wike & Cressie, 1999)— that is vector valued time series where the dimensionality of the vectors is very large having originated from sampling over an underlying continuous manifold.

Ideally, methods for connectivity analysis in the brain should be able to address the full four dimensional spatiotemporal nature of the basic data (Mardia et al., 1998; Stroud, Muller, & Sanso, 2001). Additionally, they should be capable of measuring directed influence  $I_{x \rightarrow y}$  of region  $x$  on region  $y$  (Geweke, 1982; Geweke, 1984). Unfortunately, most methods for analyzing connectivity in Neuroimages fall short of these requirements.

Latent structure models such as PCA (Ruchkin, John, & Villegas, 1964; Friston, Phillips, Chawla, & Buchel, 1999) or ICA (McKeown et al., 1998) analyze the full set of voxels in brain images. They extract subsets of voxels that are

statistically dependent and therefore may serve to identify functionally coupled brain subsystems. These models however are not designed for determining the directional connectivity associated with causal inference. Information on timing of events for example is not used to determine possible causal influences. They may be characterized as having high spatial but no temporal resolution.

Structural Equation Modeling (McIntosh & Gonzalez-Lima, 1994) on the other hand does face up to the issue of inferring directional influences and is firmly grounded in the modern statistical analysis of Causality (Pearl, 1998) via graphical models. Initial studies (McIntosh et al., 1994) in Neuroimaging, being based on non dynamical PET data, ignored temporal information. The concept of Granger Causality (Granger, 1969; Hosoya, 1991) does make use of temporal information in order to establish a measure of directed influence. This measure has been imported from the field of econometrics for use in the analysis of electrophysiological measurements (Baccala & Sameshima, 2001; Freiwald et al., 1999; Hesse, Moller, Arnold, & Schack, 2003; Kaminski, Ding, Truccolo, & Bressler, 2001; Bressler, Ding, & Yang, 1999; Bernasconi & Konig, 1999). Recently Dahlhaus and coworkers (Dahlhaus, Eichler, & Sandkuhler, 1997; Dahlhaus, 2000) have combined the notion of Granger Causality analysis with that of graphical models.

The measurement of Granger Causality analysis is usually based on a Multivariate Autoregressive Model of the data (MAR), be it linear (Penny & Roberts, 2002; Geweke, 1982; Geweke, 1984; Gersch & Yonemoto, 1977) or nonlinear (Freiwald et al., 1999). The recent introduction of linear and bilinear

MAR models for fMRI data (Harrison et al., 2003) has opened the way for measuring linear and nonlinear Granger Causality in this type of data. However, the specific type of modeling used in the cited references only allows only a very limited number of time series to be included in the analysis, resulting in models that have very moderate spatial resolution. This forces the a priori choice of either privileged certain voxels to be analyzed, or alternatively the selection of regions of interest over which average values of activity must be obtained. What is lacking is the development of fully Spatio-Temporal for AR modeling (ST-MAR.

As an concrete and motivational illustration of what has just been said, consider one of the fMRI time series that was gathered concurrently with EEG in order to analyze the origin of resting brain rhythms (Goldman, Stern, Engel, & Cohen, 2001; Goldman, Stern, Engel, & Cohen, 2002; Martinez-Montes, Valdes-Sosa, Miwakeichi, Goldman, & Cohen, 2003). . As described in those papers, significant correlations were found between time-varying spectral components in different EEG bands and the BOLD signal. Figure 1 (left) shows the map of EEG-BOLD correlations for the alpha rhythm. This figure reveals widely distributed anatomical systems that are apparently involved in the generation of this oscillation. These same locations are also identifiable on the basis of the fMRI information alone as shown by further study of the BOLD signal from the voxel with the highest (negative) correlation with alpha power. The correlation map of this voxel with all others was obtained and is shown in the lower left panel of Figure 1. It is interesting to note that this map looks very similar to the one on the left, except for a sign inversion. When faced with this type of data, the question

immediately arises as to which voxel is driving which. But no currently available MAR model can be fit to this amount of data in which the number of time series is much larger than the number of time points.

This paper will propose a ST-MAR precisely for this type of situation. This model is a generalization of the "smoothness priors" approach to MAR introduced by (Kitagawa & Gersch, 1985) but now applied to spatial aspects in the framework of Functional Data Analysis or FDA; (Ramsay & Dalzell, 1991; Ramsay & Silverman, 1997). A fully Spatio-Temporal MAR model (ST-MAR) is developed within the framework of Functional Data Analysis. For spatial data each row of a matrix  $\mathbf{A}_k$  shall be termed the *influence field* of a given voxel. A Bayesian ST-MAR model is specified in which the influence fields for all voxels are required to vary smoothly over space. This requirement is enforced by means of a penalization of spatial roughness of the influence fields calculated with a discrete version of the spatial Laplacian operator. A massive reduction in dimensionality of computations is achieved via the Singular Value Decomposition, making an interactive exploration of the model feasible. The exploratory use of the model is illustrated the data fMRI time series presented above (Goldman et al., 2001; Goldman et al., 2002).

### ***Bayesian Multivariate Autoregressive Model***

In what follows we shall denote vectors with lower case bold letters, matrices with upper case bold letters. We shall also use a general matrix-variate notation for Gaussian and related distributions (i.e. inverse Wishart) which was

introduced by Dawid (Dawid, 1981) to avoid the use of the vectorization (vec) operator and Kronecker products previously necessary for the Bayesian analysis of multivariate regressions (see the cited paper for details).

Let the dynamic neuroimaging data set be considered a vector valued

$$\text{time series: } \mathbf{y}_t = \begin{bmatrix} y_{1;t} \\ \vdots \\ y_{s;t} \\ \vdots \\ y_{Ns;t} \end{bmatrix}_{Ns \times 1}, \text{ where } s = 1, \dots, Ns \text{ indexes the voxels and } t = 1, \dots, Nt$$

the time instants at which samples are gathered. We shall posit a Multivariate Autoregressive Model (MAR) for the  $\mathbf{y}_t$ :

$$\mathbf{y}_t = \sum_{k=1}^p \mathbf{A}_k \mathbf{y}_{t-k} + \mathbf{e}_t \quad (1)$$

where  $p$  denotes the model order, the  $\mathbf{A}_k$  are the matrices of autoregressive coefficients (of dimension  $Ns \times Ns$ ) and  $\mathbf{e}_t$  the model innovations (Geweke, 1982). We shall assume that the innovations are multivariate normal vectors with mean zero and covariance matrix  $\mathbf{V}$ , that is:  $\mathbf{e}_t \sim N(\mathbf{0}, \mathbf{V}_{Ns \times Ns})$ . Note that  $\mathbf{V}$  contains information about instantaneous interactions between voxels whereas the  $\mathbf{A}_k$  are reflecting the directed linear interactions between voxels. In effect the coefficient  $a_{i,j}^{(k)}$  of  $\mathbf{A}_k$  is the (linear) contribution of voxel  $j$  on voxel  $i$  at time lag  $k$ .

For small to moderate  $Ns$  there are a number of estimation procedures for the coefficient matrices  $\mathbf{A}_k$  based on ordinary least squares (Neumaier &

Schneider, 2001; Schneider & Neumaier, 2001; Gersch et al., 1977). However ordinary least squares methods fail when the number of parameters to be estimated is very large relative to the number of observations and when the time series are highly correlated.

A more general approach is the Bayesian framework (Kitagawa et al., 1985; Penny et al., 2002; Harrison et al., 2003), which is adopted in this paper and which shall be immediately explained. Equation (1) may be rewritten as a Multivariate Regression by defining  $\mathbf{A} = [\mathbf{A}_1, \dots, \mathbf{A}_p]$ ,  $\mathbf{Y} = [\mathbf{y}_{p+1}, \dots, \mathbf{y}_{N_t}]_{N_s \times N_t}$  and

$$\mathbf{X} = \begin{bmatrix} \mathbf{y}_1, \dots, \mathbf{y}_{N_t-1} \\ \vdots \\ \mathbf{y}_p, \dots, \mathbf{y}_{N_t-p} \end{bmatrix}_{N_s \cdot p \times N_t}$$

in which case:

$$\mathbf{Y} = \mathbf{A} \mathbf{X} + \mathbf{E} \quad (2)$$

The likelihood of the data  $\mathbf{Y}$  given the parameter set  $\Theta = [\mathbf{A}, \mathbf{V}]$  and the data at previous time lags  $\mathbf{X}$  is denoted by  $P(\mathbf{Y}|\mathbf{A}, \mathbf{V}; \mathbf{X})$  and is determined by the specifications  $\mathbf{Y} = \mathbf{A} \mathbf{X} + \mathbf{E}$  and  $\mathbf{E} \sim N(\mathbf{0}, \mathbf{V}_{N_s \times N_s}, \mathbf{I}_{N_s})$  which lead to the following expression:

$$P(\mathbf{Y}|\mathbf{A}, \mathbf{V}; \mathbf{X}) = \frac{1}{|2\pi\mathbf{V}|^{\frac{N_t-p}{2}}} \exp \left[ -\frac{1}{2} \text{tr} \left( \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{A} \mathbf{X}) (\mathbf{Y} - \mathbf{A} \mathbf{X})^T \right) \right] \quad (3)$$

In the non Bayesian approach the estimators for  $\Theta = [\mathbf{A}, \mathbf{V}]$  are obtained by maximizing(3). The Bayesian formulation of the MAR also posits an a priori



distribution for  $\Theta$ ,  $P(\Theta|\mathbf{X})$  which encapsulates our prior knowledge about the autoregressive coefficients and the innovation covariance matrix. A specification that is widely used, the conjugate Normal-Inverse Wishart formulation (Minka, 2000) which is achieved by selecting  $\mathbf{A} \sim N(\mathbf{0}, \mathbf{V}, \alpha \Sigma)$  and  $\mathbf{V} \sim W^{-1}(\beta \Sigma, N_o)$  where  $\Sigma, \alpha, \beta$  are hyper parameters that also need to be specified. The a priori distributions for the autoregressive coefficients and the innovation covariance are expressed in terms of a notation for the matrix-variate probability densities as described by Dawid (Dawid, 1981). For example, the requirement that  $\mathbf{A} \sim N(\mathbf{0}, \mathbf{V}, \alpha \Sigma)$  is equivalent to saying

that  $\mathbf{A} = \underset{(Ns \times Ns)}{\mathbf{C}^t} \underset{(Ns \times Ns)}{\mathbf{U}} \underset{(Ns \times Ns)}{\mathbf{B}}$ , where  $\mathbf{U}$  are i.i.d. normal random variates and

$\mathbf{V} = \mathbf{A}^t \mathbf{A}$  and  $\Sigma = \mathbf{B}^t \mathbf{B}$ . In other words  $\Sigma$  is strongly related to the a priori covariances of the rows of the autoregressive coefficient matrices. (Minka, 2000): Our formulation differs from that of (Penny et al., 2002) in that these authors specify a "vague" prior for  $\mathbf{V}$ . For relations between the two approaches see Minka (Minka, 2000). To summarize, the a priori density is:

$$P(\mathbf{A}, \mathbf{V} | \mathbf{X}) = N(\mathbf{A} | \mathbf{0}, \mathbf{V}, \alpha \Sigma) W^{-1}(\mathbf{V} | \beta \Sigma, N_o) \quad (4)$$

Combining (3) and (4) we have that the posterior distribution is (see (Minka, 2000):

$$P(\mathbf{A}, \mathbf{V} | \mathbf{Y}, \mathbf{X}) = NW^{-1}\left(\mathbf{0}, \mathbf{S}_{yx} \mathbf{S}_{xx}^{-1}, \mathbf{S}_{xx}, \mathbf{S}_{y|x} + \beta \Sigma, \bar{N} + N_o\right) \quad (5)$$

where  $\mathbf{S}_{xx} = \mathbf{X}\mathbf{X}^T + \alpha^{-1}\boldsymbol{\Sigma}^{-1}$ ,  $\mathbf{S}_{yx} = \mathbf{Y}\mathbf{X}^T$ ,  $\mathbf{S}_{yy} = \mathbf{Y}\mathbf{Y}^T$ ,  $\mathbf{S}_{y|x} = \mathbf{S}_{yy} - \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{yx}^T$ . It

is to be noted that the normal- inverse Wishart conjugate distribution specification guarantees that the a posteriori distribution belongs to the same family as the likelihood and the a priori distribution.

From (5) the Bayesian MAR maximum a posteriori (MAP) estimators are easily obtained in closed form and are equal to:

$$\begin{aligned}\hat{\mathbf{A}} &= \mathbf{S}_{yx}\mathbf{S}_{xx}^{-1} \\ \hat{\mathbf{V}} &= \frac{\mathbf{S}_{y|x} + \beta\boldsymbol{\Sigma}}{\bar{N} + N_o}\end{aligned}\quad (6)$$

The Bayesian formulation just described depends critically on the choice of  $\boldsymbol{\Sigma}$ . Previous specifications on this matrix have been concerned mainly with regularizing temporal properties of MAR. Kitagawa et al. (Kitagawa et al., 1985) crafted  $\boldsymbol{\Sigma}$  to request smoothness of MAR coefficients in either the time or frequency domain. This was the same concern of (Penny et al., 2002; Harrison et al., 2003), a main objective being avoidance of over fitting the model by increasingly penalizing higher lag coefficients

### ***Spatio Temporal MAR***

In Neuroimaging the actual underlying model is:

$$y(s, t) = \sum_{k=1}^p \int a_k(s, s') y(s', t-k) ds' + e(s, t) \quad (7)$$

where integration is over brain manifolds. Equation (1) is now a discretized version of (7). In our concrete example  $N_s=12,642$  since we are dealing with a

large part of the fMRI image volume. That is, at each instant of time observation vector comprises all signals sampled over the points in the grid shown in Figure 2. This grid is the intersection of the MRI sampling grid with segmented brain tissue (valid brain voxels). However, in principle, the amount of data points could increase to infinity with improved MRI techniques. Not only is the spatial dimension massive but it is to be expected that nearby spatial points will be highly correlated.

In order to overcome these problems and develop a ST-MAR that is valid no matter how fine the spatial sampling rate, we propose to regularize the discrete Equation (1) by the use of a roughness penalty approach to enforce a degree of smoothness of the autoregressive coefficients. To be more specific consider  $j$ -eth rows of the different  $\mathbf{A}_k$  matrices. Each is defined over all valid brain voxels and is shall be termed, for each lag  $k$ , the **k lag influence field**. Our model specification will impose smoothness over the influence fields for all voxels and lags by defining the covariance matrix  $\Sigma = \mathbf{L}^{-2}$ .  $\mathbf{L}$  denotes a discrete version of the spatial Laplacian operator and is the square root of a roughness penalty matrix (Ramsay et al., 1997) that will punish spatial roughness of the rows of all influence fields. It is defined over the grid of valid brain voxels (Figure 2) and is encoded via a MATLAB sparse matrix (Figure 3). The definition of the roughness penalty is:

$$\mathbf{L} = \left\{ l_{i,j} \right\}_{1 \leq i,j \leq N_s} \begin{cases} l_{i,i} = 6 \\ l_{i,j} = -1 & \text{iff } i \text{ and } j \text{ are neighbors} \\ l_{i,j} = 0 & \text{iff } i \text{ and } j \text{ are NOT neighbors} \end{cases} \quad (8)$$

Note that we have chosen  $\mathbf{L}$  to be symmetric, of full rank and well conditioned.

A system for the exploratory analysis of ST-MAR has been written in MATLAB (©Mathworks). Though at first sight the evaluation of the estimators (6) would seem to be a daunting task, it has been possible to achieve an interactive system that can operate in real time. Some of the technical solutions employed to make this possible are now detailed.

1.-Due to the form of the penalty matrix it is possible to transform the multivariate regression problem (2) to the case where  $\Sigma = \mathbf{I}$  in which case this problem takes on the form of a standard Tikhonov regularization, for which there has been considerable work and software developed (Hansen, 1999). In particular, in this context, an easily evaluated criterion Generalized Cross Validation (GCV) allows a computationally inexpensive automatic selection of the hyper parameters  $p$ ,  $\alpha$  and  $\beta$ . (Golub, Heath, & Wahba, 1979).

2.-Massive dimensionality reduction is possible by means of the Singular Value Decomposition (SVD) of  $\mathbf{X}$ . The multivariate regression problem may then be transformed so the computational complexity is then determined not by  $N_s$  but by  $N_t$ . Our approach to analyzing a data matrix where the number of variables is much larger than the number of data points is similar but not identical with that described by West (West, 2002).

3.-Calculation of each the influence fields for each lag and voxel is carried out on demand. This avoids storing the full set of autoregressive matrices.

Instead two matrices of dimension  $N_s \times N_t \times p$  suffice to reconstruct any given influence field for any lag.

Instead of trying to determine if individual autoregressive coefficients are zero the influence fields will be treated as Neuroimages. The current implementation, Statistical Parametric Mapping of the influence fields is currently carried out by thresholding an approximate t statistic image derived for each influence field. Thresholds are set according to Random Field Theory (Worsley et al., 1996). The t image is based on the Jackknife . (Efron, 1986): Leave one out samples are created using and jackknife pseudo values are obtained which are then used to calculate the t image. The usual bipolar scales are used for these images in order to depict both positive and negative significant areas in the t image.

### ***Application to the fMRI data set***

An example of an exploratory analysis with the ST-MAR model is now provided. The data analyzed is the BOLD signals from the concurrent EEG/fMRI data that has already been described in the Introduction. Information on voxel interrelationships is shown in Figure 1.

The EEG was sampled at 200 Hz from an array of 16 bipolar pairs, with an additional channel for the EKG and scan trigger. The fMRI time series was measured in 6 slice planes (4 mm, skip 1 mm) parallel to the AC-PC line, with the second from the bottom slice through AC-PC. More details about this data set can be found in Goldman et al. (Goldman et al., 2002). Informed consent was

obtained from the volunteer based on a protocol approved previously by the UCLA Office for the Protection of Research Subjects. As mentioned previously  $N_s=12,642$  and  $N_t=108$ .

For this data the GCV criterion indicated that the most appropriate model order for the ST-MAR model was  $p=1$  in equation (1). According to the description in the previous section, the user interactively can select a voxel and observe the SPM for the jackknifed t image. In the lower right hand panel of Figure 4 an arrow indicates a voxel in the thalamus for which the influence field was then calculated. The thresholded t image (for global significance level of 0.05) is shown in Figure 4. It should be noticed that for this particular point significant positive influences are concentrated around the thalamus and midline. A large negative influence of the thalamic voxel selected is found in frontal regions.

## ***Discussion***

The field of Neuroimaging provides data that in principle is actually defined over an underlying spatial continuum. Thus more accurate sampling will only increase the number of highly correlated variables that are measured on a always insufficient number of subjects and conditions. The need for spatial regularization of some sort is a recurring problem in the statistics of Human Brain Mapping (Purdon, Solo, Weisskoff, & Brown, 2001; Kustra & Strother, 2001). This spatial regularization is the basic tenet of FDA (Ramsay et al., 1997). To our knowledge this is the first attempt to use the FDA approach to obtain a MAR

defined over a spatial domain. The methods are quite general have been implemented in real time by carrying out computation in a reduced dimensional space. The Bayesian formulation allows other a priori knowledge to be integrated into the model in a principled way. The procedure proposed concentrates on the examination of influence fields for given voxel and lags and therefore brings at least part of the information necessary for the evaluation of Granger Causality into the domain of well known methods for Neuroimaging statistics.

The use of a conjugate normal-inverse Wishart prior was adopted in order to obtain closed solutions for the estimators for the influence fields. This choice is asymmetric. If it is reasonable to require that the influence of a voxel other points in the brain be similar if those points are near, then the converse is also valid. That is that influences on a given voxel from two points that are nearby be also similar. This a requirement on the smoothness of the *columns* of the matrices  $A_k$  instead of on the rows as enforced in this paper. Additionally, the a priori modeling of the covariance matrix deserves more attention. In either case, different choices than those of this paper will lead to non conjugate priors, a direction in which Bayesian regression has already gone (Brown, Fearn, & Vannucci, 1999). An additional point that might require modification is the use of the GCV criterion for hyper parameter selection. A more consistent approach might be to use the Bayesian evidence for this purpose as in (Penny et al., 2002; Harrison et al., 2003).

The method proposed is intended for exploratory analysis only and thus is of use only when there are external criteria to guide the selection of the voxels for

which influence fields would want to be determined. A fully automatic search for influence fields is however quite possible and in fact is just another example of the variable selection problem in multiple regression (Brown et al., 1999; Brown, Vannucci, & Fearn, 2002).

A number of extensions of this approach are possible and probably necessary; the Bayesian formulation will accommodate all these. We mention some examples just to illustrate the possibilities. If the prior distributions are selected properly, slow changes of the  $\mathbf{A}_k$  over time may be modeled, extending this work to non-stationary processes, a Spatio-Temporal analog of (Hesse et al., 2003). In a similar fashion allowing the  $\mathbf{A}_k$  to be a smooth functions of the previous state of the system would accommodate the modeling of nonlinear dynamical systems, a Spatio-Temporal analog of (Freiwald et al., 1999). The use of anatomical constraints may be easily introduced, be it to limit the voxels for which influence fields are estimated or to establish a priori constraints based on fiber tract information.

Finally a major challenge must be mentioned. The motivating example in the introduction is from a concurrent EEG/fMRI experiment in which the EEG and BOLD time series in all truth live in totally different temporal scales. Developing concepts for multiscale Spatio-Temporal Granger causality would allow multimodal image fusion for connectivity evaluation.



## ***Acknowledgments***

I wish to thank Rolando Biscay-Lirio for many fruitful discussions that were essential for the development of the ideas expressed in this paper. Jose Miguel Bornot and Ernesto Palmero helped with the preparation of the programs and with the data processing. Eduardo Aubert modified his Neuroimaging visualization system to include the interactive selection of voxels to interrogate for causal effects and to production the final graphical outputs. I also wish to thank Mark Cohen and Robin Goldman for the use of the fMRI data set used in this paper and for their collaboration in general.

## ***References***

Baccala, L. A. & Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. *Biological Cybernetics*, 84, 463-474.

Bernasconi, C. & Konig, P. (1999). On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings. *Biological Cybernetics*, 81, 199-210.

Bressler, S. L., Ding, M. Z., & Yang, W. M. (1999). Investigation of cooperative cortical dynamics by multivariate autoregressive modeling of event-related local field potentials. *Neurocomputing*, 26-7, 625-631.

Brown, P. J., Fearn, T., & Vannucci, M. (1999). The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach. *Biometrika*, 86, 635-648.

Brown, P. J., Vannucci, M., & Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 64, 519-536.

Buchel, C. & Friston, K. (2000). Assessing interactions among neuronal systems using functional neuroimaging. *Neural Networks*, 13, 871-882.

Buchel, C. & Friston, K. (2001). Interactions among neuronal systems assessed with functional neuroimaging. *Revue Neurologique*, 157, 807-815.

Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, 51, 157-172.

Dahlhaus, R., Eichler, M., & Sandkuhler, J. (1997). Identification of synaptic connections in neural ensembles by graphical models. *Journal of Neuroscience Methods*, 77, 93-107.

Dawid, A. P. (1981). Some Matrix-Variate Distribution-Theory - Notational Considerations and A Bayesian Application. *Biometrika*, 68, 265-274.

Efron, B. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression-Analysis - Discussion. *Annals of Statistics*, 14, 1301-1304.

Freiwald, W. A., Valdes, P., Bosch, J., Biscay, R., Jimenez, J. C., Rodriguez, L. M. et al. (1999). Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. *Journal of Neuroscience Methods*, 94, 105-119.

Friston, K., Phillips, J., Chawla, D., & Buchel, C. (1999). Revealing interactions among brain systems with nonlinear PCA. *Human Brain Mapping*, 8, 92-97.

Friston, K. J. (1994). Functional and Effective Connectivity in Neuroimaging: a Synthesis. *Human Brain Mapping*, 2, 56-78.

Gersch, W. & Yonemoto, J. (1977). Parametric Time-Series Models for Multivariate Eeg Analysis. *Computers and Biomedical Research*, 10, 113-125.

Geweke, J. (1982). Measurement of Linear-Dependence and Feedback Between Multiple Time-Series. *Journal of the American Statistical Association*, 77, 304-313.

Geweke, J. F. (1984). Measures of Conditional Linear-Dependence and Feedback Between Time-Series. *Journal of the American Statistical Association*, 79, 907-915.

Goldman, R., Stern, J., Engel, J., & Cohen, M. (2001). Tomographic mapping of alpha rhythm using simultaneous EEG/fMRI. *Neuroimage*, 13, S1291.

Goldman, R. I., Stern, J. M., Engel, J., & Cohen, M. S. (2002). Simultaneous EEG and fMRI of the alpha rhythm. *Neuroreport*, 13, 2487-2492.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized Cross-Validation As A Method for Choosing A Good Ridge Parameter. *Technometrics*, 21, 215-223.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross- Spectral Methods. *Econometrica*, 37, 414-&.

Hansen, P. C. (1999). Regularization tools Version 3.0 for Matlab 5.2. *Numerical Algorithms*, 20, 195-196.

Harrison, L., Penny, W. D., & Friston, K. (2003). Multivariate autoregressive modeling of fMRI time series. *Neuroimage*, 19, 1477-1491.

Hesse, W., Moller, E., Arnold, M., & Schack, B. (2003). The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies. *Journal of Neuroscience Methods*, 124, 27-44.

Hosoya, Y. (1991). The Decomposition and Measurement of the Interdependency Between 2Nd-Order Stationary-Processes. *Probability Theory and Related Fields*, 88, 429-444.

Kaminski, M., Ding, M. Z., Truccolo, W. A., & Bressler, S. L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics*, 85, 145-157.

Kitagawa, G. & Gersch, W. (1985). A Smoothness Priors Time-Varying Ar Coefficient Modeling of Nonstationary Covariance Time-Series. *IEEE Transactions on Automatic Control*, 30, 48-56.

Kustra, R. & Strother, S. (2001). Penalized discriminant analysis of [O-15]-water PET brain images with prediction error selection of smoothness and

regularization hyperparameters. *Ieee Transactions on Medical Imaging*, 20, 376-387.

Lee, L., Harrison, L. M., & Mechelli, A. (2003). A report of the functional connectivity workshop, Dusseldorf 2002. *Neuroimage*, 19, 457-465.

Mardia, K. V., Goodall, C., Redfern, E., & Alonso, F. J. (1998). The Kriged Kalman filter - Rejoinder. *Test*, 7, 277-285.

Martinez-Montes, E., Valdes-Sosa, P., Miwakeichi, F., Goldman, R., & Cohen, M. Concurrent EEG/fMRI Analysis by Multi-way Partial Least Squares. *Neuroimage*, (in press).

McIntosh, A. R. & Gonzalez-Lima, F. (1994). Structural Equation Modeling and its Applications to Network Analysis in Functional Brain Imaging. *Human Brain Mapping*, 2, 2-22.

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T. P., Kindermann, S. S., Bell, A. J. et al. (1998). Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6, 160-188.

Minka, T. (2000). Bayesian linear regression.  
<http://www.stat.cmu.edu/~minka/papers/learning.html> [On-line]. Available:  
<http://web.media.mit.edu/~tpminka/papers/linear.html>

Neumaier, A. & Schneider, T. (2001). Estimation of parameters and eigenmodes of multivariate autoregressive models. *Acm Transactions on Mathematical Software*, 27, 27-57.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27, 226-284.

Penny, W. D. & Roberts, S. J. (2002). Bayesian multivariate autoregressive models with structured priors. In *Proceedings-Vision Image and Signal Processing*, 149, 33-41.

Purdon, P. L., Solo, V., Weisskoff, R. M., & Brown, E. N. (2001). Locally regularized spatiotemporal modeling and model comparison for functional MRI. *Neuroimage*, 14, 912-923.

Ramsay, J. O. & Dalzell, C. J. (1991). Some Tools for Functional Data-Analysis. *Journal of the Royal Statistical Society Series B- Methodological*, 53, 539-572.

Ramsay, J. O. & Silverman, B. W. (1997). *Functional Data Analysis*. Springer.

Ruchkin, D. S., John, E. R., & Villegas, J. (1964). Analysis of Average Evoked Potentials Making Use of Least Mean Square Techniques. *Annals of the New York Academy of Sciences*, 115, 799-&.

Schiessl, I., Stetter, M., Mayhew, J. E. W., McLoughlin, N., Lund, J. S., & Obermayer, K. (2000). Blind signal separation from optical imaging recordings with extended spatial decorrelation. *IEEE Transactions on Biomedical Engineering*, 47, 573-577.

Schneider, T. & Neumaier, A. (2001). Algorithm 808: ARfit - A matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *Acm Transactions on Mathematical Software*, 27, 58-65.

Stroud, J. R., Muller, P., & Sanso, B. (2001). Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 63, 673-689.

Valdés, P., Riera, J., & Casanova, R. (2000). Spatio Temporal Distributed Inverse Solutions. In Aine C.J., Y. Okada, G. Stroink, S. J. Swithenby, & C. C. Wood (Eds.), *Biomag 96: Proceedings of the Tenth International Conference on Biomagnetism*. ( Springer Verlag.

West, M. (2002). Bayesian Factor Regression Models in the "Large p, Small n" Paradigm. Working Papers of the Institute of Statistics and Decision Science, Duke University [On-line].

Wikle, C. K. & Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, 86, 815-829.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, 4, 58-73.

## ***Figure Legends***

### **Figure 1**

Correlation Maps for fMRI time series obtained with concurrent EEG recording. Left side: Correlations between power fluctuations in the alpha range and BOLD signals at each voxel. Only those correlations calculated at voxels in segmented brain tissue are shown. Right side: Correlations between BOLD signal in all brain voxels and the signal at the voxel having the highest correlation with alpha power fluctuations (minimum correlation in left side figure). Both figures are thresholded for an overall significance level of  $p < 0.05$ .

### **Figure 2**

Grid for which analyses were carried out in this paper. In particular a discrete version of the spatial Laplacian operator was obtained from this grid (Figure 3) by means of expression (8) in the text.

### **Figure 3**

Graphical display of the Laplacian matrix used as a roughness penalty to enforce spatial smoothness of the influence fields (rows of autoregressive coefficients) in the ST-MAR model. This is the default MATLAB display of sparse matrices where each entry in the matrix which is zero is shown as a black point and every non zero element is displayed as a white point. Note the extreme sparseness of the matrix which facilitates efficient computation.



## Figure 4

Statistical Parametric Map of the Jackknife t image of the influence field for a point in the thalamus. This point is marked by an arrow in the lower right panel. The optimal model only had one time lag. The t image was calculated from Jackknife pseudo values and is thresholded for an overall significance level of  $p < 0.05$ .

**Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex**

# Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex

Winrich A. Freiwald<sup>a,c,\*</sup>, Pedro Valdes<sup>b</sup>, Jorge Bosch<sup>b</sup>, Rolando Biscay<sup>b</sup>,  
Juan Carlos Jimenez<sup>b</sup>, Luis Manuel Rodriguez<sup>b</sup>, Valia Rodriguez<sup>b</sup>,  
Andreas K. Kreiter<sup>a</sup>, Wolf Singer<sup>c</sup>

<sup>a</sup> Institute for Brain Research, University of Bremen, FB2, P.O. Box 330440, D-28334 Bremen, Germany

<sup>b</sup> Cuban Neuroscience Center, Ave 25 No. 5202 esquina 158 Cubanacán, P.O. Box 6880, 6990 Ciudad Habana, Cuba

<sup>c</sup> Max-Planck-Institute for Brain Research, Deutschordenstr. 46, D-60528 Frankfurt/Main, Germany

Received 2 July 1999; accepted 9 August 1999

## Abstract

Information processing in the visual cortex depends on complex and context sensitive patterns of interactions between neuronal groups in many different cortical areas. Methods used to date for disentangling this functional connectivity presuppose either linearity or instantaneous interactions, assumptions that are not necessarily valid. In this paper a general framework that encompasses both linear and non-linear modelling of neurophysiological time series data by means of Local Linear Non-linear Autoregressive models (LLNAR) is described. Within this framework a new test for non-linearity of time series and for non-linearity of directedness of neural interactions based on LLNAR is presented. These tests assess the relative goodness of fit of linear versus non-linear models via the bootstrap technique. Additionally, a generalised definition of Granger causality is presented based on LLNAR that is valid for both linear and non-linear systems. Finally, the use of LLNAR for measuring non-linearity and directional influences is illustrated using artificial data, reference data as well as local field potentials (LFPs) from macaque area TE. LFP data is well described by the linear variant of LLNAR. Models of this sort, including lagged values of the preceding 25 to 60 ms, revealed the existence of both uni- and bi-directional influences between recording sites. © 1999 Elsevier Science B.V. All rights reserved.

**Keywords:** Non-linear dynamics; Granger causality; Multivariate non-linear autoregression; Bootstrap test for non-linear time series; Local field potential; Inferotemporal cortex

## 1. Introduction

Visual information processing in the mammalian brain is based on a multitude of cortical and subcortical structures. Within the macaque cortex more than thirty visual areas have been described (Felleman and van Essen, 1991), a number likely to be paralleled in other higher mammals, including humans. Neuroanatomical, and electrophysiological evidence suggests, that these cortical areas are further subdivided into anatomical compartments composed of neurons with distinct phys-

iological properties (Kaas and Krubitzer, 1991). Thus, multiple neuronal populations in different areas process different aspects of a visual stimulus. Since receptive fields of cortical cells usually behave like broadly tuned filters in a high dimensional feature space (Martin, 1994; van Essen et al., 1992), a given stimulus, which has different features like spatial position in the visual field, velocity, disparity, colour and form cues, will activate large neural populations within the same and in different cortical areas. These distributed responses have to be integrated into a coherent representation. The establishment of this representation requires extensive interactions between different neuronal populations within the same and in different cortical areas, since there is no final integration area in the brain onto which all processing pathways would converge.

\* Corresponding author. Tel.: +49-421-2189481; fax: +49-421-2189004.

E-mail address: freiwald@brain.uni-bremen.de (W.A. Freiwald)

The structural properties of cortical networks support such extensive interactions. Connections between cortical neurons are generally characterised by a high degree of divergence and convergence. Each cortical area is sending output connections to and is receiving input connections from several other cortical areas. These connections are so numerous that about one third of all possible connections between visual areas have been discovered and roughly one half of them are expected to exist (Felleman and van Essen, 1991). Based on these connectivity patterns between cortical areas, their strength, the spatial arrangement of areas and the relatedness of their functional properties, different schemes for their arrangement into processing pathways have been proposed (Ungerleider and Mishkin, 1982; Felleman and van Essen, 1991; Goodale and Milner, 1992; Scannell et al., 1995; Hilgetag et al., 1996). These pathways are characterised by extensive feedback connections, lateral connections to areas at the same processing level and connections by-passing intermediate levels of the hierarchy (see, e.g. Rockland and van Hoesen (1994)). Recent physiological data show that feedback projections can exert substantial effects onto earlier processing stages (Hupé et al., 1998). Large temporal overlap of the response periods of neurons even in areas at very different levels of the processing hierarchy (Nowak and Bullier, 1997) further support mutual influences. Thus, current neuroanatomical and neurophysiological evidence suggests extensive mutual interactions between distributed groups of neurons.

Despite these results, the mechanisms which serve to integrate the activities of different neurons into a coherent representation are still a much debated issue. A recent concept of information processing in the cortex, extending Hebb's cell assembly concept (Hebb, 1949), stresses the importance of the relative timing of action potentials to express relatedness of responses (von der Malsburg, 1981; Singer et al., 1990). According to this temporal binding hypothesis, neurons belonging to the same assembly should synchronise their responses, while cells belonging to different assemblies should fire asynchronously. Indeed, many experimental findings in the visual cortex are in agreement with this theory, including the existence and stimulus dependency of inter- and intra-areal synchronisation (see, e.g. Singer and Gray (1995); Engel et al., (1997) for recent reviews). According to this conceptual framework as well as to similar ones (Johannesma et al., 1986; Gerstein et al., 1989; Abeles, 1991; Aertsen et al., 1991; Sporns et al., 1991; Ahissar et al., 1992; Prut et al., 1998), neural interactions change in relation to current processing requirements defined by external stimuli and the internal behavioural state of the animal. Much previous work related to these concepts has focused on the strength of neural interactions as indicated by correla-

tion measures. Less attention has been paid to the direction of these interactions as a further dynamic property of functional connectivity. However, in our opinion, directed influences (or causal relations) that individual neurons (Abeles, 1982; Gerstein and Aertsen, 1985) or larger neural groups exert on each other and their variation in time might be of prime importance for cortical information processing. This idea seems to follow naturally from considerations of visual perception. Recent psychophysical research provided evidence that the perception even of the most elementary aspects of a visual scene may depend on factors like attention, past experience, or the segmentation of the visual scene into different objects (Braddick, 1996). In these situations top-down processing should be more prominent than in other instances, e.g. in the case of rapid processing, in which the system might essentially operate in a feed-forward manner (Thorpe et al., 1996). Accordingly, the relative influence of a 'higher level' neural group on a second, 'lower level' one might be stronger in the former condition than in the latter. Even during the response to a stimulus, the pattern of these relative influences between individual neurons or ensembles of neurons might change over time. Thus, in analysing information processing in the visual system, there is a strong interest to study the interactions of neuronal groups, i.e. to infer the direction of these influences from simultaneous electrophysiological recordings.

To define this problem formally, let us denote by  $x_t, y_t$  the values of electrical recordings at time  $t$  obtained from any of two sites. Let us also denote the vector of observations from both sites at time  $t$  as  $\mathbf{z}_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix}$ . Henceforth we shall use lower boldface type to indicate vectors and upper boldface type to indicate matrices. With this notation in place, our problem can then be formalised as defining a measure  $I(y \rightarrow x)$  which will quantify the influence of time series  $y_t$  on time series  $x_t$ .

### 1.1. First generation influence measures: linear instantaneous influences

A first generation of methods (Gerstein et al., 1978) assessed neural interactions by means of correlation methods, based on the use of linear regression. There have been many recent papers along these lines in the neuro-imaging literature, specific instances being path analysis (McIntosh and Gonzalez-Lima, 1994), partial least squares (McIntosh et al., 1996) and the general concept of 'functional connectivity' (Friston, 1994). These methods are based upon two implicit assumptions:

1. Interactions between neural ensembles are linear.
2. Interactions between neural ensembles are instantaneous, that is they depend only on the current state of the system.

Both of these assumptions may be summarised by the following equation:

$$\begin{aligned}x_t &= ay_t + \zeta_t \\y_t &= bx_t + \zeta_t\end{aligned}\quad (1)$$

where the coefficients  $a$  and  $b$  express the linear instantaneous relationships between series  $y$  and series  $x$ . These may be written in matrix notation as:

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_t + \varepsilon_t \quad (2)$$

where we have used the following notation:

$$\mathbf{z}_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix} \quad \varepsilon_t = \begin{bmatrix} \zeta_t \\ \zeta_t \end{bmatrix}$$

The time series  $\varepsilon_t$  is known as the ‘innovation’ and can be viewed alternatively as a white noise process driving the system or as the error of prediction of one time series given the other. Note that all relations involve only the current time  $t$ ; this is what is meant by instantaneous interactions. First generation influence measures are defined as association coefficients that quantify how much of the total variation of the time series are explained by instantaneous linear relationships.

### 1.2. Second generation influence measures: Granger causality

The second assumption of first generation influence measures, that of instantaneous neural interactions, is clearly not realistic since it ignores:

- The delay of transmission of information from one neural site to another.
- The fact that the evolution of the system may depend not only on the immediate past as is evidenced by the rich temporal structure of neural time series. Therefore, more realistic signal models substitute Assumption 2 above by:

3. The evolution of the state of the system may be described as a function of a finite number of past states.

On the basis of this assumption, Eq. (2) may be generalised by stating a dependence of  $\mathbf{z}_t$  not only on its own value, but also upon a set of  $p$  past vectors. (We will refer to  $p$  as ‘the number of lagged values’ in the remainder of this paper, and accordingly use ‘lagged values’ or ‘lagged vectors’.) These can be stacked into ‘delay matrices’  $\mathbf{Y}_t = [y_t, y_{t-1}, \dots, y_{t-k}, \dots, y_{t-p}]$ ,  $\mathbf{X}_t = [x_t, x_{t-1}, \dots, x_{t-k}, \dots, x_{t-p}]$ , and  $\mathbf{Z}_t = \begin{bmatrix} \mathbf{X}_t \\ \mathbf{Y}_t \end{bmatrix}$ , which contain all the information of both time series  $p$  points into the past.

The most frequently used linear model is the Multivariate Linear Autoregressive model:

$$\mathbf{z}_t = \sum_{k=1}^p \mathbf{A}_k \cdot \mathbf{z}_{t-k} + \varepsilon_t \quad (3)$$

Based on the Multivariate Linear Autoregressive model a second generation of measures of influence has been proposed (Gersch, 1970, 1972; Akaike, 1974; Franaszczuk et al., 1985). These not only take into account the correlation structure within and between the observed time series, but they also allow use of the ‘arrow of time’ to devise influence measures to statistically assess causality as introduced by Granger and co-workers (Granger, 1963, 1969, 1980; Granger and Lin, 1995). Granger reasoned thus: if time series  $x_t$  is influencing  $y_t$ , then adding past values of  $x_t$  to the regression of  $y_t$  will improve its prediction. This principle was originally formulated in a very general way, encompassing both linear and non-linear systems. However, Granger pointed out the difficulty of using non-linear models (Granger and Newbold, 1977) by stating: ‘Thus for purely pragmatic reasons, the ‘optimal prediction’ ... should be replaced by ‘optimal linear prediction’ (p. 226). Almost all specific measures of Granger causality have therefore been based on linear models.

To be specific, consider the prediction of  $x_t$  based only on its own past:

$$x_t = \sum_{k=1}^p a_k x_{t-k} + \zeta_t \quad (4)$$

In this case the innovation series  $\zeta_t$  will have a variance  $\sigma_{x|X}^2$  where the suffix indicates that the error variance is that of series  $x_t$  predicted only by its own delay matrix  $\mathbf{X}_t$ . Now consider the following model, which adds the past values of  $y_t$  as predictors of  $x_t$ :

$$x_t = \sum_{k=1}^p a_k x_{t-k} + \sum_{q=1}^r b_k y_{t-q} + \phi_t \quad (5)$$

Note that the number of delays of series  $y_t$  used to predict series  $x_t$  is  $r$  and thus does not have to be equal to  $p$ . In this case the prediction error is now  $\phi_t$  which will have a variance  $\sigma_{x|Z}^2$  where the suffix indicates that series  $x_t$  is now predicted by the complete delay matrix  $\mathbf{Z}_t$  which includes the past of both series. Based on these definitions, Granger introduced the following (linear) influence measure (Granger, 1969):

$$I^{\text{LIN}}(y \rightarrow x) = \ln \left( \frac{\sigma_{x|X}^2}{\sigma_{x|Z}^2} \right) \quad (6)$$

Note that this measure of influence has the right properties. If the past of series  $y_t$  does not improve the prediction of series  $x_t$  then  $\sigma_{x|Z}^2$  will be equal to  $\sigma_{x|X}^2$  and the influence measure will be zero. Any improvement in prediction leads to a decrease in the denominator in Eq. (6) and therefore increases the value of the influence measure. A symmetrical definition of the influence of series  $x_t$  on  $y_t$  is possible. In fact, Geweke and others have generalised these definitions to multivariate time series and have defined influence measures between two sets of time series conditional on a third set of time series (Geweke, 1982, 1984).

Bernasconi and co-workers (Bernasconi et al., 1998; Bernasconi and König, 1999) have applied this type of influence measure to electrophysiological recordings. These authors also carried out a spectral decomposition of the causality measures and provided empirical confidence intervals for this spectrum by means of the bootstrap. The results obtained indicate that influence measures are indeed a useful tool for studying neural effective connectivity.

### 1.3. Third generation influence measures: non-linear Granger causality

The above work presupposes that neural systems are linear. Neither the Hodgkin and Huxley equations of single cell neurophysiology, nor the modelling of synaptic interactions result in linear equations. Whether the ensemble behaviour of neural masses scales to a linear approximation is a matter of great importance to be determined empirically. The recent trend in signal modelling in neuroscience has been quite in the opposite direction to linear modelling. Results obtained with analytical methods derived from 'chaos theory' (see Elbert et al. (1994) for a review) have provided evidence for the essentially non-linear nature of large-scale EEG-, ECoG- and MEG-signals, even though the existence of underlying chaotic dynamics may not be demonstrable (Valdes et al., 1999). If the time series are non-linear then methods based on linear regression as those described above may be misleading.

There have been several previous attempts to generalise the first generation influence measures to the non-linear case as exemplified in the work of Lopes da Silva and Mars (1987) by using both information theory concepts (Pijn et al., 1990) and correlation concepts based on non-linear regression. Both works demonstrated that in specific instances the assumption of linear interactions was misleading and that rather than the relationship expressed in Eq. (2) the following expression should be used:

$$\mathbf{z}_t = \mathbf{F}(\mathbf{z}_t) + \varepsilon_t \quad (7)$$

where  $\mathbf{F}$  is a non-linear relationship. This model and the aforementioned measures based on it suffer from the shortcomings of all first generation methods enumerated in the previous section.

A third generation of influence measures is obtained by the application of Granger's most general concept of causality (Granger and Newbold, 1977), in the context of a specific non-linear multivariate model:

$$\mathbf{z}_t = \mathbf{F}(\mathbf{Z}_t) + \varepsilon_t \quad (8)$$

in which  $\mathbf{F}$  is not necessarily linear. The difficulty of

this task has been the specification of a tractable framework for non-linear time series analysis. One such framework was proposed by Ozaki (1985) and later generalised by Tong (1990). This consists in specifying a linear Autoregressive model in which the coefficients  $\mathbf{A}_k$  will now depend on the previous states of the system:

$$\mathbf{z}_t = \sum_{k=1}^p \mathbf{A}_k(\mathbf{Z}_t) \cdot \mathbf{z}_{t-k} + \varepsilon_t \quad (9)$$

(This is a generalisation of Eq. (3) which has coefficients  $\mathbf{A}_k$  independent of the delay matrix  $\mathbf{Z}_t$ .)

A number of recent papers implementing Granger causality measures are based on particular non-linear time series models (Teräsvirta, 1998; Warne et al., 1999). It must be stressed, however, that the model selected for implementing causality measures must be matched to the dynamic characteristics of the time series studied. Recent work (reviewed in Valdes et al. (1999)) has shown that not all non-linear models are capable of capturing the complex characteristics of neural signals. The class of models that offered a good trade-off between computational complexity and descriptive properties were the use of locally weighted polynomial non-parametric regression (Fan and Gijbels, 1995, 1996). Bell et al. (1996, 1998) have devised Granger causality measures for a specific class of additive local polynomial models. In a series of recent papers Valdes and co-workers (reviewed in Valdes et al. (1999)) have applied Local Linear polynomial regression to the analysis of neural signals. On the basis of this technique they have implemented a specific measure of Granger causality for the analysis of non-linear multivariate neurophysiological signals and have carried out the preliminary evaluation of these measures (Valdes et al., 1996). This family of models includes ordinary linear Autoregression as a special case.

The purpose of this paper is fivefold

1. To describe a general framework that encompasses both linear and non-linear modelling of neurophysiological time series by means of Local Linear Non-linear Autoregressive models (LLNAR).
2. Within this framework to describe new tests for a) non-linearity of time series and for b) non-linearity of neural interactions, both based on the LLNAR model.
3. To introduce a specific measure of Granger Causality for directed influences based on the LLNAR model and a test of significance for this measure.
4. To show the advantages of this measure of causality for non-linear data.
5. To show examples of the use of LLNAR with non-linear reference data and local field potentials (LFPs).

## 2. Material and methods

### 2.1. The electrophysiological data

To test the new data analysis methods, recordings were taken from the visual cortex of awake macaque monkeys. Since numerous previous crosscorrelation studies have shown that the likelihood to find synchronous activity generally declines with cortical separation of recording sites (see, e.g. (Ts'o et al., 1986; Ts'o and Gilbert, 1988; Krüger and Aiple, 1988; Engel et al., 1990)), we chose to apply our methods to data obtained from recordings from within the same cortical area in order to maximise chances for finding signs of interaction. Amongst the many different visual cortical areas, we chose to analyse field potential recordings from area TE for the following reasons: First, the large receptive fields of the cells in this part of the brain are indicative for integrative mechanisms in this area. Second however, the large receptive fields allow for many independent stimuli to stimulate the same neuron. Thus, the problem to handle related signals coherently and at the same time functionally separate them from unrelated signals is aggravated. This suggests the existence of dynamic mechanisms which organise interactions between different neurons. Third, in previous correlation studies in this area we found that a high percentage of cells is firing synchronously in response to a stimulus (Freiwald et al., 1998). This result implies that the probability to find interactions between LFP signals should be high, since they are the manifestations of coherently firing local groups of neurons.

#### 2.1.1. Behavioural procedure

Two male macaque monkeys (*Macaca mulatta*) were trained to perform a visual fixation task (Wurtz, 1969). This task required each monkey to maintain fixation at a spot of light with a diameter of  $0.3^\circ$  that appeared on a CRT screen at a distance of 57 cm or 114 cm to the eyes. Within 3 s after the appearance of the light spot the monkey had to start fixation and subsequently press a lever. Fixation had to be maintained for an interval of 5–7 s after which the light spot dimmed slightly and the monkey was required to release the lever within 500 ms. Successful performance of each trial was rewarded with a drop of juice or water, and after a 2 s waiting period the next trial was started. If the animal made an eye movement of more than  $0.7^\circ$  away from the fixation spot while the lever was pressed, or if it released the lever before the dimming period, the trial was aborted and a prolonged waiting period of about 4 s started without a reward. After implantation of the head holder, the animal's head was restrained during training and recording sessions, and eye movements were monitored with an infrared eye-tracking system.

#### 2.1.2. Surgery

Each monkey was implanted surgically under aseptic conditions with a head holder and a recording cylinder of 20 mm diameter. Anaesthesia was induced with an injection of ketamine (10 mg/kg, i.m.), and after tracheal intubation continued with 1–3% isoflurane in oxygen/nitrous oxide (30/70). To aid the positioning of the recording cylinder, the monkeys had been scanned before the surgery with magnetic resonance imaging (MRI), and the stereotaxic co-ordinates of STS and PMTS had been determined using the BRAINVOYAGER software (Goebel, 1996). The vertically oriented cylinder was centred above the estimated AP position of the anterior end of the PMTS. Head holder, cylinder, and screws were fixed and interconnected with dental acrylic. A small craniotomy of 2 mm diameter was placed in the centre of the cylinder. Postoperative treatment included systemic application of antibiotics for 5 days.

#### 2.1.3. Recording procedure

Local field potentials (LFPs) and multi- or single-unit activity were recorded in the posterior part of the inferotemporal cortex (TE), immediately anterior to area TEO with two to four varnish-coated tungsten microelectrodes. The impedances of the electrodes were 1–2 M $\Omega$  at 1 kHz. The electrodes were advanced independently through a 23 gauge guide tube. Initially, a short guide tube was used to record the depth profile. The correct location for recording sites was determined from the transitions through layers of cortex, white matter and sulci and from the response properties of the neurons encountered. Based on these results, a longer guide tube was chosen to ensure the final positioning of the electrodes inside the target area. Its tip was located 7 mm above the closest point of the recording area. To place the electrodes for simultaneous recordings from spatially separate columns in TE they were slightly bent on the last millimetres before their tips and oriented to move in slightly different directions. This resulted in differences of travelling distances of the electrodes of 2 mm or more. The separation of the recording sites could then be estimated to be at least equal to this difference of travelling distances. Only recording sites with a minimal spatial separation of 2 mm or more were considered for further analysis. This distance is four times the spatial range of LFPs ( $\approx 500 \mu\text{m}$ ) because of volume conductance (Schillen et al., 1992). Therefore, synchronisation of any two simultaneously recorded time series is very unlikely the trivial result of current spread from a single source to these two recording sites.

The signal from each electrode was amplified by a variable factor, filtered with a bandpass (1–300 Hz) to extract the local field potential and a second bandpass (0.5–4 kHz) to extract action potentials. The field

potential data was then A/D converted with a sampling rate of 512 or 1024 Hz (A/D converter board DT 2821-G (DataTranslation)) which was controlled by the Discovery (DataWave) data acquisition system and stored to disk. No online analysis was performed during electrode positioning or data acquisition to avoid any bias in the selection of recording sites.

The correct location of the recording sites in posterior part of area TE has been histologically verified in one of the experimental animals.

All procedures used in this study were performed in accordance with the guidelines for the welfare of experimental animals issued by the federal government of Germany and conformed to the guidelines of the National Institutes of Health for the care and use of laboratory animals.

#### 2.1.4. Generation of visual stimuli

Light stimuli were generated on a CRT display with a  $1024 \times 768$  pixel resolution, subtending a visual angle of  $35^\circ \times 26^\circ$  in the case of 57 cm monitor distance to the eyes and  $18^\circ \times 13^\circ$  in the case of 114 cm monitor distance. The monitor operated with a frame rate of 80 Hz (non-interlaced). This rate is well above the temporal resolution of macaque cones (Boynton and Baron, 1975). Independent control measurements in area MT had previously shown (Kreiter and Singer, 1992), that the 80 Hz frame rate does not influence the temporal structure of the spike train.

The stimulus set consisted of fractal patterns (Miyashita et al., 1991) and pictures of animals and plants. Sizes ranged from  $2^\circ$  to  $10^\circ$  of diameter and luminance from 1 to 4 cd/m<sup>2</sup> with a background illumination of 0.1 cd/m<sup>2</sup>. The five pixel lines closest to the borders of each picture were reduced in luminance to avoid sharp luminance contrasts. Stimulus and presentation positions were chosen so that responses were elicited in at least one of the recording sites. The stimulus was presented in one of two temporal schemes. In the first, the stimulus was turned on 1 s after trial onset and stayed on for 4 s until the end of the trial. In the second scheme, two stimuli were presented subsequently at the same location with presentation and inter-picture delay times of 1 or 1.5 s.

## 2.2. The analysis of electrophysiological data

### 2.2.1. Data processing

For all the field potential data recorded, power spectra were computed. These showed that there was very little power beyond 100 Hz in the data. To save on computer time for the following computations, the signals were therefore resampled with 200 Hz by Cubic spline interpolation (with the tension parameter  $\sigma = 1.0$ ). Whenever necessary, digital notch filters were applied to remove 50 Hz line or 80 Hz monitor artefacts

(before the resampling of the data). All data were z-transformed by subtraction of the mean and division by the standard deviation (SD). Therefore, in what follows, the SD is always equal to one.

### 2.2.2. Local Linear Non-linear Autoregression (LLNAR)

We will now proceed to describe the LLNAR model as a generalisation of *linear* Autoregression. For linear Autoregression, the estimation of the model described in Eq. (4) is carried out by least squares, that is by finding the coefficients  $a_k$  that minimise the estimated innovation variance  $\sigma_{x_t|X}^2$ :

$$E_{x_t|X}^{\text{LIN}} = \sum_{t=1}^{N_t} \left( x_t - \sum_{k=1}^p a_k \cdot x_{t-k} \right)^2 \quad (10)$$

Thus, the linear model assumes that whatever the previous state  $\mathbf{X}_t$  may have been, the Autoregressive coefficients  $a_k$  are constant.

In order to model non-linearity, assume that the  $a_k$  depend on the  $\mathbf{X}_t$ . Thus, a different linear regression model is assumed for each point of the state-space (Tong, 1990). This allows flexible modelling of many types of non-linear systems. The coefficients  $a_k(\mathbf{X})$  are estimated by a ‘local regression’ (Fan and Gijbels, 1996) which gives greater weight to those data pairs  $\langle x_t, \mathbf{X}_t \rangle$  with delay matrices near  $\mathbf{X}$ . This is achieved by modifying Eq. (10) into the following expression:

$$E_{x_t|X}^{\text{LLNAR}} = \sum_{t=1}^{N_t} \left( x_t - \sum_{k=1}^p a_k(\mathbf{X}) \cdot x_{t-k} \right)^2 K_h(\|\mathbf{X} - \mathbf{X}_t\|) \quad (11)$$

where  $K_h(x) = \exp\left(-\frac{1}{2}\left\|\frac{x}{h}\right\|^2\right)$  is a Gaussian shaped weighting function. The ‘spread’ of the weighting depends on  $h$ , which we shall term the ‘bandwidth of the local linear smoother’ or simply ‘bandwidth’ (Marron, 1992). The bandwidth  $h$  is related to ‘effective degrees of freedom’ of the nonparametric fit (Hastie and Tibshirani, 1990). When  $h$  is large the number of ‘effective’ parameters is that of the linear model,  $p$ . When  $h$  is small then the number of effective parameters increases.

This is the Local Linear Non-linear Autoregressive time series model (LLNAR), designated as such because it is approximately linear in the neighbourhood of a given previous state of the system. Note that as  $h \rightarrow \infty$  then LLNAR models reduce to the ordinary Autoregressive model that is therefore included as a specific case. This model has been shown to adequately describe non-linear characteristics of EEG data (Hernández et al., 1996).

The model depends on the selection of the bandwidth  $h$  and also on the number of delays  $p$  to be used as predictors. At first thought these could be selected by estimating the usual variance of the prediction error (using the state dependent Autoregressive coefficients):



$$\hat{\sigma}_{x|X}^2 = \frac{1}{N_t} \sum_{t=1}^{N_t} \left( x_t - \sum_{k=1}^p \hat{a}_k(\mathbf{X}_t) \cdot x_{t-k} \right)^2 \quad (12)$$

where for any population quantity  $c$ ,  $\hat{c}$  will denote its estimator.

However, Eq. (12) is a biased estimator since the data predicted is included in the ‘training set’. In fact, it is zero when  $h$  is zero since then each observation is predicted by itself!

A number of techniques have been suggested for model order determination including AIC, BIC, FPE (for a review see Tong (1990)). However, Yao and Tong (1994) have argued in favour of using the cross-validation error (CVE) as a goodness of fit measure for tuning model parameters in non-linear time series. From the literature of nonparametric regression estimation it is well known that CVE avoids overfitting (Hastie and Tibshirani, 1990). In the case of the LLNAR model CVE is a function of the model order  $p$  and the bandwidth  $h$ , this functional relation being denoted as  $\text{CVE}(p, h)$ . Yao and Tong (1994) and Bell et al. (1998) have carried out simulations that demonstrate that the CVE performs well as an order determination procedure using local polynomial regression, even in the case when the data is actually linear.

The CVE is an unbiased estimate of the prediction error obtained by the following procedure. Successively each pair  $\langle x_t, \mathbf{X}_t \rangle$  is deleted from the complete sample, the LLNAR model is fitted, and using this model a prediction is obtained for the deleted data pair. The average square of the resulting residuals is the CVE. An examination of  $\text{CVE}(p, h)$  for a suitable set of values will indicate not only the optimal bandwidth  $h$ , but also the optimal choice of  $p$ . Typically  $\text{CVE}(p, h)$  is larger for small  $p$  when the order of the model is not high enough to describe the data. Then as  $p$  increases this measure is smaller until overfitting occurs, in which case the value of CVE increases as prediction performance decreases. In a similar fashion CVE is high when  $h$  is too large — unless the linear model is adequate. For nonlinear systems CVE decreases until it becomes too small for good generalisation performance between the data points. The  $\text{CVE}(p, h)$  function therefore shows a global minimum which is used to select  $p$  and  $h$ .

We calculate  $\text{CVE}(p, h)$  for all values of  $p$  from 1 to 20 and for  $h$  defined on a logarithmically spaced grid in the range  $0.01 \cdot s \cdot \sqrt{p} \leq h \leq 4 \cdot s \cdot \sqrt{p}$ , where  $s$  is the standard deviation of the time series to be examined. Additionally, the CVE for  $h_\infty$ ,  $h = \infty$  (linear model) is also computed.  $(p_{\min}, h_{\min})$  are the values for which the CVE function attains its minimum, which will be denoted by  $\text{CVE}(p_{\min}, h_{\min})$ ,  $\text{CVE}(p_{\min})$ ,  $\text{CVE}(h_{\min})$  or simply CVE according to the parameter that is being discussed.

### 2.2.3. Test for non-linearity in time series

In this section a novel test for assessing linearity of time series is proposed. In the fitting procedure described just above  $h_{\min} = h_\infty$  is an indication that a linear model fits the time series best. When  $h_{\min} \neq h_\infty$ , then the following procedure is carried out. A series of surrogate time series are generated using the linear model by means of the ‘wild bootstrap’ technique (Mammen, 1999). Essentially, this technique generates artificial time series using the linear Autoregression equation. For the generation of each new realisation a new innovation time series is obtained by randomising the residuals of the linear fit. The set of the CVE of the surrogate data defines a bootstrap empirical probability distribution that characterises the null hypothesis. If  $\text{CVE}(p_{\min}, h_{\min})$  is lower than a prespecified (small) proportion of the bootstrapped linear CVE then the hypothesis of linearity is rejected. Specifically in this paper, linearity was rejected if  $\text{CVE}(p_{\min}, h_{\min})$  was smaller than the fifth percentile of the logspline estimated distribution function (see below) of the CVE of the surrogate data ( $P$  value of 0.05).

### 2.2.4. Multivariate regression

The LLNAR can be extended to encompass non-linear multivariate regression models that generalise Eq. (2). The non-linear equivalent to model Eq. (5) is fitted by the general procedure described in the previous section but applied to the pairs  $\langle z_t, \mathbf{Z}_t \rangle$ , i.e. using the past of both series one and two. Thus, the LLNAR version of prediction of  $y_{1,t}$  by the complete delay matrix  $\mathbf{Z}_t$  is given by

$$E_{x|Z}^{\text{NONLIN}} = \sum_{t=1}^{N_t} \left( x_t - \sum_{k=1}^p a_k(\mathbf{Z})x_{t-k} + \sum_{q=1}^r b_q(\mathbf{Z})y_{t-q} \right)^2 \cdot K_h(\|\mathbf{Z} - \mathbf{Z}_t\|) \quad (13)$$

As in the linear case, the number of lagged values for series two  $r$ , does not necessarily have to be the same as  $p$ . In this paper the value of  $p$  obtained from the regression of a time series on its own past is taken to be fixed. Past values of series two are then added if they lower the CVE. Thus, CVE is used in this multivariate regression for two purposes, for determining the value of  $r$  (how many past values of series two are necessary) and also for determining the optimal smoothing parameter  $\text{CVE}(r_{\min}, h_{\min})$  for the Local Linear regression. This provides valuable information about the influence of series two on one. If the CVE is not decreased by adding past values of series two, then one may conclude that this series does not influence series one. Additionally, an examination of the effect of the bandwidth on  $\text{CVE}(r_{\min}, h_{\min})$  by means of the procedure outlined in the previous section is a test for the linearity of the interaction.

### 2.2.5. Non-linear Granger influence measure

If past values of series two contribute to the prediction of series one, then it is possible to use the LLNAR framework for a quantitative evaluation of Granger causality. Define:

$$\hat{\sigma}_{x|z}^2 = \sum_{t=1}^{N_t} \left( x_t - \sum_{k=1}^p \hat{a}_k(\mathbf{Z})x_{t-k} + \sum_{q=1}^r \hat{b}_q(\mathbf{Z})y_{t-q} \right)^2 \quad (14)$$

Then, the non-linear measure for Granger Causality based on LLNAR is:

$$I^{\text{LLNAR}}(y \rightarrow x) = \ln \left( \frac{\hat{\sigma}_{x|x}^2}{\hat{\sigma}_{x|z}^2} \right) \quad (15)$$

Note that this measure includes the linear measure of influence as a special case when the bandwidth of the LLNAR is set to infinity.

According to Bell et al. (1998) the significance of the influence measure was tested by generating a bootstrap sample under the null hypothesis. This sample was obtained by the following procedure. Separate LLNAR models are fitted to each time series. Then, using the wild bootstrap procedure described above, independent pairs of surrogates time series are generated for both series one and two. Under these conditions stochastic

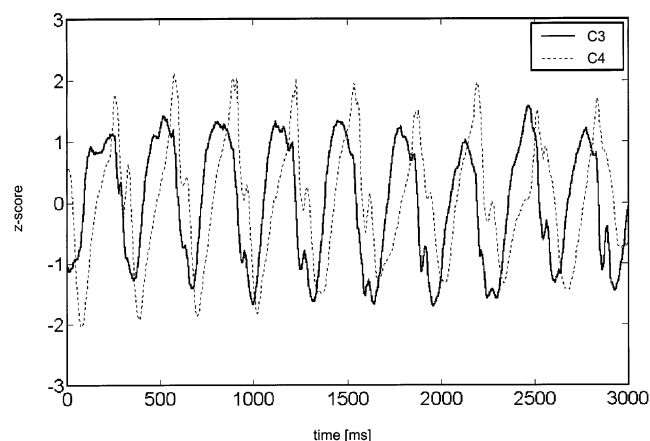


Fig. 1. Test series for analysis of non-linear dynamics. Two simultaneously recorded EEG channels (average reference) from an epileptic patient with complex partial seizures exhibiting spike and wave discharges. Each recording is scaled to zero mean and SD 1. Recordings of 3 s duration from electrode position C3 (solid line) and C4 (dashed line) of the 10/20 system are shown. Clinical and neuro-imaging data indicated the existence of a primary focus in C3 which propagates to C4. The temporal lag between the two signals, C3 leading over C4, can be seen. The best fit to the signal in C4 (as chosen by minimising the crossvalidation error,  $\text{CVE} = 0.0043$ ) was obtained by using a Local Linear Non-linear Autoregressive model (LLNAR) which depended on 11 time lags (corresponding to 55 ms) and used a Gaussian kernel with bandwidth  $h = 0.65$ . Adding five values of the past of C3 (25 ms) to the model led to a significant decrease of the CVE to 0.0033 with a bandwidth  $h = 1.10$ . Series C3 achieved a CVE of 0.0039 also for 11 lagged values with a bandwidth of 0.99. However, adding five past values of C4 only led to an insignificant decrease of the CVE to 0.0034 with a bandwidth  $h = 1.09$ .

independence of both series is ensured and in this case the population influence measure should be by construction zero. The actual influence measures are calculated for each pair of bootstrapped surrogates and a histogram of the influence measures is constructed. In this paper the null hypothesis of no influence was rejected if the observed influence measure was greater than the 95th percentile of the logspline estimated distribution function (see below) of the surrogate data ( $P = 0.05$ ).

### 2.2.6. Assessment of significance of results

All significance testing was carried out using estimates of the distribution under the null hypothesis obtained by means of the bootstrap (Efron and Tibshirani, 1993). The number of bootstrap samples always was larger than 700. Instead of using the raw histograms for determining critical values of the null distribution, estimate of this density function was obtained using the logspline technique of Kooperberg and Stone (1991) that allows better estimates of tail probability densities. The critical value for a one sided test at a given  $P$  value (in our case  $P = 0.05$ ) is the  $1 - P$  percentile of this distribution. All figures of significance tests in this paper show the raw histogram, the superimposed logspline density, the critical value for the one sided test, as well as the actual values of the test statistics.

## 3. Results

### 3.1. Results for test time series

The methods developed above were applied to two test data sets, and afterwards to LFP data. For the first test a set of 40 time series with a length of 600 data points were generated from a bivariate linear Autoregressive model with  $p = 2$ . Half the time series were generated as interdependent by construction of the autoregression matrices. The other half was generated as a set of independent time series. In all 40 cases the type of LLNAR model selected corresponded to a bandwidth of infinity, i.e. the linearity of the model was correctly identified. Additionally, the presence or absence of influence was correctly detected in all instances.

As a second test of the techniques described here, a recording of a spike and wave EEG from an epileptic patient with complex partial seizures was selected for analysis. This is a well studied time series which has been demonstrated to be highly non-linear by many different techniques (Valdes et al., 1999). Fig. 1 shows data from two channels (C3 and C4 of the 10/20 system) of the simultaneously recorded time series. Clinical and neuro-imaging data indicated the existence of a primary focus in C3, which propagates to C4.

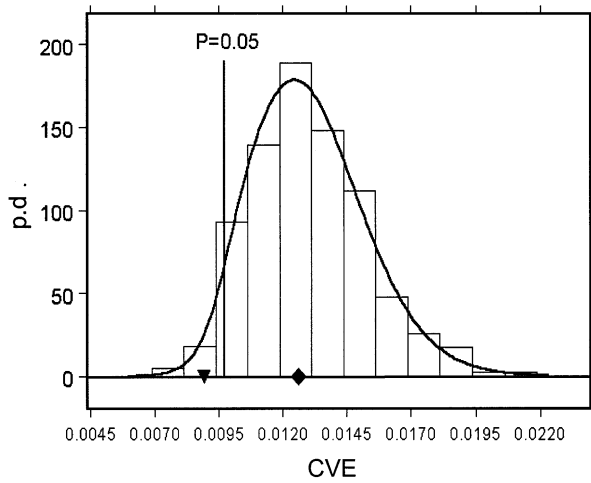


Fig. 2. Results of the test for non-linearity for the Spike and Wave data (C3). In this figure, and all other figures for statistical tests, the histogram of the test statistic for the null hypothesis obtained by means of the bootstrap is shown as a probability density distribution (p.d.) together with a log-spline estimate of the p.d. superimposed as a curve. In this case the test statistic is the CVE. The distribution of the null hypothesis is obtained fitting the LLNAR to linear surrogate time series (bootstraps). The histogram indicates the range of variation of CVE for the linear model. The log-spline density estimate is used to calculate the critical value for rejection of the null hypothesis at the  $P=0.05$  level. This level is indicated by a vertical line. The diamond indicates the CVE for a linear fit to the actual data. The triangle indicates the CVE for the non-linear model. Note that the CVE for the non-linear model is significantly lower than the range of variation of CVE for the linear model providing evidence that the analysed EEG data are indeed a non-linear time series. Similar results were obtained for C4 (not shown).

A  $p$  of 11 time lags corresponding to 55 ms of the signal's past was necessary to adequately fit the signal from C4. The bandwidth  $h_{\min}$  was found to be 0.65, which is quite low and indicates that the signal might be a non-linear one. The best fit to signals in C3 was obtained using  $p=11$ , indicating an equally complex signal. In this case an  $h_{\min}$  of 0.99 was obtained, which is similar to that of C3.

Using the new non-linearity test both signals were found to be non-linear. This is illustrated in Fig. 2 for the data from deviation C3. This figure shows that the CVE for the LLNAR model using  $h_{\min}$  (marked by the triangle in Fig. 2) was less than the critical value for rejection of the null hypothesis of linearity at the  $P=0.05$  level (vertical line) and is outside of the range of crossvalidation errors for the surrogate time series of the linear model (histogram), indicating a significant difference between the two models. Thus, the time series should be considered to be a non-linear one. Similar results were obtained for C4.

The results obtained with the non-linear influence measures were consistent with the expectations mentioned above. Five lagged values of C3 did improve the crossvalidation error of C4. This conclusion is illus-

trated in Fig. 3A, which shows that the influence measure of the non-linear model (upwards pointing triangle) is significantly higher than expected by chance.

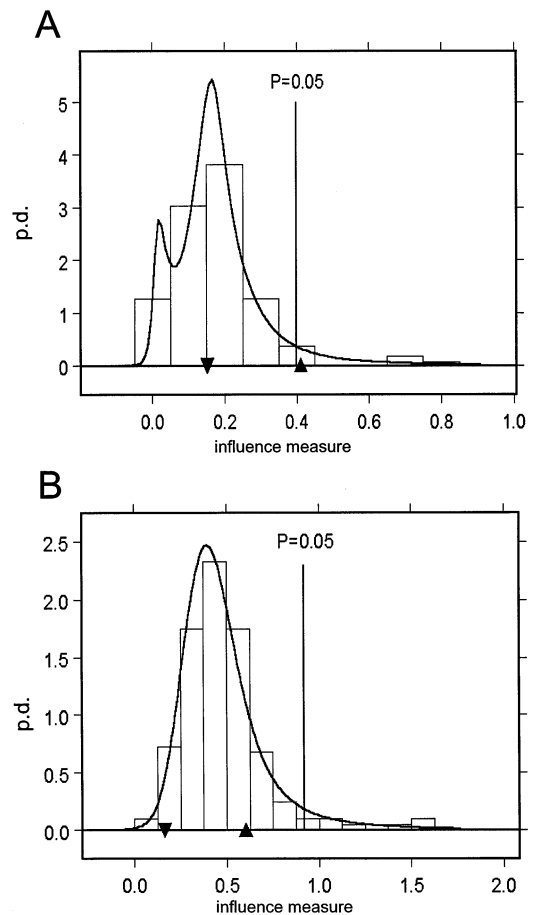


Fig. 3. Significance test of the influence measure between the signals shown in Fig. 1. The analysis of the influence of C3 on C4 is shown in A, the influence analysis of the reverse direction (C4→C3) in B. Conventions for the figures are as described in Fig. 2. In this case, however, the test statistic is the influence measure,  $I(y \rightarrow x)$ . Rejection of the null hypothesis of no influence occurs when the influence measure of the actual data is larger than the critical value for  $P=0.05$  (indicated by a vertical line in A and B) of the log-spline density estimate (solid line) of the surrogate data constructed to be independent. Shown in these figures are the histograms of bootstrapped influence measures for surrogate time series constructed to be independent, thus approximating the null hypothesis of no influence. A: The upwards pointing triangle on the right marks the value of the estimated influence of C3 on C4 based on the non-linear model. This value is significantly larger than would be expected by the null hypothesis of no influence. Thus, the null-hypothesis of no influence can be rejected. The downwards pointing triangle on the left marks the value of the estimated influence in the same direction (C3→C4), but based on the linear model. In this case, the null-hypothesis cannot be rejected. Thus, a directed influence could only be inferred from the non-linear, but not from the linear model. B: The influence of C4 on C3 was insignificant both in the linear (downwards pointing triangle) and the non-linear model (upwards pointing triangle). Thus, no past value of C4 was able to decrease CVE of C3 significantly. Taken together, this analysis suggests a uni-directional non-linear influence of deviation C3 on C4.

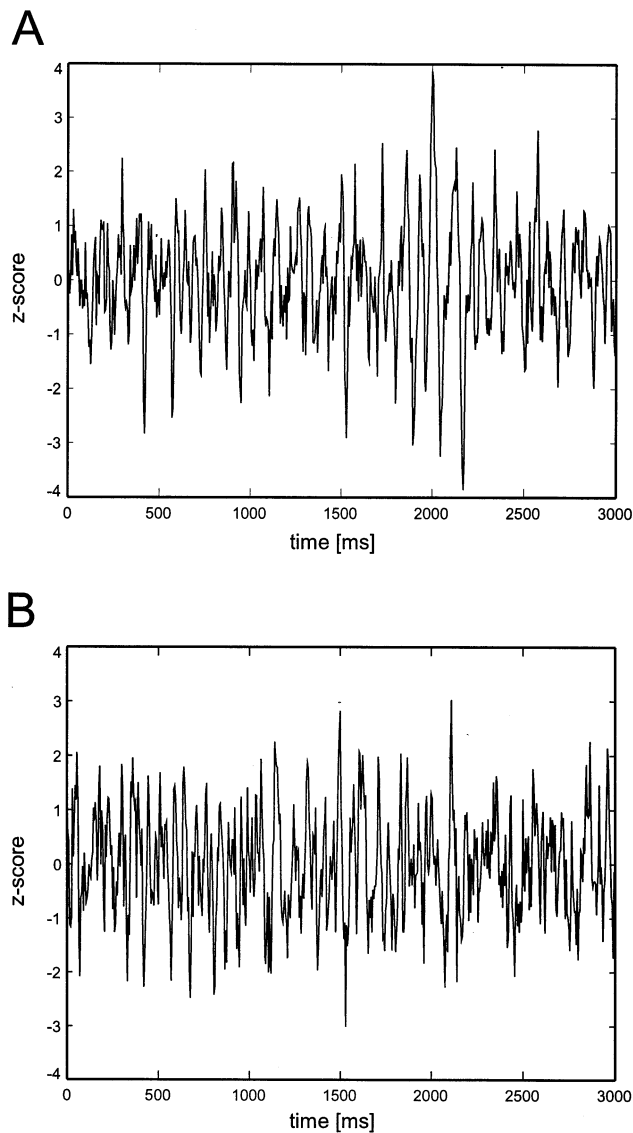


Fig. 4. Examples of local field potential (LFP) data obtained from the posterior part of area TE. Two simultaneously recorded data segments of 3 s duration each are shown in parts A and B. While phases of oscillatory activity can be observed (maybe most clearly in the interval between 1700 and 2200 ms in A), the overall appearance of this data is much more irregular than that in Fig. 1. The dominant frequencies of the data are in the alpha and beta frequency bands, while less power is contained in lower frequency bands. The best fit to the LFP shown in A (as chosen by CVE) was obtained by using a LLNAR which depended on 9 time lags (corresponding to 45 ms) and used a Gaussian kernel with bandwidth  $h = 3.127$ . The best fit to the signal in B was obtained using 12 past values (corresponding to 60 ms) and  $h = 13.81$ .

However, based on the linear model (downwards pointing triangle), no significant influence was detected. In the case of C3, also five lagged values of C4 improved the CVE. In this case (Fig. 3B), however, the test for the influence measure was not significant. Thus, for this set of data the conclusion is, that a unidirectional non-linear influence was present from C3 to C4 but not vice-versa.

### 3.2. Results for LFP data

Data from six different experiments have been analysed. From each of these data sets a random sample of 20 sweeps was selected to perform the tests of non-linearity and influence. An example of the general appearance of LFP signals from area TE is shown in Fig. 4. The time series data appear to be more irregular than that of Fig. 1 in the sense that a repeating underlying pattern is hardly visible. Oscillatory episodes can nevertheless be observed. A frequency analysis (not shown) revealed that most of the power of the signal is within the alpha and beta frequency bands, while the contribution of lower frequencies is much less pronounced.

The number of lagged past values needed to construct adequate Autoregressive models varied from five to 12. Thus, the prediction of a given data point had to be based on the 'history' of the preceding 25–60 ms. A  $p$  value of 12 is identical to the maximum considered in our calculations for reasons of computational efficiency. In at least some of these cases of maximal model length, the inclusion of even more past values might have improved the Autoregressive model.

The minimum of CVE was  $h_{\infty}$  in almost all cases. Thus, for most of the LFP time series, the linearity hypothesis could not be rejected. This result is shown in Fig. 5 for the example time series of Fig. 4A. The CVE for the LLNAR model using  $h_{\min}$  (indicated by the

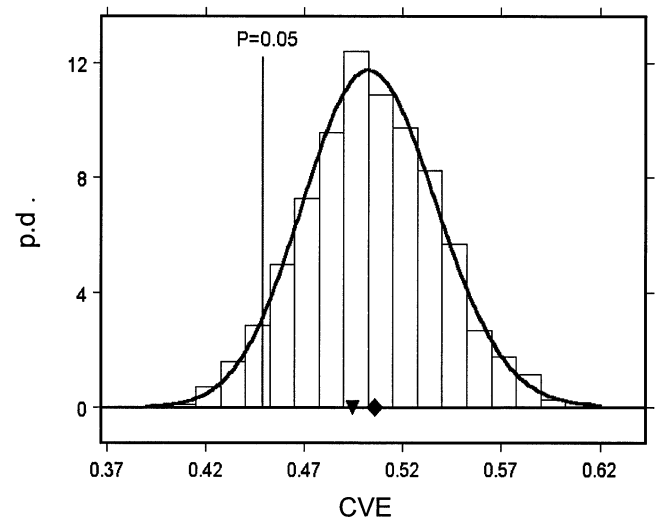


Fig. 5. Results of the test for non-linear autoregression for LFP data shown in Fig. 4A; conventions as in Fig. 2. The CVE of the linear model (diamond) and the CVE of the non-linear model (triangle) are very similar and well within the histogram showing bootstrap estimations of the variability of CVE for the linear model as well as the logspine estimate of the probability density. Therefore, both are much larger than the critical value for  $P = 0.05$  indicated by a vertical line. The null hypothesis of linearity cannot be rejected. Similar results were obtained for the LFP data shown in Fig. 4B. Thus, the LFP data should be considered to be linear.

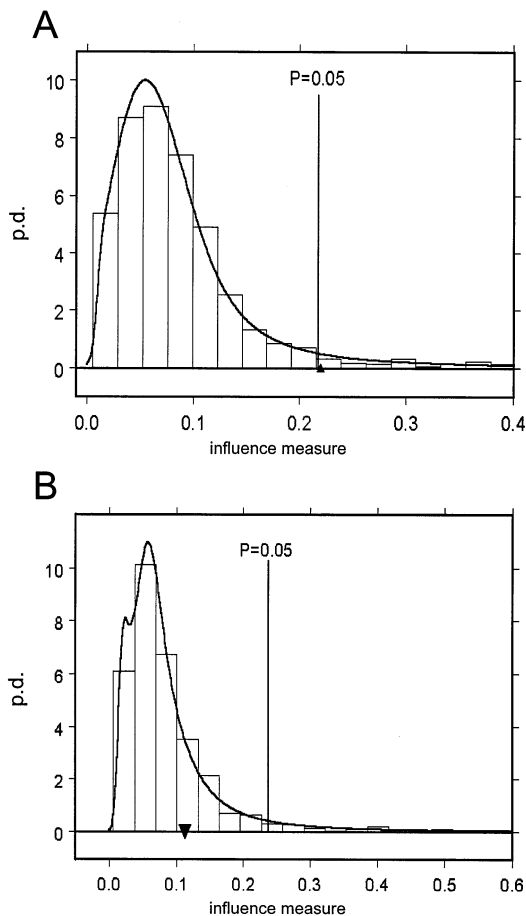


Fig. 6. Significance test of the influence measure between two simultaneously recorded LFP signals, conventions as in Fig. 3. The test statistic is the influence measure. The two signals of the recording sites will be referred to as channel 1 and channel 2. A and B show the significance tests for the influence of channel 1 on channel 2 (A) and for the influence of channel 2 on channel 1 (B). A: The estimated influence of channel 1 on channel 2 is significant at the  $P = 0.05$  level. B: The estimated influence value for the reverse direction is contained within the distribution of bootstrapped values. Therefore, the null-hypothesis (no influence of channel 2 on channel 1) cannot be rejected. Taken together these findings indicate that the interactions between channels 1 and 2 are uni-directional (channel 1  $\rightarrow$  channel 2).

triangle) was very similar to that of the linear model with  $h_{\infty}$  (indicated by the diamond) and is contained within the distribution of crossvalidation errors for the surrogate time series of the linear model shown by the histogram in Fig. 5. This indicates that the time series should be considered to be linear. In the few instances in which the minimum of CVE was not approaching  $h_{\infty}$ , the test for non-linearity always showed that the null hypothesis of linearity could not be rejected. Thus, all the observed data could be modelled as linear time series.

Directional interactions, both uni- and bi-directional ones, between pairs of LFP recordings have been found. However, not all pairs of simultaneously recorded LFPs exhibited this form of dependency. The

data shown in Fig. 4 for example did not show any signs for directional interactions. An example of a significant directed influence between two simultaneously recorded LFP data segments is shown in Fig. 6. The test of the estimated influence value of channel 1 on channel 2 indicates that this influence is significant (Fig. 6A), while the influence in the reverse direction is insignificant (Fig. 6B). The interaction therefore is a uni-directional one.

#### 4. Discussion

The major points emphasised in this article are:

1. The introduction of non-parametric non-linear Autoregressive methods, originally developed for econometrics, for the analysis of neural signals. A significance test for non-linearity of time series is presented.
2. The introduction of particular methods for detecting the non-linear character of neural time series and the presence of non-linear interactions. It is now possible to explore these non-linear characteristics of neural data with methods (LLNAR) that reduce automatically to linear techniques, if the data is linear.
3. The introduction of a measure for both linear and non-linear Granger causality and a test for its significance.
4. The demonstration of the importance of non-linear modelling. In particular, it is shown that poor estimates of causal relations may result from applying linear methods to non-linear data (Fig. 3).
5. Applying these analysis tools to LFP data from macaque area TE yielded three main findings. First, directed interactions have been found in area TE. Some of these interactions were uni-directional, others bi-directional. Second, the construction of adequate Autoregressive models required the inclusion of up to 12 lagged values. Thus, the present state of the system is influenced by its own past of up to 60 ms duration. Third, all the LFP signals could be described by linear models.

##### 4.1. Local field potential analysis

To our knowledge, the findings presented here are the first to provide evidence for the existence of directional interactions in the macaque cortex. Specifically we have found that within the same cortical area, one synchronously active neural population, which is generating the local field potential we are analysing, is exerting an influence on a second such group at another site. The existence of asymmetrical interaction patterns of neural groups within the same cortical area is quite surprising, given the fact that both recording sites are

located at the same level of the processing hierarchy and therefore no a priori expectation existed which site, if any, would exert a causal influence on the other. The existence of directional interactions would have been hard to reveal with other methods, e.g. classical cross-correlation techniques. In the examples analysed here, the cross-correlograms did not give any indication for an asymmetrical relationship as we have observed using influence measures based on the LLNAR technique. Thus, a potentially very important aspect of the relationship between the activities of neural populations would have remained unnoticed.

The finding of directional interactions between spatially separate neural groups purports to a possible role of the long range horizontal connections coupling different parts of the same visual area for directly conveying signals from one part of the area to another one. This idea is compatible with the view that these connections do not influence classical receptive field properties, but are rather related to the temporal organisation of the signal flow in cortical networks (Singer and Phillips, 1997). The existence of directed interactions between groups of synchronously firing populations of cells represents further support for the hypothesis outlined in the introduction, that temporal activity patterns are of functional relevance for visual information processing. Here, neural groups seem to be dynamically organised in a way that the synchronous activity of one group has a causal effect on the other one. An extensive study of the properties of directed influences and their relationship to stimulus properties and behavioural states of the animal shall be performed to clarify the validity of our theoretical considerations.

Our second main finding is that each state of a coherently firing group in area TE depends upon past values of up to 60 ms duration. This time period, which we will also refer to as the ‘memory duration’ or ‘memory span’ of the system, is similar for the dependency of the signal on its own past and for the influences exerted by other neural groups. A memory duration of 60 ms is long compared to the 5–8 ms found in a study of directional influences in the cat visual cortex (Bernasconi and König, 1999) and compared to time constants of cortical pyramidal cells (Koch et al., 1996). However, in the following we will point out that a) our result does fit well to earlier findings in area TE, which is part of what has been called ‘the slow brain’ (Nowak and Bullier, 1997) based on response latency measurements, and b) by comparison with results from other cortical areas indicate a possible relationship with oscillation frequencies.

Long lasting linear dependencies within and between spike trains of individual neurons in area TE have been revealed with auto- and crosscorrelation analyses (Gochin et al., 1991; Freiwald et al., 1998). In these studies, broad correlation peaks have been described

whose half width could even extend to 200 ms and more. Thus, the occurrence of a spike of one cell was found to influence the likelihood of a second cell for eliciting a spike even after this long temporal delay. An analogous finding was made for auto-correlograms, showing that similar dependencies also exist within a single unit spike train (Freiwald et al., 1998). Thus, the statistical dependencies of single unit activities are paralleled by similar phenomena at the population level, suggesting that neurons are firing in synchrony to exert influences at a larger scale onto other parts of the system.

A further observation might link the duration of memory, the value  $p$ , of the Autoregressive model to the dominant frequencies in the cortical area the LFP signals were recorded from. In area TE, the field potential signals often contain most of their power in the higher alpha and lower beta frequency range. Therefore, the duration of one such oscillatory cycle is slightly longer than the memory duration found in our data. Interestingly, a study aimed to detect coherencies in data from cat area 17, where higher frequencies in the gamma range are dominant (Pawelzik, 1994), found values for the duration of memory of about 25 ms — a value at the order of one oscillatory cycle. However, this inverse relationship of different dominant frequencies and memory duration remains speculative, since different methods were used to assess the latter quantity, and at least in our case, more data are needed to prove the existence of such a relationship. Yet, this observation might help to explain the difference of our  $p$  values to those found to be optimal for fitting linear Autoregressive models to LFP data from cat visual cortical areas **a17** and **a7** (Bernasconi and König, 1999). In this study, typically only values from the past 5–8 ms had to be considered. Since species and task differences might contribute to this discrepancy, a comparative study of simultaneous recordings in different visual cortical areas of different processing streams in the same animal might be a worthwhile endeavour to further investigate memory values and to assess their possible functional implications. However, high  $p$  values are in many cases linked to a complex structure of frequency domain features, with possibly narrow band peaks, an aspect which should also be the subject of future enquiry.

Our third main finding is that all LFP signals appear to be linear as well as the influences which exist between two such signals. Generally, the existence of dependencies detected by linear methods does not imply that the dependencies are of a linear nature. However, for the LFP we have been able to reach this conclusion by using the new tests for non-linearity described above.

The exclusive presence of linear dynamics in LFP data from area TE may seem paradoxical in view of the

fact that most underlying neural phenomena are known to be non-linear. However, many systems composed of highly non-linear components, e.g. electronic devices, exhibit an overall linear type of behaviour. What has been shown here, is that the system under study can behave in a linear fashion. This finding does not preclude the possibility that different modes of operation would appear in other contexts, e.g. different behavioural demands. Such different modes of behaviour could be regulated by the amount of noise in the system. It is well known that a deterministic non-linear system which is perturbed by an increasingly higher level of stochastic factors may reach the point in which all specific dynamical structure is lost. With lower levels of noise however, this structure can reappear.

Probably by averaging over the signals generated by several neural populations, each exhibiting a different form of non-linear temporal dynamics, a similar effect of hiding the underlying dynamical structures might occur. In this case however, increasing the size of the ensembles recorded from should further reduce the signal's complexity, while non-linearities at the level of the EEG have been reported (Elbert et al., 1994), including the pathophysiological findings presented here (Figs. 1 and 2).

Yet another problem for identifying non-linear brain activity should be considered. Different activity modes might not only be shown by different neuronal populations, but also by the same population in successive intervals. This kind of switching between a coherent oscillatory and a stochastic phase of activity has been found in cat area 17 (Bauer and Pawelzik, 1993; Pawelzik, 1994). By fitting a model to time intervals containing more than one of these phases, the noise of stochastic phases might degrade the process of system identification.

To summarise, the analyses performed in this study should be applied to signals recorded at different spatial scales, preferably from the single cell level to the level of EEG recordings. Second, they could be combined with the above-mentioned methods for identifying phases with different activity patterns. From a pragmatic point of view, linear signals have several advantages over non-linear ones. Maybe the most important one besides reduced computational demands for computing the influence measure is the following. When dealing with linear Autoregressive models, the influence measure can be subjected to a frequency decomposition procedure (Geweke, 1984; Sameshima et al., 1998; Bernasconi and König, 1999). It is then possible to evaluate the contribution of different frequencies to an observed directed influence, which is of special interest in trying to relate influence measures to the frequency content of the signals under study.

#### 4.2. Directions for future work

The present results indicate that the general framework outlined above allows an examination of the non-linearity of neural time series and their interactions. However, these techniques must be refined in several directions.

1. When modelling non-linearity, one is immediately afflicted with the 'curse of dimensionality' something that is quite prominent in non-parametric modelling. Three directions are possible to mitigate this problem: additional structure may be imposed on the non-parametric model by specifying additive or multiplicative relations among sets of lags; subset selection methods can be used to retain only those time lags that have predictive value; specific parametric components may be incorporated into the model in order to decrease the variability of the estimators of non-linear influence as well as to enhance the interpretability of the resulting model.
2. This paper only considers influences between two time series. The results of such an analysis can be misleading since other neural structures may be acting simultaneously on those being measured and this may distort the analysis. The suggestion to use measures of influence which are defined after the effect of other time series is partialled out (Geweke, 1984) is currently being adapted to non-linear Granger causality.
3. Furthermore, attempts are being made to generalise the current framework to include non-stationary neural time series. Another important extension of the presented methods is a generalisation for the analysis of point process time series which would allow for the investigation of trains of single cell spike trains.

#### Acknowledgements

The authors are grateful to Johanna Klon-Lipock and Petra Janson for excellent technical assistance. Detlef Wegener's help with organising the data analysis is gratefully acknowledged as well as the assistance of Sabine Melchert and Sunita Mandon during the literature search. Special thanks are due to the two anonymous referees and to Corrado Bernasconi whose comments during the review process helped us to clarify and generally improve an earlier manuscript. The anatomical MRI scans have been obtained together with Rainer Goebel, Claudia Goebel, Lars Muckli and Matthias H.J. Munk. This work was supported by HFSP Grant RG-20/95 B, 'Oscillatory Event-Related Brain Dynamics', and SFB 517, 'Neurocognition'.

## References

- Abeles M. Local Cortical Circuits. An Electrophysiological Study. Berlin, Heidelberg, New York: Springer Verlag, 1982.
- Abeles M. Corticonics. Cambridge, UK: Cambridge University Press, 1991.
- Aertsen A, Vaadia E, Abeles M, Ahissar E, Bergman H, Karmon B, et al. Neural interactions in the frontal cortex of a behaving monkey — signs of dependence on the stimulus context and behavioral state. *J Hirnforsch* 1991;32(6):735–43.
- Ahissar E, Vaadia E, Ahissar M, Bergman H, Arieli A, Abeles M. Dependence of cortical plasticity on correlated activity of single neurons and behavioral context. *Science* 1992;257:1412–5.
- Akaike H. A new look at the statistical model identification. *IEEE Trans Automatic Control* 1974;19:716–23.
- Bauer H-U, Pawelzik K. Alternating oscillatory and stochastic dynamics in a model for a neuronal assembly. *Physica D* 1993;69:380–93.
- Bell D, Kay J, Malley J. A nonparametric approach to nonlinear causality testing. *Econ Lett* 1996;51:7–18.
- Bell D, Kay J, Malley J. Nonparametric regression and nonlinear causality testing: a Monte-Carlo study. *Scott J Political Econ* 1998;45:528–52.
- Bernasconi C, König P. On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings. *Biol Cybern* 1999;81:199–210.
- Bernasconi C, von Stein A, König P. On the direction of interareal interactions in the cat visual system. *Soc Neurosci Abstr* 1998;24:834.3.
- Boynton RM, Baron WS. Sinusoidal flicker characteristics of primate cones in response to heterochromatic stimuli. *J Opt Soc Am* 1975;65:1091–100.
- Braddick O. Only one speed per object. *Nature* 1996;381:117–8.
- Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall, 1993.
- Elbert T, Ray WJ, Kowalik ZJ, Skinner JE, Graf KE, Birbaumer N. Chaos and physiology — deterministic chaos in excitable cell assemblies. *Physiol Rev* 1994;74(1):1–47.
- Engel AK, König P, Gray CM, Singer W. Stimulus-dependent neuronal oscillations in cat visual cortex — inter-columnar interaction as determined by cross-correlation analysis. *Eur J Neurosci* 1990;2(7):588–606.
- Engel AK, Roelfsema PR, Fries P, Brecht M, Singer W. Role of the temporal domain for response selection and perceptual binding. *Cerebral Cortex* 1997;7:571–82.
- Fan J, Gijbels I. Local polynomial modelling and its applications. London: Chapman and Hall, 1996.
- Fan J, Gijbels J. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J Royal Statist Soc B* 1995;57:371–94.
- Felleman DJ, van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1991;1:1–47.
- Franaszczuk PJ, Blinowska KJ, Kowalczyk M. The application of parametric multichannel spectral estimates in the study of electrical brain activity. *Biol Cybern* 1985;51:239–47.
- Freiwald WA, Kreiter AK, Singer W. Oscillatory and synchronous activity states in the macaque inferotemporal cortex. *Soc Neurosci Abstr* 1998;24:355.15.
- Friston KJ. Functional and effective connectivity — a synthesis. *Hum Brain Mapp* 1994;2:56–78.
- Gersch W. Spectral analysis of EEGs by Autoregressive decomposition of time series. *Math Biosci* 1970;7:205–22.
- Gersch W. Causality or driving in electrophysiological signal analysis. *Math Biosci* 1972;14:177–96.
- Gerstein GL, Aertsen AMHJ. Representation of cooperative firing activity among simultaneously recorded neurons. *J Neurophysiol* 1985;54:1513–28.
- Gerstein GL, Bedenbaugh P, Aertsen AMHJ. Neuronal assemblies. *IEEE Trans Biomed Eng* 1989;36(1):4–14.
- Gerstein GL, Perkel DH, Subramanian KN. Identification of functionally related neural assemblies. *Brain Res* 1978;140:43–62.
- Geweke J. The measurement of linear dependence and feedback between multiple time series. *J Am Stat Assoc* 1982;77:304–13.
- Geweke J. Measures of conditional linear dependence and feedback between time series. *J Am Stat Assoc* 1984;79:907–15.
- Gochin PM, Miller EK, Gross CG, Gerstein GL. Functional interactions among neurons in inferior temporal cortex of the awake macaque. *Exp Brain Res* 1991;84:505–16.
- Goebel R. Brain Voyager: a program for analyzing and visualizing functional and structural MRI data sets. *NeuroImage* 1996;3:604.
- Goodale MA, Milner AD. Separate visual pathways for perception and action. *Trends Neurosci* 1992;15:20–5.
- Granger CWJ. Economic processes involving feedback. *Inf Control* 1963;6:28–48.
- Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 1969;37:424–38.
- Granger CWJ. Testing for causality — a personal viewpoint. *J Econ Dyn Control* 1980;2:329–52.
- Granger CWJ, Lin JL. Causality in the long run. *Econom Theory* 1995;11:530–6.
- Granger CWJ, Newbold P. Forecasting economic time series. New York: Academic Press, 1977.
- Hastie T, Tibshirani R. Generalised additive models. London: Chapman and Hall, 1990.
- Hebb DO. The organization of behavior. New York: Wiley, 1949.
- Hernández JL, Valdés PA, Vila P. EEG spike and wave modelled by a stochastic limit cycle. *NeuroReport* 1996;7:2246–50.
- Hilgetag C-C, O'Neill MA, Young MP. Indeterminate organization of the visual system. *Science* 1996;271:776–7.
- Hupé JM, James AC, Payne BR, Lomber SG, Girard P, Bullier J. Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature* 1998;394:784–7.
- Johannesma P, Aertsen A, van den Boogard H, Eggemont J, Epping W. From synchrony to harmony — ideas on the function of neural assemblies and on the interpretation of neural synchrony. In: Palm G, Aertsen A, editors. *Brain Theory*. Berlin, Heidelberg: Springer-Verlag, 1986:25–47.
- Kaas JH, Krubitzer LA. The organization of extrastriate visual cortex. In: Dreher B, Robinson SR, editors. *Vision and visual dysfunction*. Macmillan Press, 1991:302–23.
- Koch C, Rapp M, Segev I. A brief history of time (constants). *Cereb Cortex* 1996;6:93–101.
- Kooperberg C, Stone CJ. A study of logspline density estimation. *Comput Stat Data Anal* 1991;12:327–48.
- Kreiter AK, Singer W. Oscillatory neuronal responses in the visual cortex of the awake macaque monkey. *Eur J Neurosci* 1992;4:369–75.
- Krüger J, Aiple F. Multimicroelectrode investigation of monkey striate cortex — spike train correlations in the infragranular layers. *J Neurophysiol* 1988;60(2):798–828.
- Lopes da Silva FH, Mars NJI. Parametric methods in EEG analysis. In: Gevins AS, Rémond A, editors. *Methods of analysis of brain electrical and magnetic signals*. EEG handbook, vol. 1. Elsevier, 1987:243–60.
- Mammen E. Resampling methods for non and semiparametric regression. In: Schimek MG, editor. *Smoothing and regression approaches, computations and applications*. New York: Wiley, 1999.
- Marron JS. Bootstrap bandwidth selection. In: LePage R, Billard L, editors. *Exploring the limits of the bootstrap*. New York: Wiley, 1992.
- Martin KAC. A brief history of the 'feature detector'. *Cereb Cortex* 1994;4(1):1–7.



- McIntosh AR, Bookstein FL, Haxby JV, Grady CL. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage* 1996;3:143–57.
- McIntosh AR, Gonzalez-Lima F. Structural equation modeling and its application to network analysis of functional brain imaging. *Hum Brain Mapp* 1994;2:2–22.
- Miyashita Y, Higuchi S-I, Sakai K, Masui N. Generation of fractal patterns for probing the visual memory. *Neurosci Res* 1991;12:307–11.
- Nowak LG, Bullier J. The timing of information transfer in the visual system. In: Rockland KS, Kaas JH, Peters A, editors. *Cerebral Cortex*, vol. 12. New York: Plenum Press, 1997:205–41.
- Ozaki T. Non-linear time series models and dynamical systems. In: Hannan AJ, Krishnaiah PR, Rao MM, editors. *Time series in the time domain*, Handbook of Statistics, vol. 5. North-Holland, 1985:25–83.
- Pawelzik K. Detecting coherence in neuronal data. In: Domany E, van Hemmen JL, Schultz W, editors. *Models of Neural Networks II*. New York: Springer-Verlag, 1994:253–85.
- Pijn JPM. Quantitative evaluation of EEG signals in epilepsy; nonlinear associations, time delays and nonlinear dynamics, 1990; Rodopi, (Thesis).
- Prut Y, Vaadia E, Bergman H, Haalman I, Slovin H, Abeles M. Spatiotemporal structure of cortical activity: properties and behavioral relevance. *J Neurophysiol* 1998;79:2857–74.
- Rockland KS, van Hoesen GW. Direct temporal-occipital feedback connections to striate cortex (v1) in the macaque monkey. *Cereb Cortex* 1994;4:300–13.
- Sameshima K, Baccalá LA, Ballester G, Valle AC, Timo-Iaria C. Statistical testing in probing causality in the interaction of brain structures. *Soc Neurosci Abstr* 1998;24(1):133.
- Scannell JW, Blakemore C, Young MP. Analysis of connectivity in the cat cerebral cortex. *J Neurosci* 1995;15:1463–583.
- Schillen TB, König P, Löwel S, Singer W. Assessing the range of local field potentials through ocular dominance and orientation maps of cat visual cortex. In: *Proceedings of the 20th Göttingen Neurobiology Conference*, (Abstract).
- Singer W, Gray C, Engel A, König P, Artola A, Bröcher S. Formation of cortical cell assemblies. *Symp Quant Biol* 1990;55:939–52.
- Singer W, Gray CM. Visual feature integration and the temporal correlation hypothesis. *Annu Rev Neurosci* 1995;18:555–86.
- Singer W, Phillips WA. In search of common foundations for cortical computation. *Behav Brain Sci* 1997;20:657–83.
- Sporns O, Tononi G, Edelman GM. Modelling perceptual grouping and figure-ground segregation by means of active reentrant connections. *Proc Natl Acad Sci USA* 1991;88:129–33.
- Teräsvirta T. Modeling economic relationships with smooth transition regressions. In: Ullah A, Gilles DEA, editors. *Handbook of applied economic statistics*. New York: Marcel Dekker, 1998.
- Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. *Nature* 1996;38:520–2.
- Tong H. *Nonlinear time series: a dynamical system approach*. Oxford: Clarendon Press, 1990.
- Ts'o DY, Gilbert CD, Wiesel TN. Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. *J Neurosci* 1986;6(4):1160–70.
- Ts'o DY, Gilbert DC. The organization of chromatic and spatial interactions in the primate striate cortex. *J Neurosci* 1988;8:1712–27.
- Ungerleider LG, Mishkin M. Two cortical visual systems. In: Ingle DJ, Goodale MA, Mansfield RJW, editors. *Analysis of visual behavior*. Cambridge: MIT Press, 1982:549–86.
- Valdes P, Bosch J, Jimenez JC, Trujillo N, Biscay R, Morales F, et al. The statistical identification of nonlinear brain dynamics — a progress report. In: Pradhan N, Rapp PE, Sreenivasan R, editors. *Nonlinear Dynamics and Brain Functioning*. Nova Science, 1999.
- Valdes PA, Bosch J, Biscay R, Jimenez JC, Virues T, Macías F, et al. Nonlinear measures of influences between EEG sources. *NeuroImage* 1996;7:S634.
- van Essen DC, Anderson CH, Felleman DJ. Information processing in the primate visual system: an integrated systems perspective. *Science* 1992;255:419–23.
- von der Malsburg C. The correlation theory of brain function. *Int Rep MPI Biophys Chem* 1981;81–2:1–40.
- Warne A. *Causality in a Markov switching VAR*, Institute for International Economic Studies, Stockholm University, Sweden. 1999; <http://www.iies.su.se/awarne>.
- Wurtz RH. Visual receptive fields of striate cortex neurons in awake monkeys. *J Neurophysiol* 1969;32:727–42.
- Yao Q, Tong H. On subset selection in nonparametric stochastic regression. *Stat Sin* 1994;4:51–70.

**Decomposing EEG data into space-time-frequency components using parallel factor analysis**

## Decomposing EEG data into space–time–frequency components using Parallel Factor Analysis

Fumikazu Miwakeichi,<sup>a,\*</sup> Eduardo Martínez-Montes,<sup>b</sup> Pedro A. Valdés-Sosa,<sup>b</sup> Nobuaki Nishiyama,<sup>a</sup> Hiroaki Mizuhara,<sup>a</sup> and Yoko Yamaguchi<sup>a</sup>

<sup>a</sup>Laboratory for Dynamics of Emergent Intelligence, RIKEN Brain Science Institute, Saitama 351-0198, Japan

<sup>b</sup>Neuroscience Department, Cuban Neuroscience Center, Habana, Cuba

Received 17 July 2003; revised 12 March 2004; accepted 17 March 2004

Finding the means to efficiently summarize electroencephalographic data has been a long-standing problem in electrophysiology. A popular approach is identification of component modes on the basis of the time-varying spectrum of multichannel EEG recordings—in other words, a space/frequency/time atomic decomposition of the time-varying EEG spectrum. Previous work has been limited to only two of these dimensions. Principal Component Analysis (PCA) and Independent Component Analysis (ICA) have been used to create space/time decompositions; suffering an inherent lack of uniqueness that is overcome only by imposing constraints of orthogonality or independence of atoms. Conventional frequency/time decompositions ignore the spatial aspects of the EEG. Framing of the data being as a three-way array indexed by channel, frequency, and time allows the application of a unique decomposition that is known as Parallel Factor Analysis (PARAFAC). Each atom is the tri-linear decomposition into a spatial, spectral, and temporal signature. We applied this decomposition to the EEG recordings of five subjects during the resting state and during mental arithmetic. Common to all subjects were two atoms with spectral signatures whose peaks were in the theta and alpha range. These signatures were modulated by physiological state, increasing during the resting stage for alpha and during mental arithmetic for theta. Furthermore, we describe a new method (Source Spectra Imaging or SSI) to estimate the location of electric current sources from the EEG spectrum. The topography of the theta atom is frontal and the maximum of the corresponding SSI solution is in the anterior frontal cortex. The topography of the alpha atom is occipital with maximum of the SSI solution in the visual cortex. We show that the proposed decomposition can be used to search for activity with a given spectral and topographic profile in new recordings, and that the method may be useful for artifact recognition and removal.

© 2004 Elsevier Inc. All rights reserved.

**Keywords:** Parallel Factor Analysis; EEG space/frequency/time decomposition; Principal Component Analysis; Multiway analysis; Source Spectra Imaging

### Introduction

The electroencephalogram (EEG) is the reflection upon the scalp of the summed synaptic potentials of millions of neurons (Lopes da Silva, 1987). Most investigators agree (Lachaux et al., 1999; Varela et al., 2001) that these neurons self-organize into transient networks (“neural masses”) that synchronize in time and space to produce a mixture of short bursts of oscillations that are observable in the EEG record. A statistical description of the oscillatory phenomena of the EEG was carried out first in the frequency domain (Lopes da Silva, 1987) by estimation of the power spectrum for quasi-stationary segments of data. More recent characterizations of transient oscillations are carried out by estimation of the time-varying (or evolutionary) spectrum in the frequency/time domain (Dahlhaus, 1997). These evolutionary spectra of EEG oscillations will have a topographic distribution on the sensors that is contingent on the spatial configuration of the neural sources that generate them as well as the properties of the head as a volume conductor (Nunez, 1993).

The purpose of the present study was to attempt the decomposition of multichannel time-varying EEG spectrum into a series of distinct components or modes. In the parlance of modern harmonic analysis (Chen and Donoho, 2001), we performed a space/frequency/time “atomic decomposition” of multidimensional data. In other words, we assume that each neural mass contributes a distinctive atom to the topographic frequency/time description of the EEG, so that the estimation of these atoms is possible by means of signal-processing techniques. Each atom will be defined by its topography, spectral content, and time profile; in other words, by its spatial, spectral, and temporal signatures. We expect that these extracted atoms ultimately will allow the identification of the corresponding neural masses that produce them.

There is a long history of atomic decompositions for the EEG. However, to date, atoms have not been defined by the triplet spatial, spectral, and temporal signatures but rather pairwise combinations of these components. Some of the current procedures for these analyses are reviewed below.

### Space/time atoms: PCA and ICA

Space/time atoms are the basis of both Principal Component Analysis (PCA) and Independent Component Analysis (ICA) as

---

\* Corresponding author. Laboratory for Dynamics of Emergent Intelligence, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan. Fax: +81-48-467-6938.

E-mail address: miwakel@brain.riken.go.jp (F. Miwakeichi).

Available online on ScienceDirect (www.sciencedirect.com.)

applied to multichannel EEG. PCA has been used for artifact removal and to extract significant activities in the EEG (Lagerlund et al., 1997; Soong and Koles, 1995). A basic problem is that atoms defined by only two signatures (space and time) are not determined uniquely. In PCA, orthogonality is therefore imposed between the corresponding signatures of different atoms. This, however, is a rather nonphysiological constraint. Even with this restriction, there is the well-known nonuniqueness of PCA that allows the arbitrary choice of rotation of axes (e.g., Varimax and Quartimax rotations). More recently, ICA has become a popular tool for space/time atomic decomposition (Cichocki and Amari, 2002; Hyvarinen et al., 2001). It avoids the arbitrary choice of rotation (Jung et al., 2001). Uniqueness, however, is achieved at the price of imposing a constraint even stronger than orthogonality, namely, statistical independence. In both PCA and ICA, the frequency information may be obtained from the temporal signature of the extracted atoms in a separate step.

#### Frequency/time atoms: wavelet analysis

There are many papers on the decomposition of single-channel EEG into frequency/time atoms. For this purpose, the Fast Fourier Transformation (FFT) with sliding window (Makeig, 1993) or the wavelet transformation (Bertrand et al., 1994; Tallon-Baudry et al., 1997) have been employed. In fact, any of the frequency/time atomic decompositions currently available (Chen and Donoho, 2001) could, in principle, be used for the EEG. However, these methods do not address the topographic aspects of the EEG time/frequency analysis.

#### Space/frequency/time atoms: PARAFAC

Gonzalez Andino et al. (2001) improved previous analyses by analyzing regions of the frequency/time plane where a single dipole model is an adequate spatial description of the signal, thus incorporating topographic information. Topographic frequency/time decomposition of the EEG was introduced by Koenig et al. (2001), which is the first work to estimate atoms characterized simultaneously by a frequency/time and spatial signature. In their analyses, it was possible to extract physiologically significant activity in the EEG. However, in order to achieve a unique decomposition, they imposed the mathematical constraints that the combined frequency/time signatures of all atoms were required to be of minimum norm and the spatial or topographic signatures were required to have maximal smoothness. These constraints have been found to be unnecessary for unique topographic time/frequency decomposition, a fact that has motivated the work described in this paper.

It has long been known, especially in the chemometrics literature, that unique multi-linear decompositions of multi-way arrays of data (more than two dimensions) are possible under very weak conditions (Sidiropoulos and Bro, 2000). In fact, this is the basic argument for Parallel Factor Analysis (PARAFAC). This technique was proposed independently by Harshman (1970, 1972) and by Carroll and Chang (1970) who named the model Canonical Decomposition, and recently has been improved by Bro (1998) who also provided a Matlab toolbox (available as of this writing at: <http://www.models.kvl.dk/users/rasmus/>). In PARAFAC, for three-way arrays, the data is decomposed as a sum of components (corresponding to our “atoms”), each of which is the tri-linear product of one score vector and two loading vectors (corresponding to our “signatures”). The important difference

between PARAFAC and techniques such as PCA or ICA is that the decomposition of multi-way data is unique even without additional orthogonality or independence constraints.

Thus, PARAFAC can be employed for a space/frequency/time atomic decomposition of the EEG. This makes use of the fact that multichannel evolutionary spectra are multi-way arrays, indexed by electrode, frequency, and time. The inherent uniqueness of the PARAFAC solution leads to a topographic time/frequency decomposition with a minimum of *a priori* assumptions.

Here, we use PARAFAC for the purpose of simultaneous space/frequency/time decompositions. Previous applications of PARAFAC to EEG data have analyzed only space/time, and some additional external dimensions provided by subject and drug dose, among other factors (Achim and Bouchard, 1997; Estienne et al., 2001; Field and Graupe, 1991). A special interpretation of this model is also the Topographic Components Model (TCM) (Möcks, 1988a,b), which gives justification for the PARAFAC model in the context of evoked potentials analysis, based on biophysical considerations (Möcks, 1988b). In this field, a relevant proof of the use of TCM over PCA using only synthetic noiseless data was given in Achim and Bouchard (1997).

To illustrate the usefulness of PARAFAC, we applied the decomposition of time-varying EEG spectrum to the comparison of resting EEG to that recorded while the subject performed mental arithmetic. Mental arithmetic produces theta activity in the frontal area and a suppression of alpha activity in the occipital area, while the converse occurs when the eyes are closed in the resting condition (Harmony et al., 1999; Ishihara and Yoshii, 1972; Sasaki et al., 1996). The PARAFAC atomic decomposition should be able to extract these components, localize them correctly, and detect the corresponding level of activity in these bands in each physiological state. Once estimated, the spatial and spectral signatures of the identified atoms may be used to search for similar types of activity in new data sets. Here, this procedure will be called “screening” for the presence of an atom.

Our focus is on space/time/frequency decompositions tailored to the description of oscillatory phenomena. These are not the only interesting phenomena in the EEG, transient activity being another example. The methods described in this paper may be generalized to this application by exchanging the basic dictionary that describes oscillations.

This paper is organized as follows. We first describe the experimental methods. Then, we consider the basic theoretical development of the space/frequency/time atomic decomposition

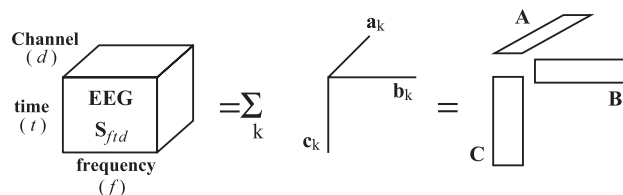


Fig. 1. Graphical explanation of the PARAFAC model. The multichannel EEG evolutionary spectrum  $S$  is obtained from a channel by channel wavelet transform.  $S$  is a three-way data array indicated by channel, frequency, and time. PARAFAC decomposes this array into the sum of “atoms”. The  $k$ th atom is the tri-linear product of loading vectors representing spatial ( $\mathbf{a}_k$ ), spectral ( $\mathbf{b}_k$ ), and temporal ( $\mathbf{c}_k$ ) “signatures”. Under these conditions, PARAFAC can be summarized as finding the matrices  $\mathbf{A} = \{\mathbf{a}_k\}$ ,  $\mathbf{B} = \{\mathbf{b}_k\}$ , and  $\mathbf{C} = \{\mathbf{c}_k\}$ , which explain  $S$  with minimal residual error.

and the use of estimated factors to screen for activity in new data segments. The results and discussion follow.

**Methods**

*Data acquisition*

Five male right-handed subjects (mean age  $25.8 \pm 3.96$  years) that produced clear theta activities during a mental task were studied in this work. All subjects signed an informed consent form approved by the RIKEN Human Subject Protection Committee before EEG recording. All subjects were required to concentrate, for 3 min, on mental arithmetic (subtraction by serial 7 from 1000) with closed eyes. They were asked the final residual number at the end of the task. The resting EEG with closed eyes was also recorded for comparison. During the recording, we provided no visual nor auditory stimulation for the subjects.

EEG recordings were carried out with a standard 64-channel system (NeuroScan Syn Amps Model 5083) referred to linked

earlobes. The EEG data were sampled at 500 Hz and bandpass filtered from 1 to 30 Hz.

*Theory*

In our application to EEG data, the data matrix  $S_{(N_d \times N_f \times N_t)}$  is the three-way time-varying EEG spectrum array obtained by using the wavelet transformation, where  $N_d$ ,  $N_f$ , and  $N_t$  are the number of channels, steps of frequency, and time points, respectively. For the wavelet transformation, a complex Morlet mother function was used (Jensen and Tesche, 2002; Kronland-Martinet and Morlet, 1987; Tallon-Baudry et al., 1997). The energy  $S(d, f, t)$  of the channel  $d$  at frequency  $f$  and time  $t$  is given by the squared norm of the convolution of a Morlet wavelet with the EEG signal  $v(d, t)$

$$S(d, f, t) = |w(t, f) * v(d, t)|^2, \tag{1}$$

where the complex Morlet wavelet,  $w(t, f)$  is defined by  $w(t, f) = \sqrt{\pi} \sigma_b \exp\left(-\left(\frac{t}{\sigma_b}\right)^2\right) * \exp(i2\pi ft)$  with  $\sigma_b$  being the

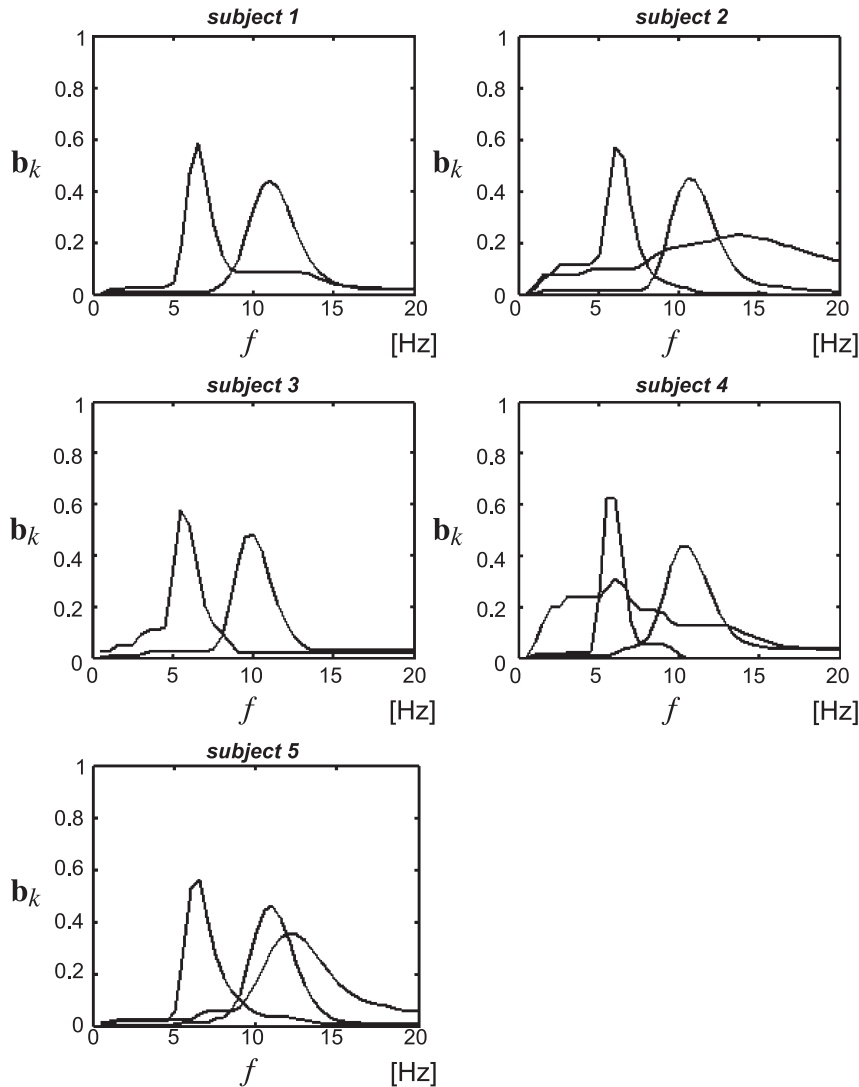


Fig. 2. Spectral signature  $b_k$  of atoms of Parallel Factor Analysis (PARAFAC) for each subject. Note the recurrent appearance of frequency peaks in the theta and alpha bands. The horizontal axis is frequency in Hz and the vertical axis is the normalized amplitude.

bandwidth parameter. The width of the wavelet,  $m = 2\pi\sigma_b f$  is set to 7 in this study.

We closely follow here the detailed description of PARAFAC found in Bro (1998). The basic structural model for a PARAFAC decomposition of the data matrix  $\mathbf{S}_{(N_d \times N_f \times N_t)}$  of elements  $S_{dft}$  is defined as:

$$\hat{S}_{dft} = \sum_{k=1}^{N_k} a_{dk} b_{fk} c_{tk} \quad (2)$$

The problem is to find the loading matrices,  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$ , whose elements are  $a_{dk}$ ,  $b_{fk}$ , and  $c_{tk}$ . In our application, each component  $k$  will be designated as an “atom” and the corresponding vectors  $\mathbf{a}_k = \{a_{dk}\}$ ,  $\mathbf{b}_k = \{b_{fk}\}$ ,  $\mathbf{c}_k = \{c_{tk}\}$  will be the spatial, spectral, and temporal signatures of each atom (Fig. 1). The uniqueness of the solution is guaranteed when  $\text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B}) + \text{rank}(\mathbf{C}) \geq 2N_k + 2$ . As can be seen, this is a less-stringent condition than either orthogonality or statistical independence (Sidiropoulos and Bro, 2000). The decomposition (Eq. (2)) is achieved by finding  $\min_{a_{dk} b_{fk} c_{tk}} \|\hat{S}_{dft} - \sum_{k=1}^{N_k} a_{dk} b_{fk} c_{tk}\|$ . Since the  $\hat{S}_{dft}$  are spectra, this minimization must be carried out under the non-negativity constraint for the loading vectors. This particular variant of PARAFAC has been developed by Bro (1998). PARAFAC produces the vectors  $\mathbf{a}_{k(N_d \times 1)}$ , which is the  $k$ th component loading vector that can be seen as topographical maps,  $\mathbf{b}_{k(N_f \times 1)}$  is the spectrum for  $k$ th component and  $\mathbf{c}_{k(N_t \times 1)}$  is the temporal signature for component  $k$ .

The main advantage of this method is that it provides us with a unique decomposition of the time-varying EEG spectrum corresponding to the best model in the least-squares sense. The only indeterminacy in the least-square solution is the order of the atoms and the relative scaling of the signatures. On the other hand, it has also been proved that if the data is approximately tri-linear, then the algorithm will show the true underlying phenomena, if the correct number of components is used and if the signal-to-noise ratio is appropriate (Harshman, 1972; Kruskal, 1976, 1997).

An important point is the selection of the most appropriate number  $N_k$  of components. Several methods have been developed for this purpose only (Bro, 1998). The Core Consistency Diagnostic (Corcondia) is an approach that applies especially to PARAFAC models, and has been shown to be a powerful and simple tool for determining the appropriate number of components in multiway models. In this work, we will use not only Corcondia but also the evaluation of systematic variation left in the residuals of the model.

#### Validation of the method

As described by Harshman (1984) and Bro (1998), the validation of a particular analysis can be seen as part of the analysis itself and can be divided into levels: zero-fit diagnostics (related to data before fitting any model, selection of proper model); one-fit diagnostics (validate the consistency of the model applied); many-fit diagnostics (comparisons between different models, use of statistical inferences on the results). Given some general considerations of the PARAFAC modeling of the time-varying EEG spectrum, we shall make a deeper analysis of the appropriateness of this procedure following these levels and the general guidelines for validating the application of multi-way models given in Bro (1998).

The usual way to assess the multiway (three-way in this case) nature of the data in study is the exploration of results provided by bi-linear analysis of the data. In particular, the application of PCA

to the unfolded three-way array will provide a matrix of loadings in which a global behavior can be detected, indicating the existence of dimensional structure in the explored dimensions. For data similar to those treated here, this is clearly shown in Estienne et al. (2001), and with a more complete analysis in Field and Graupe (1991). Another way of assessing the tri-linear structure of the data is by means of the Core Consistency Diagnostic (Corcondia) (Bro, 1998; Estienne et al., 2001). This is a tool provided automatically in the implementation of PARAFAC and other related multi-way algorithms contained in the Matlab Toolbox used here. Corcondia was utilized for successfully demonstrating the presence of multiway structure in our data.

In this work, we have chosen PARAFAC among several multi-way models, (e.g., PARAFAC2, PARATUCK2, TUCKER1, TUCKER3) (Bro, 1998). This is an application of Occam’s razor as PARAFAC is the simplest and most restricted model. As we only consider it in our analysis, whether other versions of PARAFAC or TUCKER models can give better results in terms of explanation of the systematic variation of the data and interpretability of the results remains an open question.

On the other hand, other drawbacks of the application of PARAFAC model include the need for careful preprocessing of the data and for checking residuals, leverages and other parameters in the search of constant factors, outliers, and degeneracy. We do not detail these problems here, as such analyses appear in the literature, e.g., exhaustive ways of exploring degeneracy and model mis-specifications can be found in Field and Graupe (1991). The Matlab Toolbox used here provided us of these tools for many-fit diagnostics (residuals, leverages, Corcondia, convergence, explained variation).

What is missing in the present study is a rigorous analysis of the uniqueness of the model in our case, but it is well-known that through the use of PARAFAC, uniqueness is almost always present.

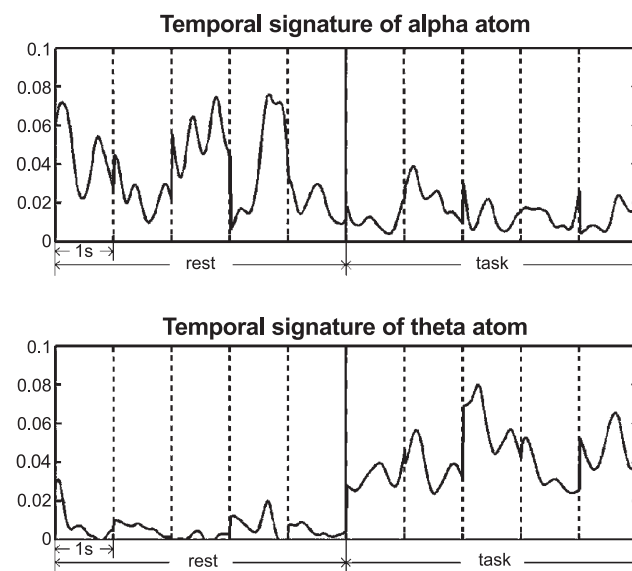


Fig. 3. Temporal signatures,  $\mathbf{c}_k$ , of theta and alpha atoms of Parallel Factor Analysis (PARAFAC) for a typical subject. The first five segments were chosen randomly from the rest condition; the second five segments were selected so as to contain typical theta bursts. Each segment is 1-s long, containing 100 time frames. The horizontal axis is time  $t$ , and the vertical axis is the value of  $\mathbf{c}_k$ , which is dimensionless.

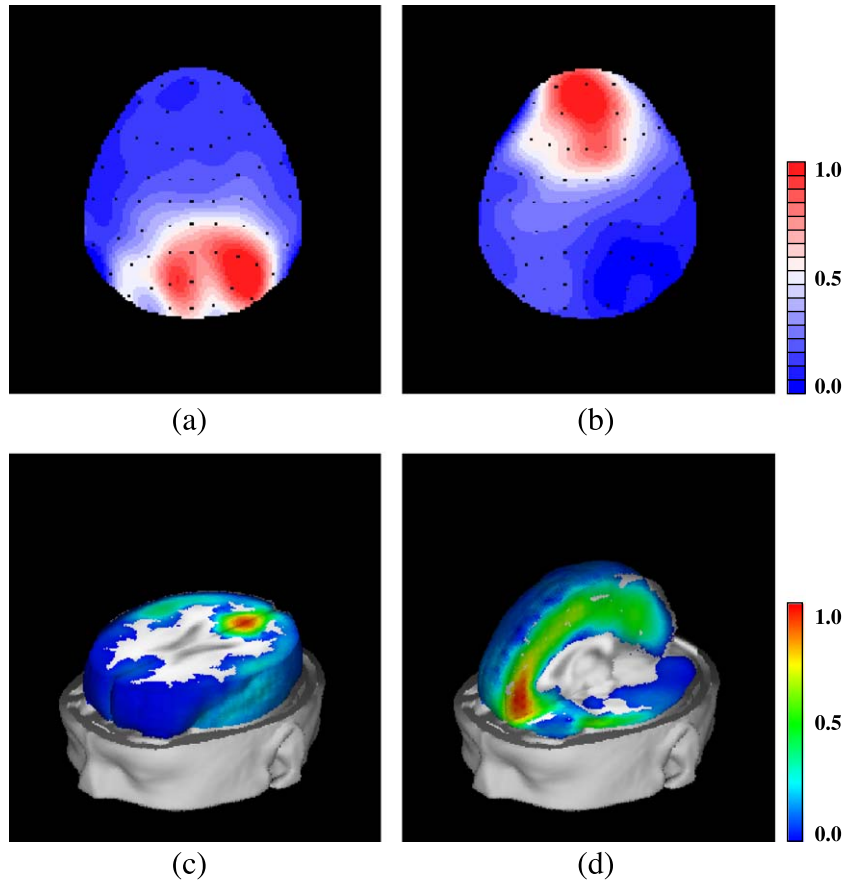


Fig. 4. Spatial signatures,  $\mathbf{a}_k$ , for the theta and alpha atoms of Parallel Factor Analysis (PARAFAC) of a typical subject. In (a) and (b), each signature is displayed as a topographic map; (c) and (d) are the corresponding Source Spectra Imaging solutions. The cross sections of brain were prepared for better visualization of the maximally activated regions. These are illustrated with a normalized color scale of the magnitude of  $\mathbf{J}_k$ .

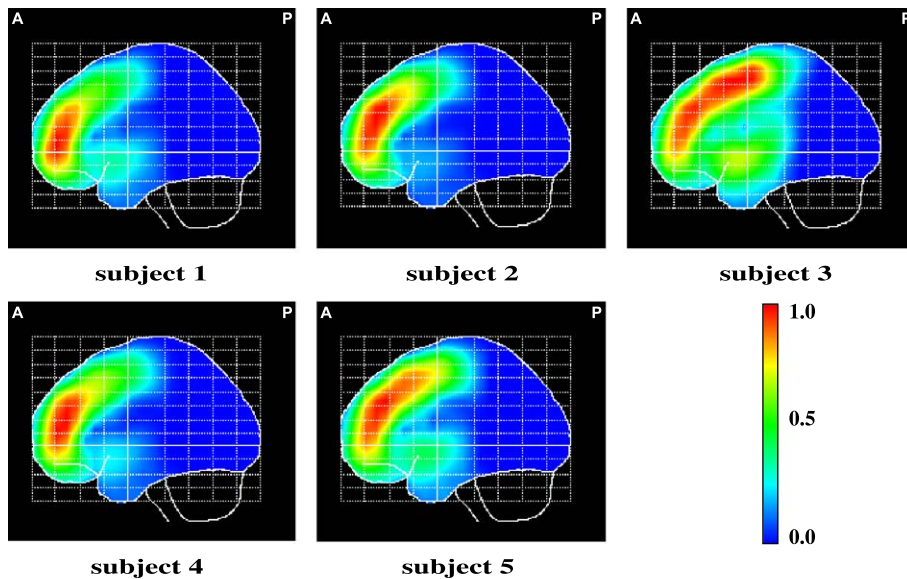


Fig. 5. Maximum-intensity projections of the Source Spectra Imaging solution of the spatial signature,  $\mathbf{a}_k$ , of the theta atom for each subject.

Although a sufficient condition was given above, usually, uniqueness can be assessed by checking the convergence of the algorithm and the interpretability of the results. A strong sufficient condition, but easier to verify, is that no two loading vectors are linearly dependent.

#### Screening and artifact detection

PARAFAC can be used not only for extracting significant activities from EEG, but also for searching for the presence of atoms in a new data set, which were not used for estimating the loadings and can be either from the same or from a different subject. If the spatial and spectral signatures of an atom are fixed, they can be used as templates for screening. Formally, after estimating atoms in a training data set, this can be reconstructed as

$$\hat{S} = \mathbf{C}(\mathbf{B} | \otimes | \mathbf{A})^T, \quad (3)$$

here,  $\mathbf{B} | \otimes | \mathbf{A} = [\mathbf{b}_1 \otimes \mathbf{a}_1 \mathbf{b}_2 \otimes \mathbf{a}_2 \dots \mathbf{b}_{nk} \otimes \mathbf{a}_{nk}]$  is the Khatri-Rao product of  $\mathbf{B}$  and  $\mathbf{A}$ , and represents the convolution of space and frequency.  $\mathbf{X}^T$  denotes the transpose of matrix  $\mathbf{X}$ . For the definition of a template, not all atoms are necessary; that is, for the sake of screening, atoms that are not of interest may be eliminated. Let  $\mathbf{B}'$  and  $\mathbf{A}'$  be fixed spectral and spatial signatures (with some atoms

possibly eliminated). The temporal signature,  $\mathbf{C}'$ , can then be estimated by using least squares in a new data set  $\mathbf{X}$ ,

$$\mathbf{C}'^T = (\mathbf{P}^T \mathbf{P})^{-1} \mathbf{P}^T \mathbf{X}; \quad \mathbf{P} = (\mathbf{B}' | \otimes | \mathbf{A}')^T \quad (4)$$

$\mathbf{P}$  can be regarded as a template for screening for the presence of the atoms of interest. The new temporal signature,  $\mathbf{C}'$ , will then be an estimate of the detected activities corresponding to each atom in the new data set.

If certain atoms obtained by PARAFAC decomposition contain artifact (e.g., eye movements, eye blinking, electromyogram, etc. . .) their space/frequency reconstructions can be used as templates for an artifact detector. The reconstruction, obtained by eliminating the component that corresponds to artifacts, will be an artifact removal method.

#### Inverse solution for the spatial signatures of atoms: source spectra imaging

Each column  $\mathbf{a}_k$  of matrix  $\mathbf{A}$  can be seen as the topography of atom  $k$ . Thus, it would be desirable to obtain the sources inside the brain that can produce these topographies to highlight more precise anatomical details. The difficulty here is that the spatial signatures are all positive values, as they are the differential topographic

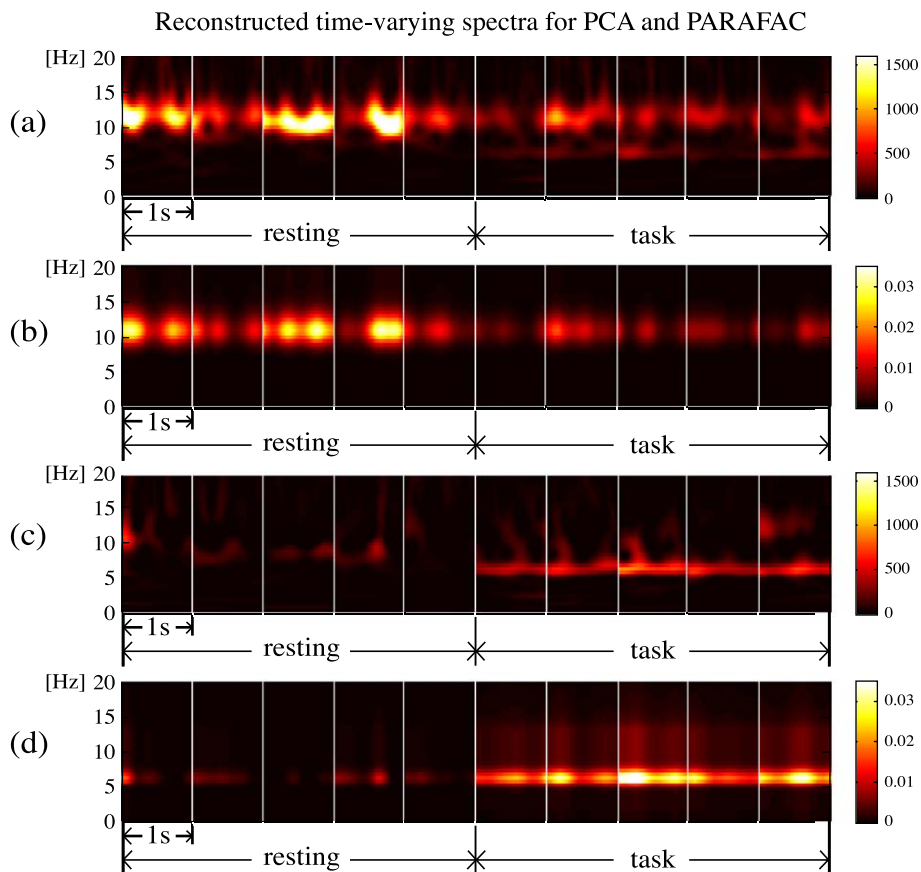


Fig. 6. Panels (a) and (b) illustrate the reconstructed decomposed component by Varimax rotated Principal Component Analysis (PCA) corresponding to the two largest eigenvalues. Panels (c) and (d) are the reconstructed alpha and theta atoms of PARAFAC decomposition in the frequency/time plane, respectively. The first five segments were randomly selected from the resting condition; the second five segments were selected so as to contain typical theta bursts. Each segment is 1-s long and contains 100 time frames.



profiles of EEG spectra, i.e., the variances of complex Fourier coefficients, and therefore the inverse solutions in this case are not the ordinary estimates for the current densities. However, a simple exploratory analysis can be performed in which, under certain assumptions and simplifications, the underlying sources for these topographies can be obtained by an inverse solution. Furthermore, these sources will be shown to be the spectra of electric current densities.

The well-known relation between the electric current densities inside the brain and the electric potentials measured by a set of electrodes on the scalp is:

$$\mathbf{V}_f = \mathbf{K}\mathbf{J}_f \quad (5)$$

Here we have written Eq. (5) directly in the frequency domain, i.e.,  $\mathbf{V}_f$  ( $N_d \times 1$ ) and  $\mathbf{J}_f$  ( $3N_v \times 1$ ) are the vectors of Fourier coefficients of the voltages and electric current density time series, respectively.  $N_v$  is the number of voxels of a regular grid inside the brain. The matrix  $\mathbf{K}$  ( $N_d \times 3N_v$ ) is the electric lead field, which is unaffected by the Fourier transformation. As the absolute value of the electric potential has no physical meaning, the average value of voltages was taken as the reference. From Eq. (5), we can find the spectra of voltages as  $\alpha = \text{diag}(\mathbf{V}_f \mathbf{V}_f^*)$ :

$$\alpha = \text{diag}(\mathbf{K}\mathbf{J}_f\mathbf{J}_f^*\mathbf{K}^T) \quad (6)$$

If we assume that there is no correlation between the current densities in different voxels, i.e.,  $\mathbf{J}_f\mathbf{J}_f^*$  is a diagonal matrix, we can obtain their spectra as  $\gamma = \text{diag}(\mathbf{J}_f\mathbf{J}_f^*)$ . Eq. (6) then becomes:

$$\alpha = \mathbf{K}^{\wedge 2}\boldsymbol{\gamma} \quad (7)$$

where  $\mathbf{K}^{\wedge 2}$  indicates the operation of squaring each element of the matrix,  $\mathbf{K}$ . This represents a linear relation between  $\alpha$  (spatial

signatures obtained by PARAFAC decomposition), and the spectra of current sources that generate the scalp voltages. For the sake of simplicity, it may be assumed that the spectrum vector of the current density has the same magnitude in all directions for each voxel. Therefore, Eq. (7) may be rewritten as:

$$\alpha = \mathbf{M}\boldsymbol{\mu} \quad (8)$$

Here, matrix  $\mathbf{M}$  ( $N_d \times N_v$ ) was obtained by averaging every three columns of matrix  $\mathbf{K}^{\wedge 2}$ , and  $\boldsymbol{\mu}$  ( $N_v \times 1$ ) is the spectrum of current densities for each voxel.

From Eq. (8), the spectra of current sources can be found by an inverse solution procedure. Note here that spectra  $\alpha$  and  $\boldsymbol{\mu}$  are non-negative vectors, allowing us to solve Eq. (8) as a minimum least squares problem under the non-negativity constraint for  $\boldsymbol{\mu}$ . Eq. (8) is undetermined; thus, we shall constrain the solution to be the smoothest one. In this case, the underlying sources ( $\mu_k$ ) for the topographic signature ( $\mathbf{a}_k$ ) of atom  $k$ , can be obtained from

$$\boldsymbol{\mu}_k = \arg \min_{\mu_k \geq 0} \|\mathbf{a}_k - \mathbf{M}\boldsymbol{\mu}_k\|^2 + \lambda \|\mathbf{L}\boldsymbol{\mu}_k\|^2 \quad (9)$$

where  $\mathbf{L}$  is the discrete Laplacian operator as described in Pascual-Marqui et al. (1994) and  $\lambda$  is a regularization parameter.

In general, we shall call this approximation to the source reconstruction for spectra of voltages the ‘‘Source Spectra Imaging’’ (SSI) solution. Despite the assumption of independence of electric current densities, finding the SSI solution for the spatial signatures implies the a priori assumption of spatial smoothness of the spectrum of current densities. Moreover, for this work, the SSI solution for each atom was obtained by imposing anatomical constraints using the Montreal Neurological Institute Probabilistic Brain Atlas as described in Casanova et al. (2000). This set of

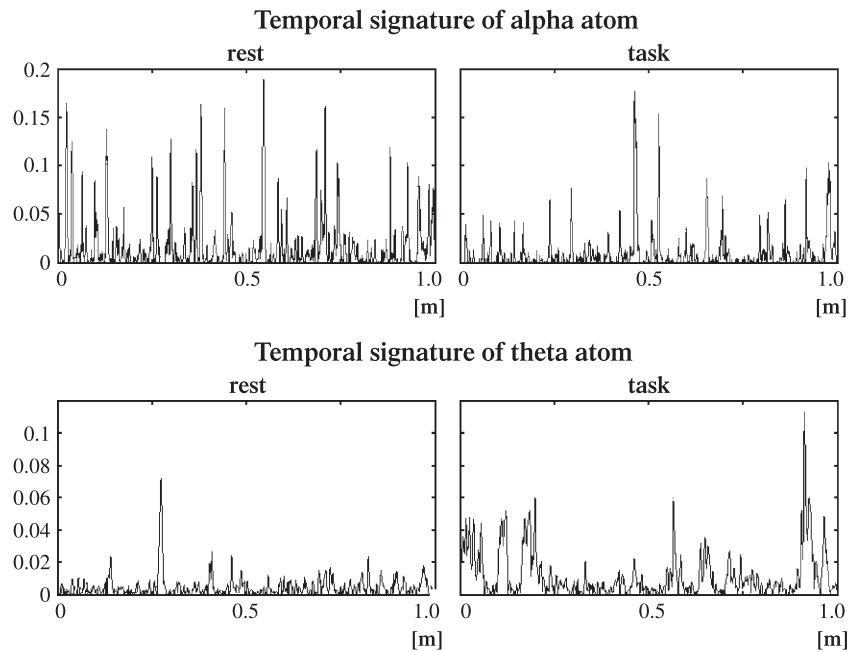


Fig. 7. Theta and alpha activities detected by the screening procedure, based on previously identified spatial and spectral signatures for each component. The screening was applied for 1 min of continuous data sets in the resting and task condition of a new data set of Subject 2. There are more theta bursts and less alpha bursts in the task state, while the converse relationship holds in the resting state.

assumptions has been amply used widely in the so-called distributed inverse solutions, which find the best applicability in the reconstruction of activation of wide areas in the brain. Advantages and shortcomings of these methods have been discussed extensively in the literature (Fuchs et al., 1999; Pascual-Marqui 1995, 1999; Pascual-Marqui et al., 1994).

## Results

### Parallel Factor Analysis

To evaluate the performance of PARAFAC for extracting alpha and theta activities in EEG, two different states were prepared in a benchmark data set. For this purpose, 10 segments of 1 s each were selected from the wavelet-transformed data (after wavelet transformation the time-varying EEG spectrum data set was subsampled to 100 Hz to reduce the computational cost of PARAFAC). Clear alpha activity is observed continuously during resting and task condition; however, strong theta activity appears only intermittent-

ly during task condition. Therefore, five segments were selected randomly from the resting condition and the other five segments were selected from the portions that contain typical theta bursts during the task. The segments were concatenated into the benchmark data set consecutively to form the three dimensional matrix  $S_{(N_d \times N_f \times N_p)}$ .

In the PARAFAC decomposition of  $S$ , two atoms appeared for all subjects with spectral signature peaking in the alpha and theta range (Fig. 2).

The analyzed frequency range was 0.5–20 Hz step by 0.5 Hz, which is sufficient to extract theta and alpha activity. The use of the Concordia index suggested that in three subjects, these two atoms were sufficient to explain the data set. In two subjects, an additional atom was needed. The Concordia was more than 90% in all cases (optimally it should be 100%). The alpha and theta peaks were around 11 and 7 Hz, respectively. Subject 1 is typical of those who showed strong alpha and little theta activity during rest conditions. Temporal signatures (Fig. 3) show that during the task condition, this subject produced strong theta activities and reduced alpha activity.

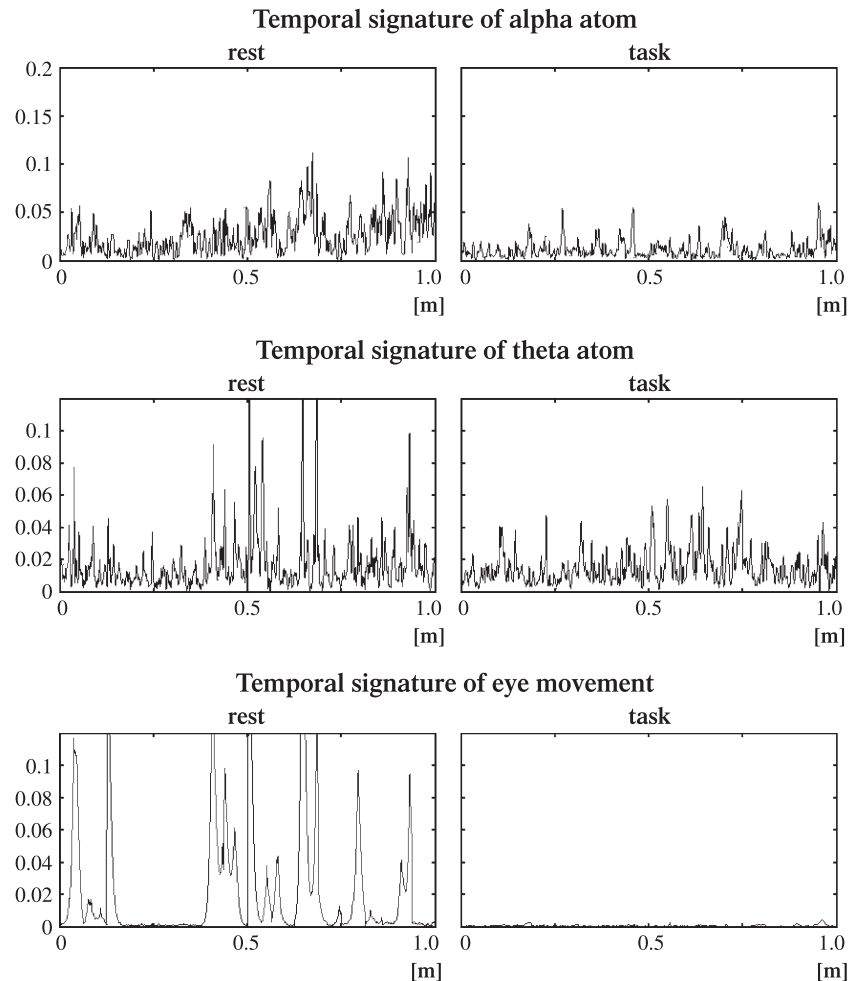


Fig. 8. To train the screening algorithm for the detection of eye movement artifacts, PARAFAC was applied to a data set containing typical theta and alpha activity, as well as eye blinks. Using the spatial and spectral signatures of these three components as a template, we screened 1 min of continuous data sets obtained in the resting condition, which was not used for estimating factors. The data set recorded from Subject 1 contained many eye blink artifacts in the resting condition and fewer in the task condition. On visual inspection of the raw data, it appears that there are more theta bursts during rest than during the task condition. The PARAFAC atomic decomposition showed that this is due to eye blink, as the many peaks in the temporal signature of theta and eye blink atom coincide.

Figs. 4(a) and (b) show the spatial signatures of the alpha and theta atoms as topographic maps for this subject.

The alpha and theta atoms appear in occipital and frontal area, respectively. Figs. 4(c) and (d) are the corresponding SSI solutions for these spatial signatures. The sources for the alpha and theta atoms are in the calcarine sulcus and in the anterior middle frontal cortex, respectively. These spatial distributions were relatively stable for all subjects. Fig. 5 shows the estimated SSI solutions for the spatial signatures of all subjects, corresponding to the theta atom. The activated region had a predominantly frontal distribution in all subjects.

#### Principal Component Analysis

To compare it with PARAFAC, we also carried out PCA of  $\mathbf{S}$ . For this purpose, the data were transformed into a matrix by unfolding the three dimensional array. Results from our PARAFAC decomposition of  $\mathbf{S}$  were matched with corresponding results obtained by applying PCA to the unfolded data set. Figs. 6(a) and (c) show the reconstructed components in the frequency/time plane that correspond to the two largest eigenvalues of PCA. The first component showed strong alpha activities during the resting condition. The second component shows strong theta activities and reduced alpha activity during the task condition. These components had a marked resemblance to the frequency/time reconstructed plane of the alpha and theta atoms of PARAFAC (Figs. 6(b) and (d)).

The peaks of these activities, as well as the order of appearance of the atoms, were the same in PARAFAC and PCA decompositions. The topographies of the PCA components were also very similar to the spatial signature of the PARAFAC atoms (Figs. 4(a) and (b)).

#### Screening

Using the screening procedure described above, it was possible to use PARAFAC to search for the presence of atoms in a new data set, which were not used for estimating the loadings. In this study, we consider new data from the same subject and only theta and alpha atoms were of interest. The spatial and spectral signatures for the templates of theta and alpha atoms were estimated by using Subject 2 data as a benchmark. Reconstruction of the temporal signature for new data was carried out by screening 1 min of continuous data in the resting and task conditions. Fig. 7 shows the appearance of pronounced theta bursts and the decrease of alpha bursts in the task state. In the resting state, the theta burst disappeared and the alpha bursts increased.

#### Artifact detection

If PARAFAC is applied to a data set that contains artifact, some of the atoms will correspond to such activity (e.g., eye blink, eye movement, EMG, etc. . .). Using these atoms as templates, artifact detection can be carried out by the screening procedure. As an example, PARAFAC was applied first to a training data set from Subject 1 that contained theta and alpha oscillations as well as eye movement artifact (this was assessed empirically by an experienced electrophysiologist). The number of atoms was chosen such they could be identified easily as theta, alpha, and eye movement artifact. Using the spatial and spectral signatures of these three atoms as templates, 1 min of continuous data in resting and task conditions were screened. Fig. 8 shows the corresponding temporal signature of the three atoms for Subject 1.

The data set recorded from this subject contained many eye movement artifacts in the resting condition and far fewer in the task condition. A superficial analysis would lead to the conclusion that there are more theta bursts during the resting than the task condition. However, these are probably due to the presence of artifacts, because there are many coincident peaks in the temporal signatures of the theta and artifact atoms.

#### Discussion

This paper introduces a new type of space/frequency/time atomic decomposition of the EEG. It takes advantage of the fact that three-way arrays of data may be decomposed into a sum of atoms of which is a trilinear combination of factors or signatures. This decomposition will be unique if the number of atoms is less than half the sum of the ranks of the three matrices formed by concatenating the signatures. The application of this concept to obtain unique space/frequency/time decomposition for the EEG is possible by arranging the multichannel evolutionary spectrum of the EEG in a three-way data array with dimensions indexed by channel, frequency, and time. The underlying theoretical requirement is that of a moderate amount of linear independence for atom topographies, spectra, and time courses. This is a much milder requirement than previous models underlying space/time atomic decompositions (PCA or ICA). This is the first intrinsically unique space/frequency/time atomic decomposition proposed in the literature.

A physiological interpretation of the model presented here is intuitively appealing. It assumes neural sources with a fixed geometrical relation to the sensors that produce oscillatory activity with a fixed spectral whose amplitude is temporally modulated. This model is not a completely general; for example, a frequency-modulated chirp would require a large number of components, such that the rank condition would be violated.

On the other hand, at most three space/frequency/time atoms are necessary for an adequate description of the EEG data analyzed in this paper. The use of the Corcondia index facilitates the selection of the number of components, an issue that is still difficult for most decomposition methods including PCA and ICA. Also, for the data set analyzed in this paper, two of the spectral signatures had a clear and common interpretation as theta and alpha oscillatory activity. Other components were not so constant and were sometimes difficult to interpret. It may be that more a priori information must be built into the model to avoid identification ambiguity. In this regard, PARAFAC shares with ICA the lack of inherent ordering of the extracted components. In the case of ICA, clustering techniques have been applied to identify common modes (Makeig et al., 2002). In the future, this approach might be used also for the space/frequency/time decomposition.

Our work also shows that the temporal signatures of the theta and alpha atoms may be used as indicators of physiological states. A comparison with a PCA-Varimax analysis shows that the results of the latter may sometimes be similar to those of PARAFAC in terms of description of space and frequency/time profiles. PARAFAC, however, provides a more parsimonious description of the data in a qualitatively simpler manner.

An important application of the space/frequency/time atomic decomposition is the screening of new data sets for the presence of particular atoms. In other words, PARAFAC offers the opportunity to screen recordings for bursts of oscillatory activity with a given

topographic and spectral content. The results shown here demonstrate the feasibility of this technique, not only to detect physiological activity but also for the ever-present problem of artifact removal.

One limitation of the implementation of the method presented here is the estimation by the least-squares techniques. Embedding the model in a Bayesian framework would allow more flexibility in incorporating a priori knowledge and a principled testing of different hypotheses about signatures within and between subjects.

#### *Are these results 'real'?*

As noted above, it has been proven that if the data is approximately trilinear, if the correct number of components is used, and if the signal-to-noise ratio is appropriate, then the true underlying phenomena will be found with PARAFAC (Harshman 1972; Kruskal, 1976, 1997). Also, there have been examples in which the PARAFAC model coincides with a physical model, e.g., fluorescence excitation–emission, chromatography with spectral detection, and spatiotemporal analysis of multichannel-evoked potentials (Field and Graupe, 1991).

The usual and stronger way to validate the truthfulness of results given by the application of multiway models is by split-half analysis (Harshman, 1984; Harshman and De Sarbo, 1984). Due to the uniqueness of the PARAFAC model, the same loadings must be obtained in the non-split modes from models of any suitable subset of the data. This analysis was not accomplished in this work. Although we do not have definitive proof that our results reflect exactly the real physical phenomena underlying the data, there are some aspects of the method we can lean upon for assessing the robustness of the model.

First, the algorithm is implemented such that we can select different initial values. We obtain the same results by applying the method with initial values given by direct trilinear decomposition of the data as we do by random guesses. Second, changing the convergence criterion over four orders of magnitude did not affect the results. Finally, the interpretability of the results, their agreement with previous studies of this kind of electrophysiological experiment, and their robustness with constraints to the loadings like non-negativity and orthogonality; as well as the small variability among subjects, all give additional evidence in this regard.

From this perspective, we think that PARAFAC space/frequency/time atomic decomposition of multichannel evolutionary spectrum of the EEG can reliably and uniquely extract meaningful and significant physiological activities, although this does not ensure that the results correspond to the physical sources that generated the data. Furthermore, the application of this technique requires careful preprocessing of the data, exploration of outliers and degenerate solutions, use of constraints, selection of appropriate model order, and validation of the results as this cannot be accomplished easily without prior knowledge of, or a theoretical basis for, of the expected results. PARAFAC should be simply considered another promising addition to the Neuroimaging analysis toolkit.

#### **Acknowledgments**

The authors want to thank Prof. Mark S. Cohen, Director of Functional MR Imaging, Ahmanson-Lovelace Brain Mapping

Center, UCLA School of Medicine, for his very helpful advice and suggestion for this work.

#### **References**

- Achim, A., Bouchard, S., 1997. Toward a dynamic topographic components model. *Electroencephalogr. Clin. Neurophysiol.* 103, 381–385.
- Bertrand, O., Bohorquez, J., Pernier, J., 1994. Time–frequency digital filtering based on an invertible wavelet transform: an application to evoked potential. *IEEE Trans. Biomed. Eng.* 41, 77–88.
- Bro, R., 1998. Multi-way Analysis in the Food Industry: Models, Algorithms and Applications. PhD Thesis. University of Amsterdam (NL) and Royal Veterinary and Agricultural University (DK).
- Carroll, J.D., Chang, J., 1970. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika* 35, 283–319.
- Casanova, R., Valdes-Sosa, P., Garcia, F.M., 2000. Frequency domain distributed inverse solution. In: Aine, C.J., Okada, Y., Stroink, G., Swinthenby, S.J., Wood, C.C. (Eds.), *Biomag 96: Proceedings of the Tenth International Conference on Biomagnetism*. Springer Verlag.
- Chen, S., Donoho, D., 2001. Atomic decomposition by basis pursuit. *SIAM Rev.* 43, 129–159.
- Cichocki, A., Amari, S., 2002. Adaptive Blind Signal and Image Processing. John Wiley & Sons, Ltd.
- Dahlhaus, R., 1997. Fitting time series models to non-stationary processes. *Ann. Stat.* 25, 1–37.
- Estienne, F., Matthijs, N., Massart, D.L., Ricoux, P., Leibovici, D., 2001. Multi-way modelling of high-dimensionality electroencephalographic data. *Chemom. Intell. Lab. Syst.* 58, 59–71.
- Field, A.S., Graupe, D., 1991. Topographic component (Parallel Factor) analysis of multichannel evoked potentials: practical issues in trilinear spatiotemporal decomposition. *Brain Topogr.* 3, 407–423.
- Fuchs, M., Wagner, M., Kohler, T., Wischman, H.A., 1999. Linear and nonlinear current density reconstructions. *J. Clin. Neurophysiol.* 16, 267–295.
- Gonzalez Andino, S.L., Grave de Peralta Menendez, R., Lantz, C.M., Blank, O., Michel, C.M., Landis, T., 2001. Non-stationary distributed source approximation: an alternative to improve localization procedures. *Hum. Brain Mapp.* 14, 81–95.
- Harmony, T., Fernandez, T., Silva, J., et al., 1999. Do specific EEG frequencies indicate different processes during mental calculation? *Neurosci. Lett.* 266, 25–28.
- Harshman, R.A., 1970. Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multi-modal factor analysis. *UCLA Work. Pap. Phon.* 16, 1–84.
- Harshman, R.A., 1972. Determination and proof of minimum uniqueness conditions for PARAFAC1. *UCLA Work. Pap. Phon.* 22, 111–117.
- Harshman, R.A., 1984. "How can I know if it's 'real'?" A catalog of diagnostics for use with three-mode factor analysis and multidimensional scaling. In: Law, H.G., Snyder, C.W., Hattie, J.A., McDonald, R.P. (Eds.), *Research Methods for Multimode Data Analysis*. Praeger, New York, pp. 566–591.
- Harshman, R.A., De Sarbo, W.S., 1984. An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques. In: Law, H.G., Snyder, J.A., Hattie, J.A., McDonald, R.P. (Eds.), *Research Methods for Multimode Data Analysis*. Praeger, New York, pp. 602–642.
- Hyvarinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. John Wiley & Sons, Inc.
- Ishihara, T., Yoshii, N., 1972. Multivariate analytic study of EEG and mental activity in juvenile delinquents. *Electroencephalogr. Clin. Neurophysiol.* 33, 71–80.
- Jensen, O., Tesche, C.D., 2002. Frontal theta activity in human increased with memory load in a working memory task. *Eur. J. Neurosci.* 15, 1395–1399.

- Jung, T.P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., Sejnowski, T.J., 2001. Analysis and visualization of single-trial event-related potential. *Hum. Brain Mapp.* 14, 166–185.
- Koenig, T., Marti-Lopez, F., Valdes-Sosa, P., 2001. Topographic time-frequency decomposition of the EEG. *NeuroImage* 14, 383–390.
- Kronland-Martinet, R., Morlet, J., 1987. Analysis of sound patterns through wavelet transforms. *Int. J. Pattern Recogn. Artif. Intell.* 1, 273–302.
- Kruskal, J.B., 1976. More factors than subjects, test and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika* 41, 281–293.
- Kruskal, J.B., 1977. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.* 18, 95–138.
- Lachaux, J.P., Rodriguez, E., Martinerie, J., Varela, F.J., 1999. Measuring phase synchrony in brain signals. *Hum. Brain Mapp.* 8, 194–208.
- Lagerlund, T.D., Sharbrough, F.W., Busacker, N.E., 1997. Spatial filtering of multichannel electroencephalographic recordings through principal component analysis by singular value decomposition. *J. Clin. Neurophysiol.* 14, 73–83.
- Lopes da Silva, F., 1987. EEG analysis: theory and practice. In: Neidermeyer, E., Lopes da Silva, F. (Eds.), *Electroencephalography*. Urban and Schwarzenberg.
- Makeig, S., 1993. Auditory event-related dynamics of the EEG spectrum and effects of exposure to tones. *Electroencephalogr. Clin. Neurophysiol.* 86, 283–293.
- Makeig, S., Westerfield, M., Jung, T.-P., et al., 2002. Dynamic brain sources of visual evoked responses. *Science* 295, 690–694.
- Möcks, J., 1988a. Decomposing event-related potential: a new topographic components model. *Biol. Psychol.* 26, 199–215.
- Möcks, J., 1988b. Topographic components model for event-related potentials and some biophysical considerations. *IEEE Trans. Biomed. Eng.* 35, 482–484.
- Nunez, P.L., 1993. *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford Univ. Press.
- Pascual-Marqui, R.D., 1995. Reply to comments by Hamalainen, Ilmoniemi and Nunez. In *source localization: continuing discussion of the inverse problem*. *Skrandies W ISBET Newsletter*, vol. 6, pp. 16–28. ISSN 0947-5133.
- Pascual-Marqui, R.D., 1999. Review of methods for solving the EEG inverse problem. *Int. J. Bioelectromagn.* 1, 1.
- Pascual-Marqui, R.D., Michel, C.M., Lehmann, D., 1994. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *Int. J. Psychophysiol.* 18, 49–65.
- Sasaki, K., Tsujimoto, T., Nishikawa, S., Nishitani, N., Ishihara, T., 1996. Frontal mental theta wave recorded simultaneously with magnetoencephalography and electroencephalography. *Neuroscience* 26, 79–81.
- Sidiropoulos, N.D., Bro, R., 2000. On the uniqueness of multilinear decomposition of  $N$ -way arrays. *J. Chemom.* 14, 229–239.
- Soong, A.C., Koles, Z.J., 1995. Principal-component localization of the sources of the background EEG. *IEEE Trans. Biomed. Eng.* 42, 59–67.
- Tallon-Baudry, C., Bertrand, O., Delpuech, C., Pernier, J., 1997. Oscillatory  $\gamma$ -band (30–70 Hz) activity induced by a visual search task in humans. *J. Neurosci.* 17, 722–734.
- Varela, F., Lachaux, J.P., Rodriguez, E., Martinerie, J., 2001. The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* 2, 229–239.

## **Introduction: multimodal neuroimaging of brain connectivity**

## Introduction: multimodal neuroimaging of brain connectivity

Pedro A Valdés-Sosa, Rolf Kötter and Karl J Friston

*Phil. Trans. R. Soc. B* 2005 **360**, 865-867

doi: 10.1098/rstb.2005.1655

---

### References

[This article cites 33 articles](#)

<http://rstb.royalsocietypublishing.org/content/360/1457/865.full.html#ref-list-1>

### Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click [here](#)

---

To subscribe to *Phil. Trans. R. Soc. B* go to: <http://rstb.royalsocietypublishing.org/subscriptions>

---

## Introduction: multimodal neuroimaging of brain connectivity

A major conceptual tenet of modern neuroscience, stated explicitly in the work of its founders (Freund 2002), is that the computational properties of the brain are a direct consequence of its circuitry. This insight has been cumulatively validated over the years, but has received unprecedented attention in the past decades. This is owing to several factors. In the first place, technological advances have transformed the acquisition of data about neural connections from a slow paced, tentative groping, into a high throughput process of massive multimodal data acquisition (Kotter 2001; Buzsaki 2004) that encompasses morphological, neurochemical and functional variables. With the advent of modern neuroimaging methods, much of this data can now be observed *in vivo* (Aine 1995; Savoy 2001). In the second place, these advances in measurements have occurred on par with theoretical breakthroughs that now allow the formal analysis of large complex networks (Albert & Barabasi 2002; Hilgetag *et al.* 2002; Newman 2003). In the third place, successful efforts in large-scale science, (brought to the limelight by the Human Genome Project) have established new paradigms of electronic collaboration, data sharing and processing that are being applied to the data acquired (Van Essen 2002).

As a response to this situation, a series of multi-disciplinary workshops have been organized around the theme of *brain connectivity*, first in Dusseldorf (Lee *et al.* 2003) and later in Cambridge (Bullmore *et al.* 2004). These workshops assessed the data accumulated, the methods by which they were gathered and analysed, and generated general theoretical conclusions. They also charted out areas in which further work was necessary. As a consequence, a third workshop was carried out during 26–30 April 2004 in Havana, organized by the Cuban Neuroscience Centre (<http://www.hirnforschung.net/download/bcw04.html>). Emphasis on this occasion was placed on the use of *in vivo* neuroimaging with magnetic resonance imaging (MRI) and electroencephalogram (EEG), to determine both anatomical and physiological connectivity. Advances in statistical methodology to determine physiological connectivity was a theme of intense debate, especially regarding the analysis of multimodal EEG–fMRI experiments. The theoretical bases of connectivity studies were addressed with discussions of detailed modelling of neural systems at multiple spatial and temporal scales. Special attention was dedicated to the validation of hypothetical neural connections.

When discussing with Professor Semir Zeki the subject matter of these discussions, we were encouraged to put together a theme issue of the *Philosophical Transactions of the Royal Society B* dedicated to brain connectivity. For this purpose, we invited a distinguished series of authors to expand their thoughts, inspired by the Havana workshop and subsequent exchanges, in order to give a coherent overview of current work in this area.

The first four papers (Tuch *et al.* 2005; Perrin *et al.* 2005; Parker & Alexander 2005; Behrens & Johansen-Berg 2005) provide a state-of-the-art revision of the use of diffusion MRI techniques for the *in vivo* estimation of anatomical connectivity. A new method—*q*-ball imaging—for measuring the diffusion of water with MRI is explained and validated both in animal preparations and phantoms. Another candidate method—PAS-MRI—is used to drive probabilistic fibre tracking for the first time, and the use of probabilistic tracking methods for the segmentation of brain structures is outlined.

These papers on anatomical connectivity are followed by those on physiological connectivity as reflected by EEG or fMRI time-series. The work by Worsley *et al.* (2005), Dodel *et al.* (2005) and Salvador *et al.* (2005) and explain methods for the determination of functional connectivity, the relatively assumption-free estimation of the correlation between different brain areas (Friston 1994). The performance of random field theory, the theoretical underpinning of neuroimaging statistics, for testing massive sets of correlations is studied. Methods for studying the conditional independence of brain structures are developed both in the time and frequency domain. Of note is the introduction of graphical models (Wermuth & Lauritzen 1990; Cowell *et al.* 1999) as a theoretically sound basis for the study of functional connectivity.

With more structured time-series models, Kamiński (2005), Eichler (2005), Valdés-Sosa *et al.* (2005) and Penny *et al.* (2005) attempt to estimate causal relations or effective connectivity in developments that combine modern causality theory (Glymour *et al.* 1988; Pearl 2000; Spirtes *et al.* 2000) with classical time-series analysis. The importance of including all sources of signals into a common multichannel system when estimating causal relations was stressed.

The following papers (Tass 2005; Beckmann *et al.* 2005; Koenig *et al.* 2005; Riera *et al.* 2005) develop methods for physiological connectivity analysis by means of EEG recordings, fMRI recordings or concurrent EEG/fMRI experiments. The latter pose challenging modelling issues, but also promise

One contribution of 21 to a Theme Issue ‘Multimodal neuroimaging of brain connectivity’.



increased spatial and temporal resolution by the fusion of information between these modalities.

Placing this area of research on a sounder theoretical basis, Robinson *et al.* (2005); Breakspear & Stam (2005) and Harrison *et al.* (2005) develop multiscale, stochastic models of neural dynamics.

As important as it is to develop statistical methods for theoretical models of brain connectivity, it is essential to devise strategies to validate them. This aspect is addressed by the last two papers in the issue. Horwitz *et al.* (2005) develop realistic computational models that explore physical limits on inference about connectivity. Paus (2005) explores the use of transcranial magnetic stimulation (TMS) perturbation as a means of confirming casual relations in brain systems.

Careful perusal of this series of papers brings to mind areas in which further work must be done. The comparison of *in vivo* diffusion MRI-based tractography information with physiological connectivity measures in the same subjects has not been carried out systematically. The more ambitious use of diffusion MRI tractography probability distributions as prior information for the estimation of functional and effective connectivity has yet to be achieved. While progress in modelling neural systems at the dynamical level is encouraging, the formulation and validation of adequate observation equations leaves much to be desired. In all, this field will remain an exciting area of research in the future.

Pedro A. Valdés-Sosa<sup>1</sup>

Rolf Kötter<sup>2</sup>

Karl J. Friston<sup>3</sup>

March 2005

<sup>1</sup>*Cuban Neuroscience Centre, Avenue 25 No. 15202 esquina 158, Cubanacan, Playa, PO Box 6412/6414, Area Code 11600, Ciudad Habana, Cuba*  
(peter@cneuro.edu.cu)

<sup>2</sup>*Computational and Systems Neuroscience Group, C. & O. Vogt Brain Research Institute and Institute of Anatomy II, Heinrich Heine University, Universitätsstrasse 1, 40225 Düsseldorf, Germany*

<sup>3</sup>*Leopold Müller Functional Imaging Laboratory, Wellcome Department of Imaging Neuroscience, Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, UK*

## REFERENCES

- Aine, C. J. 1995 A conceptual overview and critique of functional neuroimaging techniques in humans. 1. MRI/fMRI and Pet. *Crit. Rev. Neurobiol.* **9**, 229–309.
- Albert, R. & Barabasi, A. L. 2002 Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97.
- Beckmann, C. F., DeLuca, M., Devlin, J. T. & Smith, S. M. 2005 Investigations into resting-state connectivity using independent component analysis. *Phil. Trans. R. Soc. B* **360**, 1001–1013. (doi:10.1098/rstb.2005.1634.)
- Behrens, T. E. J. & Johansen-Berg, H. 2005 Relating connective architecture to grey matter function using diffusion imaging. *Phil. Trans. R. Soc. B* **360**, 903–911. (doi:10.1098/rstb.2005.1640.)
- Breakspear, M. & Stam, C. J. 2005 Dynamics of a neural system with a multiscale architecture. *Phil. Trans. R. Soc. B* **360**, 1051–1074. (doi:10.1098/rstb.2005.1643.)
- Bullmore, E., Harrison, L., Lee, L., Mechelli, A. & Friston, K. 2004 Brain connectivity workshop, Cambridge UK, May 2003. *Neuroinformatics* **2**, 123–125.
- Buzsaki, G. 2004 Large-scale recording of neuronal ensembles. *Nat. Neurosci.* **7**, 446–451.
- Cowell, R. G., Dawid, A. P., Lauritzen, S. L. & Spiegelhalter, D. J. 1999 *Probabilistic networks and expert systems*. New York: Springer.
- Dodel, S., Golestani, N., Pallier, C., El Kouby, V., Le Bihan, D. & Poline, J.-B. 2005 Condition-dependent functional connectivity: syntax networks in bilinguals. *Phil. Trans. R. Soc. B* **360**, 921–935. (doi:10.1098/rstb.2005.1653.)
- Eichler, M. 2005 A graphical approach for evaluating effective connectivity in neural systems. *Phil. Trans. R. Soc. B* **360**, 953–967. (doi:10.1098/rstb.2005.1641.)
- Freund, T. F. 2002 Changes in the views of neuronal connectivity and communication after Cajal: examples from the hippocampus. *Changing Views of Cajals Neuron* **136**, 203–213.
- Friston, K. J. 1994 Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapp.* **2**, 56–78.
- Glymour, C., Scheines, R., Spirtes, P. & Kelly, K. 1988 Tetrad-discovering causal-structure. *Multivariate Behav. Res.* **23**, 279–280.
- Harrison, L. M., David, O. & Friston, K. J. 2005 Stochastic models of neuronal dynamics. *Phil. Trans. R. Soc. B* **360**, 1075–1091. (doi:10.1098/rstb.2005.1648.)
- Hilgetag, C., Kötter, R., Stephan, K. E. & Sporns, O. 2002 Computational methods for the analysis of brain connectivity. In *Computational neuroanatomy* (ed. G. Ascoli). Totowa, NJ: Humana Press.
- Horwitz, B., Warner, B., Fitzer, J., Tagamets, M.-A., Husain, F. T. & Long, T. W. 2005 Investigating the neural basis for functional and effective connectivity. Application to fMRI. *Phil. Trans. R. Soc. B* **360**, 1093–1108. (doi:10.1098/rstb.2005.1647.)
- Kamiński, M. 2005 Determination of transmission patterns in multichannel data. *Phil. Trans. R. Soc. B* **360**, 947–952. (doi:10.1098/rstb.2005.1636.)
- Koenig, T., Studer, D., Hubl, D., Melie, L. & Strik, W. K. 2005 Brain connectivity at different time-scales measured with EEG. *Phil. Trans. R. Soc. B* **360**, 1015–1023. (doi:10.1098/rstb.2005.1649.)
- Kötter, R. 2001 Neuroscience databases: tools for exploring brain structure–function relationships. *Phil. Trans. R. Soc. B* **356**, 1111–1120. (doi:10.1098/rstb.2001.0902.)
- Lee, L., Harrison, L. M. & Mechelli, A. 2003 A report of the functional connectivity workshop, Dusseldorf 2002. *NeuroImage* **19**, 457–465.
- Newman, M. E. J. 2003 The structure and function of complex networks. *SIAM Rev.* **45**, 167–256.
- Parker, G. J. M. & Alexander, D. C. 2005 Probabilistic anatomical connectivity derived from the microscopic persistent angular structure of cerebral tissue. *Phil. Trans. R. Soc. B* **360**, 893–902. (doi:10.1098/rstb.2005.1639.)
- Paus, T. 2005 Inferring causality in brain images: a perturbation approach. *Phil. Trans. R. Soc. B* **360**, 1109–1114. (doi:10.1098/rstb.2005.1652.)
- Pearl, J. 2000 *Causality*. Cambridge: Cambridge University Press.
- Penny, W., Ghahramani, Z. & Friston, K. 2005 Bilinear dynamical systems. *Phil. Trans. R. Soc. B* **360**, 983–993. (doi:10.1098/rstb.2005.1642.)
- Perrin, M., Poupon, C., Rieul, B., Leroux, P., Constantinesco, A., Mangin, J.-F. & LeBihan, D.

- 2005 Validation of  $q$ -ball imaging with a diffusion fibre-crossing phantom on a clinical scanner. *Phil. Trans. R. Soc. B* **360**, 881–891. (doi:10.1098/rstb.2005.1650.)
- Riera, J., Aubert, E., Iwata, K., Kawashima, R., Wan, X. & Ozaki, T. 2005 Fusing EEG and fMRI based on a bottom-up model: inferring activation and effective connectivity in neural masses. *Phil. Trans. R. Soc. B* **360**, 1025–1041. (doi:10.1098/rstb.2005.1646.)
- Robinson, P. A., Rennie, C. J., Rowe, D. L., O'Connor, S. C. & Gordon, E. 2005 Multiscale brain modelling. *Phil. Trans. R. Soc. B* **360**, 1043–1050. (doi:10.1098/rstb.2005.1638.)
- Salvador, R., Suckling, J., Schwarzbauer, C. & Bullmore, E. 2005 Undirected graphs of frequency-dependent functional connectivity in whole brain networks. *Phil. Trans. R. Soc. B* **360**, 937–946. (doi:10.1098/rstb.2005.1645.)
- Savoy, R. L. 2001 History and future directions of human brain mapping and functional neuroimaging. *Acta Psychol.* **107**, 9–42.
- Spirtes, P., Glymour, C. & Scheines, R. 2000 *Causation, prediction, and search*. Cambridge: The MIT Press.
- Tass, P. A. 2005 Estimation of the transmission time of stimulus-locked responses: modelling and stochastic phase resetting analysis. *Phil. Trans. R. Soc. B* **360**, 995–999. (doi:10.1098/rstb.2005.1635.)
- Tuch, D. S., Wisco, J. J., Khachaturian, M. H., Ekstrom, L. B., Kotter, R. & Vanduffel, W. 2005  $Q$ -ball imaging of macaque white matter architecture. *Phil. Trans. R. Soc. B* **360**, 869–879. (doi:10.1098/rstb.2005.1651.)
- Valdés-Sosa, P. A., Sanchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernandez, M., Bosch-Bayard, J., Melie-García, L. & Canales-Rodriguez, E. 2005 Estimating brain functional connectivity with sparse multivariate autoregression. *Phil. Trans. R. Soc. B* **360**, 969–981. (doi:10.1098/rstb.2005.1654.)
- Van Essen, D. C. 2002 Windows on the brain: the emerging role of atlases and databases in neuroscience. *Curr. Opin. Neurobiol.* **12**, 574–579.
- Wermuth, N. & Lauritzen, S. L. 1990 On substantive research hypotheses, conditional-independence graphs and graphical chain models. *J. R. Stat. Soc. B Met.* **52**, 21–50.
- Worsley, K. J., Chen, J.-I., Lerch, J. & Evans, A. C. 2005 Comparing functional connectivity via thresholding correlations and singular value decomposition. *Phil. Trans. R. Soc. B* **360**, 913–920. (doi:10.1098/rstb.2005.1637.)

## **Concurrent EEG/fMRI analysis by multiway Partial Least Squares**

## Concurrent EEG/fMRI analysis by multiway Partial Least Squares

Eduardo Martínez-Montes,<sup>a,\*</sup> Pedro A. Valdés-Sosa,<sup>a</sup> Fumikazu Miwakeichi,<sup>b</sup>  
Robin I. Goldman,<sup>c</sup> and Mark S. Cohen<sup>d</sup>

<sup>a</sup>Neurophysics Department, Cuban Neuroscience Center, Havana, Cuba

<sup>b</sup>Laboratory for Dynamics of Emergent Intelligence, RIKEN Brain Science Institute, Wako, Saitama, Japan

<sup>c</sup>Hatch Center for MR Research, Columbia University, New York, NY 10032, USA

<sup>d</sup>Ahmanson-Lovelace Brain Mapping Center, UCLA, Medical School, Los Angeles, CA 90095, USA

Received 17 July 2003; revised 12 March 2004; accepted 17 March 2004

Data may now be recorded concurrently from EEG and functional MRI, using the Simultaneous Imaging for Tomographic Electrophysiology (SITE) method. As yet, there is no established means to integrate the analysis of the combined data set. Recognizing that the hemodynamically convolved time-varying EEG spectrum,  $S$ , is intrinsically multidimensional in space, frequency, and time motivated us to use multiway Partial Least-Squares (N-PLS) analysis to decompose EEG (independent variable) and fMRI (dependent variable) data uniquely as a sum of “atoms”. Each EEG atom is the outer product of spatial, spectral, and temporal signatures and each fMRI atom the product of spatial and temporal signatures. The decomposition was constrained to maximize the covariance between corresponding temporal signatures of the EEG and fMRI. On all data sets, three components whose spectral peaks were in the theta, alpha, and gamma bands appeared; only the alpha atom had a significant temporal correlation with the fMRI signal. The spatial distribution of the alpha-band atom included parieto-occipital cortex, thalamus, and insula, and corresponded closely to that reported by Goldman et al. [NeuroReport 13(18) (2002) 2487] using a more conventional analysis. The source reconstruction from EEG spatial signature showed only the parieto-occipital sources. We interpret these results to indicate that some electrical sources may be intrinsically invisible to scalp EEG, yet may be revealed through conjoint analysis of EEG and fMRI data. These results may also expose brain regions that participate in the control of brain rhythms but may not themselves be generators. As of yet, no single neuroimaging method offers the optimal combination of spatial and temporal resolution; fusing fMRI and EEG meaningfully extends the spatio-temporal resolution and sensitivity of each method. © 2004 Elsevier Inc. All rights reserved.

**Keywords:** N-PLS; EEG/fMRI fusion; PARAFAC; Multiway analysis; SSI; SITE

### Introduction

The armamentarium of the neuroscientist now includes tools with spatial resolution ranging from centimeters to microns and temporal resolution from years to nanoseconds. Even so, no single tool provides an optimal combination of spatial and temporal resolution, and there generally exists a tradeoff in which improvement in one dimension of resolution requires compromises in the other (Churchland and Sejnowski, 1988). Extending our understanding of the functional architecture of the human brain necessarily requires a rational combination of multiple methods. Particularly attractive is the fusion of the superb temporal resolution of electroencephalography (EEG) or magnetoencephalography (MEG), with the excellent contrast and spatial resolving power of functional MRI (fMRI). Several methods of integration have been reported (Horwitz and Poeppel, 2002), each with its own approaches to analysis.

Under the assumption that the response of the brain to a set of stimuli or conditions is the same when acquired at different times, several groups (Babiloni et al., 2001; Baillet et al., 2001; Singh et al., 1998) have attempted the analysis of EEG and fMRI data, gathered separately. While this approach is not without problems (Gonzalez-Andino et al., 2001; Ioannides, 1999), there is increasing evidence that adequate modeling of multimodal data will allow the estimation of the underlying neural processes with simultaneously high spatial and temporal resolution (Trujillo et al., 2001).

More recently, methods have been described for the concurrent collection of EEG and fMRI data (Goldman et al., 2000). These methods make possible the study of dynamic relationship between fluctuations in the blood oxygenation level dependent (BOLD) signal and the properties of the electrical activity recorded on the scalp. Here, the fMRI and EEG data each necessarily provide evidence of the same underlying brain activity, although the extent to which they are measuring the same signals, or even signals from the same processes, is indeterminate.

In a method they have called Simultaneous Imaging for Tomographic Electrophysiology, or SITE, Goldman et al. (2002) created tomograms of the brain regions whose fMRI signal changes were associated with variations in alpha band power. In that work, 16 bipolar EEG channels were recorded under the eyes-closed resting state that is well known to produce elevated alpha wave activity. To match the EEG and the fMRI time courses, they then convolved the

---

\* Corresponding author. Neurophysics Department, Cuban Neuroscience Center, Avenue 25, Esq. 158, #15202, PO Box 6412, 6414 Cubanacán, Playa, Havana, Cuba. Fax: +53-7-208-6707.

E-mail address: eduardo@cneuro.edu.cu (E. Martínez-Montes).

Available online on ScienceDirect (www.sciencedirect.com.)

measured alpha power at each time point with an a priori hemodynamic response model (Cohen, 1997) and calculated the correlation between the fluctuations of alpha activity and the BOLD time course at each voxel. Alpha activity was defined as the broad band spectral power in the frequency range 8–12 Hz calculated over the 2.5-s period needed to acquire each MRI volume and averaged over the occipital derivations (*T6-O2*, *O2-P4*, *T5-O1*, *O1-P3*). Positive correlations were found in thalamic voxels as well as in the insula, while negative correlations predominated in the parieto-occipital cortex. Thus, these correlation maps showed extended thalamo-cortical structures implicated in the generation of this EEG rhythm. A deeper analysis of these results, however, leads to further questions.

Traditionally, the EEG has been decomposed into a series of fixed broad spectral bands (delta, theta, alpha, beta, gamma, ...) based more on history and discovery than on a theoretical framework. This approach, although computationally convenient, may obscure the fact that the sources of each of these characteristic oscillations may or may not be unique (Szava et al., 1994). It has been shown that the EEG can be analyzed as a partially overlapping spectral components defined with high-frequency resolution (Pascual-Marqui et al., 1988); each spectral component being interpreted as reflecting the activity in a given oscillatory network. We seek here to associate each of these components with the BOLD-fMRI activity. In keeping with standard terminology in time–frequency decompositions (Chen et al., 2001), these components will be designated as “atoms”.

While strong prior information suggests that the scalp locations best associated with alpha power fluctuations may well be near the occipital electrodes, other spectral components may have a more subtle or distributed relationship to scalp topography. Under these more general circumstances, it may be better to have a more data-driven means to estimate the linear combination of EEG measurements (or derivations) that correlate optimally with BOLD. Such estimates are likely to result in greater statistical power for the detection of EEG–fMRI relationships. Similarly, it might be desirable to look at the correlation of the EEG with a calculated optimum linear combination of all BOLD signals, rather than with each voxel separately.

Essentially, our goal has been to seek methods that best explain the spatio-temporal relationships between fMRI and the oscillatory components of the EEG without first forming a priori hypotheses as to which characteristics of the EEG are likely be of most interest. These considerations led us to search for methods of atomic decomposition of the EEG and methods for correlating the output of this decomposition with the fMRI data. There are a lot of well-known methods for data reduction of the EEG. Among them, Principal Components Analysis (PCA), Independent Components Analysis (ICA), and dictionary-based decompositions have been the most explored. They have been applied only to two-dimensional data. As the time-varying EEG spectrum is, in fact, a three-dimensional array (electrode pairs, frequencies, and time), it cannot be expressed conveniently as a matrix. The decomposition of such a multidimensional data has been better accomplished by a generalization of the Singular Value Decomposition known as Parallel Factor Analysis (PARAFAC) (Harshman, 1970), a tool that has been used previously in the analysis of evoked potentials (Field and Graupe, 1991) and pharmacological studies using high-dimensionality EEG data (Estienne et al., 2001). The most interesting advantage of the PARAFAC model is that it provides a unique decomposition without imposing

orthogonality or independence constraints to the components. It is also valued for being a parsimonious and “easily interpretable” model (Bro, 1998).

Several calibration methods (Principal Components Regression, ridge regression) can be used for correlating the EEG decomposition and the fMRI data. Although some general guidelines have been given for establishing a hierarchy among them (Kiers, 1991), there is not definitive calibration method that one can stick to, since its correct application depends strongly on the behavior of the data considered. In a straightforward application of any of these methods (e.g., Principal Components Regression), one could use the EEG spectral power estimates (principal components of time-varying EEG spectrum), for different time segments, as the independent variable to be correlated with the fMRI. This procedure in two steps (decomposing and correlating) does not ensure that we are finding the optimal relationship between the EEG and the fMRI because decomposition is based on nonphysiological assumptions (e.g., it is unreasonable to expect that the activities of individual neural generators to be mutually orthogonal). Therefore, we should search for a method that is capable of simultaneously extracting EEG spectral components or atoms (and their scalp landscapes) having maximal temporal covariance with certain BOLD profiles. One possible candidate for such a multimodal analysis is Partial Least-Squares (PLS) regression, introduced in fMRI analysis by McIntosh et al. (1996). In PLS, the fMRI data are treated as a matrix (voxels by time). PLS identifies those linear combinations of fMRI voxels that have maximal temporal covariance with linear combinations of a second matrix of independent variables, measured at the same time points. The method hinges on calculating the Singular Value Decomposition (SVD) of the covariance matrix between the fMRI and independent variables. This method has been used for spatio-temporal analysis of event-related potentials (Lobaugh et al., 2001) and simultaneous EEG and MEG data (Düzel et al., 2003).

Fortunately, the PLS technique has been extended by Bro (1996) to deal with multidimensional data, obtaining a new model known as Multiway Partial Least Squares or just N-PLS. This model consists essentially of decomposing the independent and dependent data into multilinear models such that the score vectors from these models have pairwise maximal covariance. The multilinear decomposition is made in the same way as PARAFAC, thus inheriting both advantages and limitations of that model.

In this paper, the N-PLS model will be introduced for decomposing the EEG into a sum of atoms each with a specific spatial, temporal, and spectral factors or “signatures”. Simultaneously, the fMRI data will be decomposed into the same number of atoms, each the product of spatial and temporal signatures, in such a way that the latter will have maximal covariance with the EEG temporal signature. The source localization of the EEG spatial signature (topography) of each atom will be examined, allowing separate analysis of the tomographic distribution of the EEG sources (what we will call sources of the “EEG rhythm”) and those tomographic sources obtained as the fMRI tomograms that we interpret as the “brain rhythm” generating system. It should be noted that we have limited our consideration to oscillatory components of the EEG. While important, they do not exhaust the list of interesting phenomena that might possibly relate to the fMRI. Transient waveforms, for example, are not optimally described in the time–frequency framework. In principle, the methods developed here may be extended to consider this situation.

## Methods

Consider a matrix,  $\mathbf{F}_{(N_s \times N_t)}$ , of the fMRI data ( $N_s$  voxels,  $N_t$  time points) that is recorded simultaneously with the EEG time series from  $N_d$  electrodes. Further, define the EEG signal recorded during each TR (the period needed to collect an MRI volume) as a “segment.” In the present case, the time-varying EEG spectrum,  $\mathbf{S}(\omega)_{(N_d \times N_t)}$  ( $\omega$  being the frequency), for  $N_t$  segments, was estimated via the Thomson multitaper method (Thomson, 1982). Let  $\mathbf{s}$  be a reference EEG time signal, formed by selecting a linear combination,  $\mathbf{a}$ , of the EEG electrode power in a given band of frequencies  $\Omega$ , which was then filtered by the hemodynamic response,  $\mathbf{H}$ :

$$\mathbf{s}_{(1 \times N_t)} = \mathbf{a}_{(1 \times N_d)}^T \sum_{\omega \in \Omega} \mathbf{S}(\omega)_{(N_d \times N_t)} \mathbf{H}_{(N_t \times N_t)} \quad (1)$$

where the symbol,  $\mathbf{a}^T$ , represents the transpose of vector  $\mathbf{a}$ . Then, the correlations between the fMRI matrix and the reference EEG signal:  $\mathbf{r}_{(N_s \times 1)} = \text{corr}(\mathbf{F}, \mathbf{s})$  are mapped.

In the analysis performed by Goldman et al. (2002), they chose an ad hoc linear combination,  $\mathbf{a}$  (an occipital electrode set), and frequency band (8–12 Hz) for finding the reference EEG time signal. We will extend this analysis to estimate the optimal linear combination of electrodes, and a particular spectral window defining an optimal frequency band  $\sum_{\omega \in \Omega} \mathbf{b}(\omega) \mathbf{S}(\omega)$ . Finally, we will estimate a suitable linear combination,  $\mathbf{u}_{(1 \times N_t)}^T$  of the elements of the fMRI matrix to be correlated with a particular EEG time signal.

### Parallel Factor Analysis

Recognizing that the time-varying EEG spectrum may be expressed conveniently as a three-dimensional array makes possible the use of Parallel Factor Analysis (PARAFAC) (Carroll and Chang, 1970; Harshman, 1970), a generalization of Principal Component Analysis (PCA) for dealing with multidimensional data. With PARAFAC, the time-varying EEG spectrum is decomposed (in a least-squares sense) into trilinear components, or atoms, each being the product of a spatial, spectral and temporal factors, or signatures.

Unlike PCA, PARAFAC has no rotational freedom; therefore, the decomposition is unique, even without any orthogonality constraints. It has been shown that if the data are approximately trilinear, the correct number of components is used, and the signal-to-noise ratio is adequate, then the PARAFAC algorithm will show the true underlying phenomena (Kruskal, 1976, 1977). Moreover, PARAFAC provides a unique data-determined linear combination, i.e., a reference time signal, to correlate with the fMRI data. The use of PARAFAC in analyzing three-dimensional EEG data, (space, frequency, time) is described in a companion paper (Miwa-keichi et al., 2004).

Then, applied to the time-varying EEG spectrum, which is expressed as a three-dimensional matrix  $\mathbf{S}_{(N_d \times N_w \times N_t)}$ , PARAFAC decomposition establishes an element-wise trilinear model for these data:

$$\hat{S}_{dwt} = \sum_{k=1}^{N_k} a_{dk} b_{wk} c_{tk} + \varepsilon_{dwt} \quad (2)$$

where  $d$ ,  $w$ , and  $t$  designate electrode pairs, frequency, and time, respectively, and the term  $\varepsilon_{dwt}$  represents the error. The total

number of components is  $N_k$ , each of which is designated by index  $k$ . Our problem is to find the so-called “loading matrices”,  $\mathbf{A}$ ,  $\mathbf{B}$ , and  $\mathbf{C}$  whose  $N_k$  columns are the loading vectors  $\mathbf{a}_{k(N_d \times 1)}$ ,  $\mathbf{b}_{k(N_w \times 1)}$ , and  $\mathbf{c}_{k(N_t \times 1)}$  of elements  $a_{dk}$ ,  $b_{wk}$ , and  $c_{tk}$  respectively.

We can fit the model expressed in Eq. (2) by finding

$$\min_{\mathbf{a}_{dk} \mathbf{b}_{wk} \mathbf{c}_{tk}} \left\| S_{dwt} - \sum_{k=1}^{N_k} a_{dk} b_{wk} c_{tk} \right\|^2.$$

The interpretation of the loading vectors is as follows:  $\mathbf{a}_k$  is the spatial signature of the  $k$ th atom, which is a representative topographic map, or linear combination of electrodes;  $\mathbf{b}_k$  is the spectral signature for the  $k$ th atom and  $\mathbf{c}_k$  is the temporal signature, or time course, for atom  $k$ . The only indeterminacies in the least-square solution are the order of components and the scaling of loading vectors. Thus, centering and scaling of the data are needed before decomposition, as is a convention for the signs and scale of the loadings. For PARAFAC, the resulting spectral and spatial loadings are normalized, while the non-normalized loading will be the temporal factor, reflecting the scale of the data.

It is important to select the most appropriate number,  $N_k$ , of components. The Core Consistency Diagnostic (Corcondia) is an approach for so doing that applies especially to PARAFAC models, and has been shown to be a powerful and simple tool for determining the appropriate number of components in multiway models (Bro, 1998). In this work, we use not only Corcondia but also the evaluation of the systematic variation left in the model’s residuals.

PARAFAC has been extensively used in chemometrics, psychometrics, and econometrics. In the field of spatio-temporal analysis of Event-Related Potentials, PARAFAC has been shown to be formally equivalent to the Topographic Components Model (TCM) (Möcks, 1988a,b). Field and Graupe (1991) offered some general guidelines for the correct exploration of EEG data with PARAFAC. The basic pitfall of the application of PARAFAC is that the data are actually not trilinear, and, hence, a careful preprocessing and analysis of the results must be done for assessing the validity of the model.

### Multiway Partial Least-Squares Regression

Despite being a useful tool for data explorations and to find a unique reference EEG time signal, the PARAFAC analysis leaves two important questions unanswered:

- Which frequency components are related to the fMRI signal?
- What is the optimal linear combination of EEG electrodes to correlate with the fMRI?

Partial Least-Squares regression is an automatic procedure to find the linear combination that maximizes the temporal correlation between the EEG and fMRI data (de Jong and Phatak, 1997; Martens and Naes, 1989). This method is similar to Principal Components Regression (PCR), where the independent variable is decomposed into a set of scores, and the dependent variable is regressed on these scores instead of the original variable. The main difference being that in PLS regression, both independent and dependent variables are decomposed such that these scores have maximal covariance; that is, the relevant variations of the independent variable for predicting the dependent variable are emphasized. An extension of the PLS regression model to three-way data was proposed by Ståhle (1989). Later, Bro (1996) developed a general multiway PLS (N-PLS) regression model that was shown

to be optimal according to the theory of PLS and had a particular case numerically equivalent to that of Ståhle. N-PLS seeks in accordance with the philosophy of PLS to describe the covariance of the dependent and independent variables. This is achieved by fitting multilinear models simultaneously for independent and dependent variables and for a regression model relating the two decomposition models. On the other hand, as covariance is the product of the correlation and the variances, these three measures actually are maximized collectively.

According to Bro (1996), the model is known as N-PLS or Multilinear PLS in general, and the specific model to be used in this work is called tri-PLS2. This follows from its having a three-way decomposition for the independent variable (*tri*), which will be the time-varying EEG spectrum, and a two-way or bilinear decomposition for the dependent variable (2), corresponding to the fMRI data. This can be considered as a form of PARAFAC decomposition constrained by additional conditions of maximal covariance with certain BOLD components. The structural model can be expressed as:

$$\hat{S}_{dwt} = \sum_{k=1}^{N_k} a_{dk} b_{wk} c_{tk} + e_{dwt}$$

$$\hat{F}_{st} = \sum_{k=1}^{N_k} u_{sk} v_{tk} + \varepsilon_{st}$$

where  $\varepsilon_{st}$  and  $e_{dwt}$  are elements of noise matrices and the index,  $s$ , represents the voxels or grid points inside the brain. These decomposition models are estimated iteratively, component-wise, by finding a set of normalized vectors,  $\mathbf{a}_k$ ,  $\mathbf{b}_k$ , and  $\mathbf{u}_k$  such that the least-squares score vectors,  $\mathbf{c}_k$  and  $\mathbf{v}_k$ , have maximal covariance. It is worth underscoring that the N-PLS model is unique, as it consists of successively estimated one-atom models, each of which is itself always unique. On the other hand, note that the EEG data

must first be preprocessed, both by removing muscle and motion artifacts, replacing them by linear interpolation of the data, and by convolution of the EEG spectrum with the hemodynamic impulse response function (Cohen, 1997). A graphical representation of the tri-PLS2 method is shown in Fig. 1.

The interpretation of loading vectors is straightforward. The spectral signature of the EEG for the  $k$  atom,  $\mathbf{b}_k$ , will allow the identification of those brain rhythms whose time-varying envelopes has maximal covariances with the BOLD signal. The spatial signature of the fMRI for atom  $k$ ,  $\mathbf{u}_k$ , is a tomographic map (which is not a correlation map) showing those BOLD signals whose time courses are correlated maximally with the EEG. Finally, the spatial signature of the EEG,  $\mathbf{a}_k$ , is a representative topography of atom  $k$ , extracted by asking for the maximal temporal correlation between EEG and fMRI.

The decomposition is made component-wise; that is, for each component (atom), a rank-one model is built of both the independent variable 3D matrix  $\mathbf{S}$ , and the dependent variable 2D fMRI matrix,  $\mathbf{F}$ . These models are then subtracted from the original data, and a new atom of signatures is found from the residuals. The calculation for one atom of the tri-PLS2 model is developed in detail in Appendix A. As in PARAFAC, a convention about signs and scale is needed. In this case, the non-normalized factors will be the temporal signatures of both the EEG and the fMRI data, while the other signatures are normalized. Signs were assigned to ensure that the correlation between fMRI and EEG temporal signatures for the alpha atom were positive. Moreover, in this work, it is important to obtain smooth images as atoms of the spatial signature of the fMRI. For the sake of simplicity, the raw data can be presmoothed and the same smoothed signatures will be obtained from the decomposition (Bro, personal communication). Therefore, the raw fMRI data are presmoothed to obtain smoothed atoms for the spatial signature of fMRI. Smoothing consisted of applying the nearest neighbor moving average three times to the raw fMRI data.

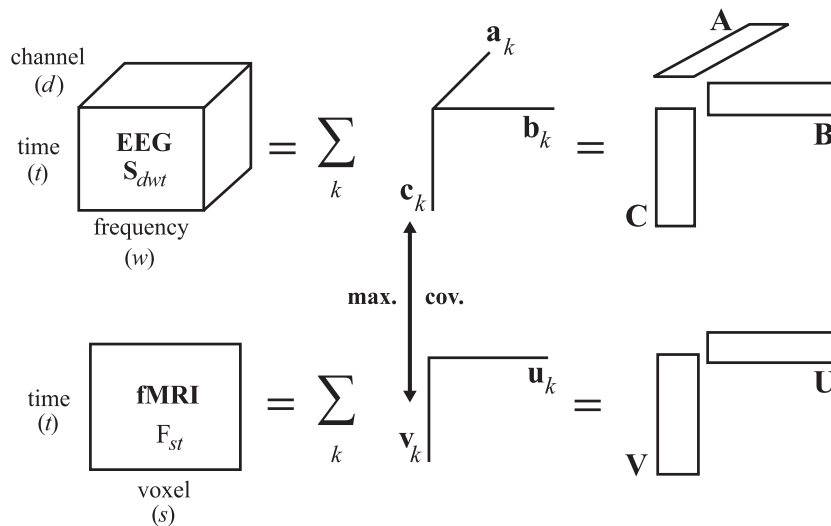


Fig. 1. Tri-PLS2 diagram. The time-varying EEG spectrum is represented as a three-dimensional array indexed by time ( $t$ ), frequency ( $w$ ), and channel ( $d$ ). The fMRI matrix is indexed by time and voxels ( $s$ ). Both data are decomposed into a sum of atoms or components. Each EEG atom have spatial ( $\mathbf{a}_k$ ), spectral ( $\mathbf{b}_k$ ), and temporal ( $\mathbf{c}_k$ ) signatures. fMRI atoms have a spatial ( $\mathbf{u}_k$ ) and temporal ( $\mathbf{v}_k$ ) signature. The extraction of atoms is performed simultaneously by constraining the temporal signatures to have maximal covariance. Joining all atoms for each signature allows them to be expressed in matricial notation, obtaining corresponding matrices  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\mathbf{C}$ ,  $\mathbf{V}$ ,  $\mathbf{U}$ .

### Assessment of N-PLS model

The advantages of N-PLS over bilinear methods are that it is much more parsimonious, easier to interpret, and less prone to noise. These advantages hold even over nonlinear calibration models (e.g., feedforward neural networks) because they are bilinear in the decomposition of the independent variable and the nonlinearity is introduced only in the relation of this decomposition with the dependent variable. Another advantage is that the algorithm is faster than other multilinear decomposition methods (e.g., PARAFAC) due to the relatively few parameters to estimate and particularly, because the tri-PLS algorithm boils down to eigenvalue problems.

However, this model has its own pitfalls. The basic problem is the appropriateness of the trilinear model. As this is a data-dependent question, there is not a general and straightforward answer. If there is no any a priori knowledge about the three-way nature of a given data, one could try different methods to see which one describes the data best. In the case of several methods fitting the data equally well, one should choose the simplest model and in this regard multilinear models are preferred over bilinear ones. On the other hand, although it has been shown that models like N-PLS seldom fail to converge and offer degenerate solutions (Bro, 1998), these are problems that can arise in multiway methods and should be taken into account in the exploration of the data.

In practice, it is convenient to apply a PARAFAC decomposition to the EEG data before applying tri-PLS2 model. This initial exploration will allow to assessing the appropriateness of the trilinear model for the time-varying EEG spectrum, to identify possible outliers in the data, and the estimation of the number of significant atoms present in the data. The implementation of PARAFAC used in this work is contained in a Matlab Toolbox developed by Bro and available on the web. It provides several diagnostic tools, such as Corcondia, residuals plots, leverages plots, convergence, and explained variance of the data, among others.

As said above, the appropriate number of components was obtained with the residual analysis and the Corcondia index. This index was also used for assessing the trilinear structure of the data as shown in Estienne et al. (2001). The analysis of leverages allowed to detecting four outliers in the time mode. These four time windows (or segments) were discarded from the data for subsequent analysis. We also removed some constant signature (nonphysiologically meaningful) in the frequency mode by an adequate centering across this mode. Furthermore, comparison between the loadings of the time-varying EEG spectrum decomposition provided by PARAFAC and those provided by tri-PLS will validate (at a preliminary level) the truthfulness of the results obtained. A detailed explanation about the use of the diagnostic tools for this exploratory analysis and discussion of the reliability of PARAFAC model can be found in Bro (1998) and Miwakeichi et al. (2004).

### Source localization analysis

The spatial signature for the time-varying EEG spectrum,  $\mathbf{a}_k$ , may be analyzed further by source reconstruction methods, such as Low-Resolution Electromagnetic Tomography (LORETA) (Pascual-Marqui et al., 1994) to find those underlying electrical sources that are correlated temporally with the BOLD signal. However,

LORETA cannot be applied directly since  $\mathbf{a}_k$  is not derived from voltages but rather from the power spectra of voltages. Therefore, in this case, we developed a procedure that allows the estimation of the spectra of the EEG sources on the basis of the spectra of the observed voltages. We shall call this type of source localization “Source Spectra Imaging” (SSI). This is based on the following assumptions:

- There is no spatial correlation between scalp voltage measurements.
- There is no spatial correlation between electric current densities inside the brain.
- The source spectra (variances of current densities in frequency domain) to be estimated will be the smoothest one in space.
- The source spectra is the same in the  $x$ ,  $y$ , and  $z$  directions.

The detailed formulation for obtaining this inverse solution can be found in a companion paper (Miwakeichi et al., 2004). It must be emphasized that the assumptions behind this inverse solution can classify it as a distributed inverse solution, whose pitfalls and drawbacks have been extensively described in the literature (Fuchs et al., 1999; Pascual-Marqui, 1999).

Moreover, the EEG data analyzed in this work corresponds to voltages measured in an array of 16 bipolar pairs; therefore, to find the SSI solution, the problem of transforming these bipolar measurements into unipolar voltages must be addressed. We must thus construct the matrix  $\mathbf{M}$  that transforms the spatial signatures of the EEG,  $\mathbf{a}_k^{\text{uni}}$ , obtained (ideally) from a unipolar array, into those measured from bipolar recordings (Eq. (3)). A partial representation of matrix  $\mathbf{M}$  is given in Eq. (4). Then,  $\mathbf{a}_k^{\text{uni}}$  is estimated by multiplying Eq. (3) by the Moore-Penrose pseudo-inverse of matrix  $\mathbf{M}$ .

$$\mathbf{a}_k = \mathbf{M}\mathbf{a}_k^{\text{uni}} \quad (3)$$

$$\mathbf{M} = \begin{array}{cccccc} Fp2 & \dots & F7 & F8 & \dots & O2 & \dots & T4 & \dots & T6 \\ 1 & \dots & 0 & -1 & \dots & 0 & \dots & 0 & \dots & 0 & Fp2 - F8 \\ 0 & \dots & 0 & 1 & \dots & 0 & \dots & -1 & \dots & 0 & F8 - T4 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & -1 & \dots & 0 & \dots & 1 & T6 - O2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \vdots \end{array} \quad (4)$$

Further, with this method, we can visualize the spatial signatures of the EEG obtained by tri-PLS2 decomposition (which would correspond to the bipolar topographies) as a topographic map on the head. Finally, it is noteworthy that these topographic maps, and their SSI solutions, are essentially dimensionless, as the former is normalized as part of the scale convention for the tri-PLS2 model.

### Statistical inference

Our first inferential problem is to determine whether there is a significant correlation between the time courses of the EEG and fMRI. This can be tested readily by permutation of the time



segments in the time-varying EEG spectrum, which will destroy any temporal correlation between the EEG and fMRI data (Galán et al., 1997). This procedure is not appropriate, however, if there is any autocorrelation in the time series of the EEG data. Using the ARFIT toolbox for Matlab (Schneider and Neumaier, 2001), we fitted an autoregressive model of order 2 (selected automatically by Schwarz's criterion) to the time course of time-varying EEG spectrum. With this information, we applied a block bootstrap method, which is adequate in the case of weak dependence of observations (time points in this case). The method consists of resampling with replacement, using blocks of consecutive time points instead of individual time points. The length of the blocks was chosen to be great enough to preserve the original dependence, so that the empirical distribution of statistics for blocks will resemble that for the original time points (Davison and Hinkley, 1997). On the other hand, it is also desirable to have as many blocks as possible. In our case, we use nonoverlapping blocks of length  $l = 2p+1$ ;  $p = 2$  being the order of the autoregressive model. Thus, by applying the tri-PLS2 method for  $N$  resampled series, we obtained  $N$  different decompositions into atoms of corresponding signatures for each modality. The correlation coefficients between the EEG and corresponding fMRI time courses for each atom were then computed. From the 95th percentile of the empirical distribution of these correlations, we established a significance level for testing of the original correlation.

Our second inferential problem is to determine which voxels in the spatial signature of the fMRI are significantly different from zero. This is important for identifying brain regions that contribute to a particular EEG-fMRI temporal correlation. Thus, for this problem, we used a simple jackknife resampling procedure (Davison and Hinkley, 1997) from which a pseudo  $t$  image was constructed. In this specific case, the jackknifed estimate was obtained as follows: the leave-one-out spatial signatures ( $\mathbf{u}_i$ ;  $i = 1 \dots N_t$ ) of the fMRI were created by leaving out time points one at a time and applying the tri-PLS2 model to the truncated data. The jackknife pseudo observations were then computed as:

$$\mathbf{u}_i^* N_t \mathbf{u} - (N_t - 1) \mathbf{u}_i; \quad i = 1 \dots N_t$$

where  $\mathbf{u}$  is the fMRI spatial signature corresponding to the complete data. This equation holds for all components although we have eliminated the subscript  $k$  for simplicity. Using the mean of the pseudo observations ( $\bar{\mathbf{u}}^* = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{u}_i^*$ ) and the standard deviations ( $\sigma_{u^*} = \sqrt{\frac{1}{N_t} \sum_{i=1}^{N_t} (\mathbf{u}_i^* - \bar{\mathbf{u}}^*)^2}$ ), the pseudo  $t$  image for each atom can be computed as  $\mathbf{t}_{\text{image}} = \sqrt{N_t} \frac{\bar{\mathbf{u}}^*}{\sigma_{u^*}}$ .

#### Experimental data

The EEG was sampled at 200 Hz from an array of 16 bipolar pairs, (Fp2-F8, F8-T4, T4-T6, T6-O2, O2-P4, P4-C4, C4-F4, F4-Fp2; Fp1-F7, F7-T3, T3-T5, T5-O1, O1-P3, P3-C3, C3-F7, F7-Fp1), with an additional channel for the EKG and scan trigger. The fMRI time series was measured in six slice planes (4 mm, skip 1 mm) parallel to the AC-PC line, with the second from the bottom slice through AC-PC. More details about this data set can be found in Goldman et al. (2002). In the work presented here, we have analyzed five simultaneous EEG/fMRI recordings from three different subjects. Informed consent was obtained from all volunteers based on a protocol approved previously by the UCLA Office for the Protection of Research Subjects.

## Results

Both PARAFAC and N-PLS techniques were applied to the recorded data sets, and yielded similar results for all subjects. There was no statistical inference about differences among subjects, so, for the purpose of this paper, we present representative data from a single subject. As a first exploration of the data, a PARAFAC model was fitted to the time-varying EEG spectrum. The appropriate number of components for this model was chosen using Corcondia (see above). The model was fitted using direct trilinear decomposition for its initial values.

Three significant atoms or components, characterized by their spectral signature, were extracted by PARAFAC (Fig. 2A). It is

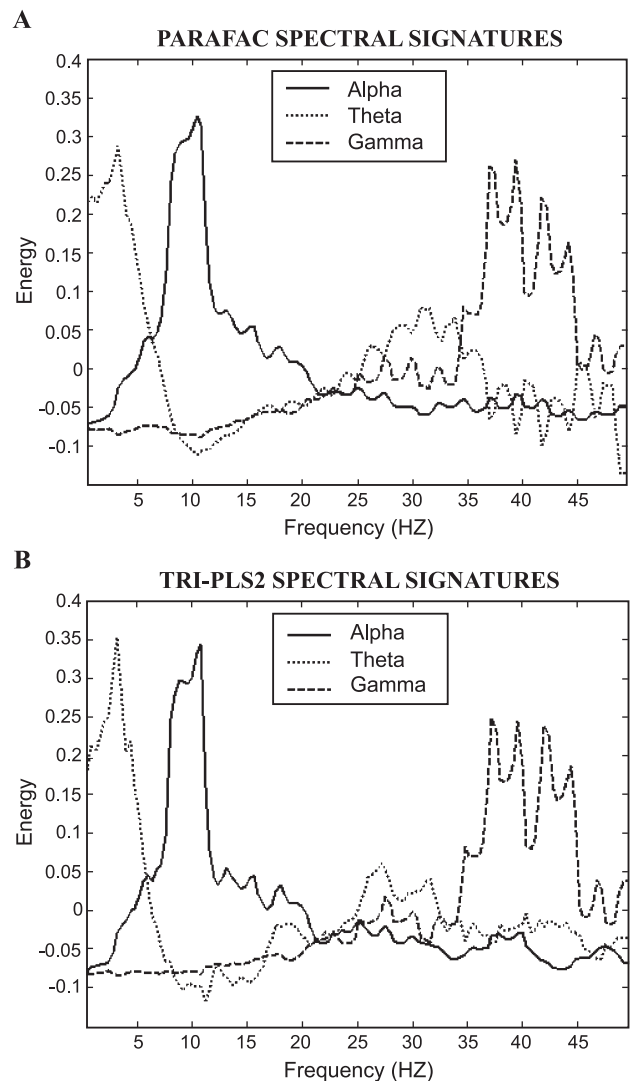


Fig. 2. Spectral signatures of the EEG decomposition. (A) Spectral signatures obtained from PARAFAC decomposition of the time-varying EEG spectrum. Three atoms were extracted. The first has a spectral peak around 10 Hz, corresponding to the well-known alpha rhythm. The second has a spectral peak around 4 Hz, which is a value usually assigned to theta activity. The third atom corresponds to a fast activity with spectral peaks from 35 to 45 Hz, in the gamma range. (B) Spectral signatures of the time-varying EEG spectrum, obtained from the tri-PLS2 model. Three atoms or components were extracted. These spectra resemble strongly those obtained from PARAFAC decomposition.

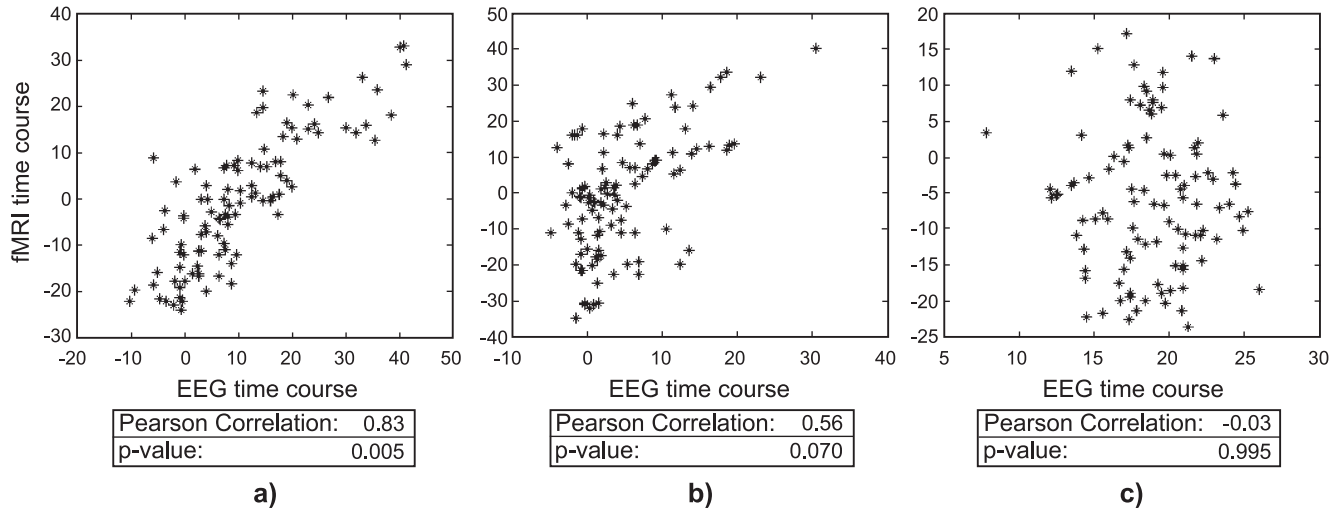


Fig. 3. Scatter plot of fMRI temporal signature against EEG temporal signature. (a) Alpha atom. A nearly linear positive dependence can be seen. The Pearson correlation value is 0.83, corresponding to  $P = 0.005$ . (b) Theta atom. The linear dependence between the EEG time course and fMRI time course has a positive correlation value of 0.56. However, it is not significant,  $p = 0.07$ . (c) Gamma atom. There is no clear linear dependence.

easy to recognize the alpha atom with its peak near 10 Hz. A slower theta activity peak is also present with a maximum around 4 Hz, as is a gamma peak in the range from 35 to 45 Hz. The Corcondia for this fit was around 93%, and the explained variation of the data was 53.5%. Moreover, PARAFAC allowed identification of outliers in the temporal signature, which were eliminated from the data for subsequent PARAFAC and posterior analyses.

Using this information, the N-PLS model was applied for only three atoms. In Fig. 2B, the spectral signatures for all atoms are shown, and they resemble strongly the spectra found by PAR-

AFAC decomposition. Fig. 3 shows scatter plots of the temporal signatures of the fMRI vs. the EEG separately for each atom. The alpha and theta activities seem to have clearly positive correlations, but gamma activity does not. The Pearson correlation values are shown for each activity band. Supporting the visual impression, correlations were highest for the alpha atom. By using the 1000 samples of the block bootstrap test described previously, only the alpha atom presents a correlation value with probability lower than 0.05. The theta atom has a non-negligible correlation value, whose empirical probability is slightly higher than the predetermined theoretical significance level of 0.05.

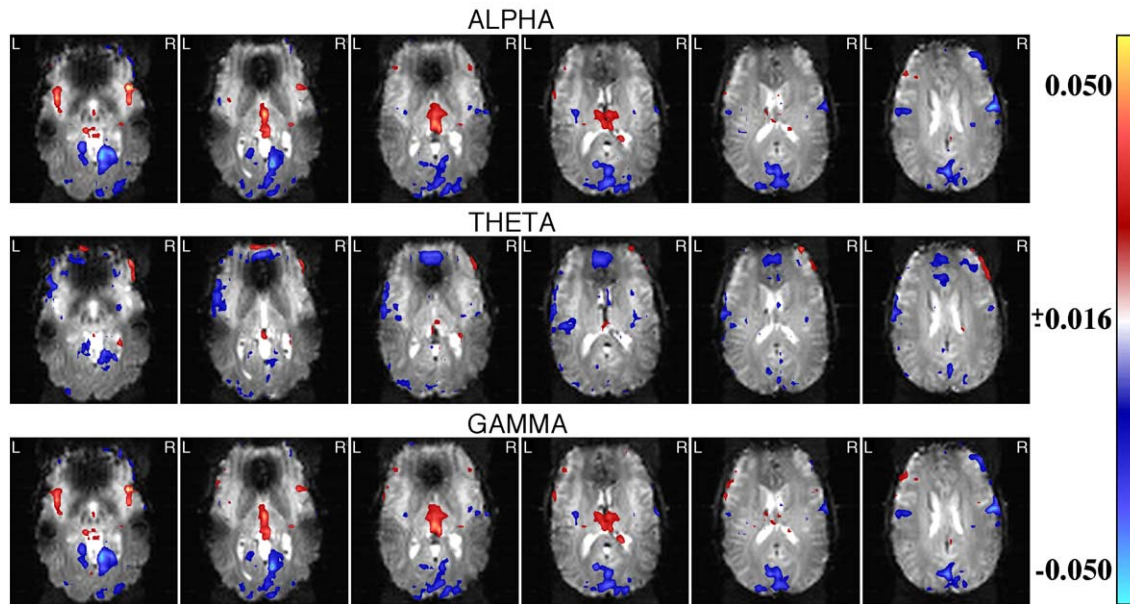


Fig. 4. fMRI spatial signatures for the three atoms. All images were plotted following a color scale from  $-0.05$  to  $0.05$ . However, the components have different minimum and maximum values. Alpha and Gamma atoms have a maximum of 0.055 and a minimum value of  $-0.066$ . Theta atom has a maximum value of 0.037 and a minimum of  $-0.066$ . The threshold was chosen conveniently to 0.016 for better visualization of the areas with higher values.

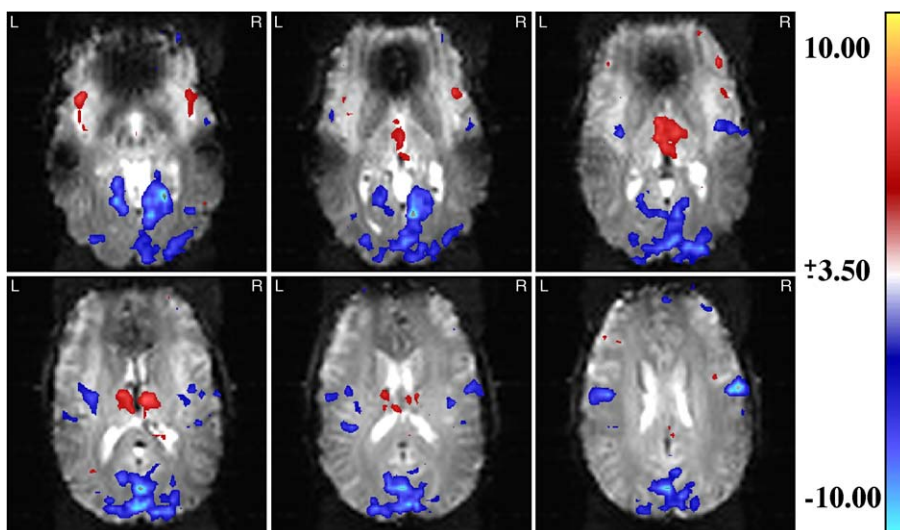


Fig. 5. Jackknifed pseudo  $t$  image for the fMRI spatial signature of the alpha rhythm atom. The jackknife procedure consisted of leaving out temporal points one at a time and applying the tri-PLS2 model to the truncated data. Then, a  $t$  value was calculated for each voxel and the resulting image was thresholded to a significance value of  $\pm 3.5$ . Blue regions (anterior median occipital, lateral occipital, occipital pole, and left and right temporal superior) represent those areas with significantly negative temporal correlation with EEG. Thalamus and insula are red representing a significant positive correlation between EEG and fMRI time courses.

Fig. 4 shows the spatial signature of the fMRI decomposition: the  $u_k$  vectors. These are shown as tomograms in which those regions that have negative temporal correlation between EEG and fMRI are blue and those that have positive temporal correlation appear in red. For the alpha atom, the fMRI spatial signature shows positive activation of thalamus and insula, while occipital and superior temporal regions are activated negatively. The theta atom showed predominantly negative activation of anterior cingulate and

occipital regions, while the gamma atom resembles the alpha component. For testing the robustness of this type of image, a pseudo  $t$  image of the alpha atom was calculated, it being the only atom having a significant temporal correlation with the EEG. This image is shown in Fig. 5, and was achieved by the jackknife procedure described above. In this figure, blue regions, (anterior median occipital, lateral occipital, occipital pole, and left and right temporal superior) represent those areas with significant negative

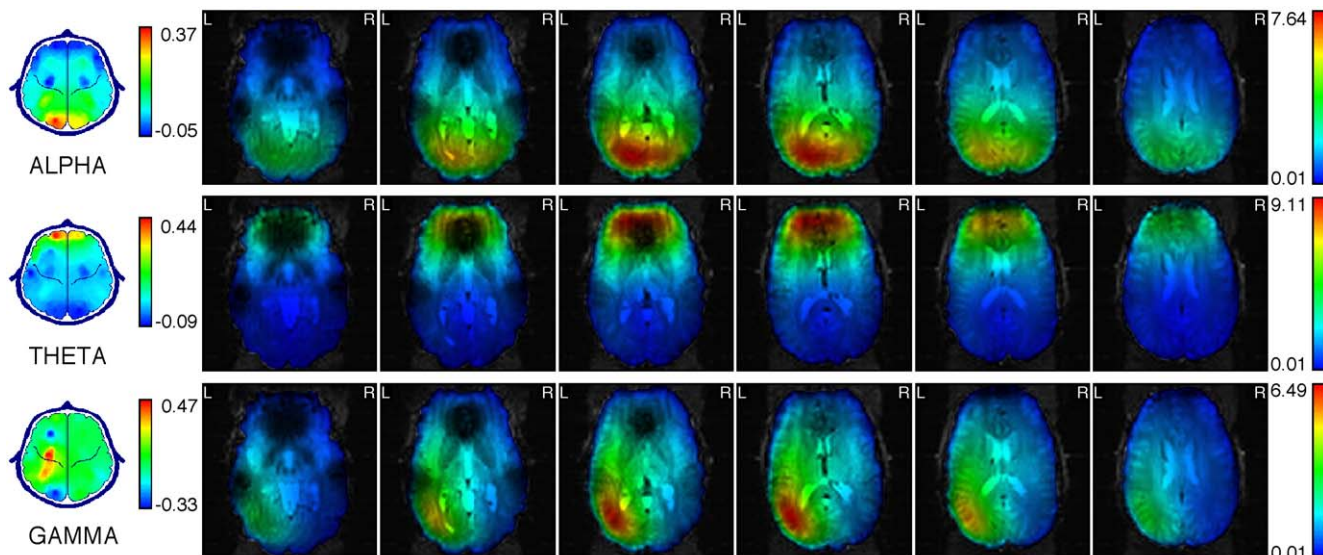


Fig. 6. Spatial signatures of the EEG and its SSI solutions. The topographical representation of spatial signatures of the EEG is shown at the far left. This map was calculated by pseudo-inverting the matrix that transforms topographies from unipolar recordings into those obtained with bipolar derivations. The alpha atom shows higher values at posterior regions, the theta topography has higher values located in frontal regions, and the gamma atom shows maximum values in the left parieto-temporal area. The corresponding SSI solutions are to the right. Maximum activation for the alpha component is located in the occipital area, with higher activation in the left hemisphere. Theta sources are in the anterior cingulate region, and the activated region for the gamma atom is located in the parieto-temporal area. Units for inverse solutions are ignored because energy values for the topographies are plotted and have been normalized as part of the N-PLS algorithm.

temporal correlation with the EEG. The areas corresponding to thalamus and insula are red, representing a significant positive correlation between EEG and fMRI time courses.

From the spatial signature,  $\mathbf{a}_k$ , of the time-varying EEG spectrum, we estimated those regions inside the brain that contribute to the EEG and that are correlated temporally with fMRI. Fig. 6 shows the topographies or EEG spatial signatures, and their corresponding SSI solutions (current density spectra) for each atom. The topography of the alpha atom shows higher values at posterior regions; the theta topography has higher values in frontal regions, and the gamma atom shows maximum topographic values in the left parieto-temporal area. The Source Spectra Imaging solution for the alpha component showed its maximum activation in the occipital area, with higher activation in the left hemisphere. Sources for theta atom are in the anterior cingulate region, and the activated region for the gamma atom is in the parieto-temporal area.

## Discussion

This paper introduces a new method, trilinear Partial Least Squares (tri-PLS2), for the analysis of concurrent EEG/fMRI recordings. This is the first use of Partial Least-Squares techniques to carry out multimodal neuroimaging fusion. Our objective is to identify the coherent systems of neural oscillators that contribute to the spontaneous EEG. Doing so requires the solution of three related problems: (i) decomposing the EEG, in the space–frequency–time domain, into a set of components or atoms, (ii) establishing the relation of these EEG components to concurrent BOLD fluctuations, and (iii) analyzing the sources of the EEG atoms. We shall consider each of these problems in turn. At the outset it should be stressed that the two phenomena—EEG and BOLD—evolve over very different time scales. In fact, we shall be analyzing the *evolutionary spectrum* (Priestley, 1965) of the EEG, a concept based on a locally stationary modeling of the electroencephalogram (Dahlhaus, 1997). It is only the *envelope* of the waves usually analyzed by electroencephalography that will be matched to BOLD.

### Atomic decomposition of the EEG

The analysis of the *evolutionary spectrum* of the EEG produces a three-dimensional data array (space–frequency–time). The first choices that come to mind for the decomposition of this array are either Principal Components Analysis (PCA) or Independent Components Analysis (ICA), a set of techniques that have received much recent attention. We decided to avoid these methods, however, for two reasons: First, they achieve a unique decomposition into atoms only by imposing arbitrary mathematical constraints (orthogonality and independence, respectively), and second, these methods are targeted toward two-dimensional arrays (matrices). In our situation, this means “unfolding” the data, stacking the time and frequency components along one dimension, and thereby destroying their distinction; keeping these different dimensions separate seems a much better alternative. A first attempt at a space–frequency–time atomic decomposition was reported in a paper by Koenig et al. (2001). In their method, the decomposition is carried out in several stages; first by the identification of time–frequency atoms, and then by the estimation of distinct topographies that are stable over time. This separation into

two stages of analysis is not conceptually necessary, and in fact is not optimal.

Our trilinear method based on Parallel Factor Analysis, introduced in the present paper, allows space–frequency–time estimation in a single step, by minimization of an explicit objective function. The resulting decomposition is intrinsically unique and specifies atoms that are defined as spectral components that vary over time and have a specific topography. A more detailed description of the combined use of PARAFAC, and distributed inverse solutions for in vivo imaging of neural oscillatory systems, is the subject of a companion paper (Miwa-keichi et al., 2004). A consistent finding in all data sets analyzed was the appearance of three components whose peaks were within the traditional theta, alpha, and gamma bands. Thus, when looking at the relations of the EEG with BOLD, it is potentially important not to constrain the analysis to a single frequency band, as was done by Goldman et al. (2002), although the present data do not show strong fMRI correlation with the EEG signal in the other bands.

It is remarkable that the restriction of maximal correlation with the BOLD signal produces spectra that are practically the same as those obtained by the PARAFAC decomposition. Based on the diagnostic tools, the physiological interpretability, and the replicability among several data sets, we can say that these are meaningful results, although we cannot ensure that they correspond with the real underlying physical phenomena. Therefore, it can be concluded that we have obtained robust and physiologically meaningful results with the use of tri-PLS algorithm.

### Relating EEG atoms to the BOLD signal

As shown here, it is possible to constrain the trilinear EEG atomic decomposition further by requiring maximal temporal correlation with the BOLD signal, a procedure that extends the classical Partial Least-Squares technique. It is important to say that the correlation found between both temporal signatures was assessed by a block bootstrap method, which is a strong diagnostic tool for obtaining reliable results. Furthermore, the spatial signature of the fMRI was also statistically validated with the use of a jackknife procedure. These kind of diagnostic tools provide additional evidence on the robustness of the model assumed, i.e., about how well the properties of the data fit the assumptions of the model.

The alpha component has a temporal relation to the BOLD signal that is significant. The regional distribution of the fMRI spatial factor corresponds closely to that described by Goldman et al. (2002), for alpha activity, thus confirming their conclusions. Since the correlation between the EEG and BOLD temporal factors are positive, it becomes clear that the image shown in Fig. 5 is equivalent to the correlation map presented by that group. In particular, there is a positive relation between thalamic and insular BOLD activity and the EEG time course for alpha component. On the other hand, the BOLD signals within parieto-occipital and somatosensory cortices are related inversely to EEG. This latter negative correlation is probably due to a decrease in the amplitude of the EEG in activated cortex in this band, resulting from the temporal resynchronization of the postsynaptic potentials of the involved neural circuits.

The extracted theta component showed moderate temporal correlations that did not reach the pre-established 0.05 level of statistical significance. An examination of the spatial distribution

of the fMRI spatial signature for this atom shows a frontal activation. It is tempting to speculate that this component corresponds to a frontal midline theta rhythm that has not been adequately resolved due to the limited spatial coverage of the brain by the fMRI protocol used. The gamma component was not correlated with the recorded BOLD signal. Once again, we cannot exclude the possibility that better spatial coverage of the brain might reveal such correlations. Further, we can speculate that gamma fluctuations might relate to dynamic and transient assemblies of systems of brain activation (Tallon-Baudry and Bertrand, 1999) that are not stable throughout the recording period.

#### Analyzing the sources of the EEG atoms

A strength of both PARAFAC and tri-PLS2 is that they identify definite topographic patterns that can be subjected to source localization. These inverse solutions interpreted together with the fMRI spatial factors provide new information on the sources of EEG rhythms.

The Source Spectra Imaging solution for the alpha component reveals activation predominantly in the parieto-occipital region. This corresponds with results on the origins of alpha rhythm that have been reported previously, both using a frequency domain dipole solution (Valdés-Sosa et al., 1998) as well as frequency domain distributed solutions (Casanova et al., 2000). An interesting fact is that the thalamus shows very little activation, in contrast to the high positive correlation found by Goldman et al. (2002), and confirmed by the tri-PLS2 fMRI spatial signature.

This dissociation between the sources of the spatial signature of the EEG atoms and the spatial signature of the fMRI of the alpha atom is likely due to the negligible contribution of the primary current sources of thalamic neurons to the scalp EEG. In this case, the observed correlations between thalamic BOLD and EEG must be indirect. For example, the thalamus is probably correlated negatively with the parieto-occipital cortex, which seems to be the location of the generators of the “EEG alpha rhythm”. Because the BOLD signal in this region is also correlated negatively with the alpha EEG spectrum, this would explain a positive correlation between alpha power and thalamic metabolic activity as an indirect effect through parieto-occipital cortex. In the terminology of Friston et al. (1996), there is a functional connectivity between the EEG and thalamus, but the effective connectivity path would not be direct, being mediated instead by the parieto-occipital cortex. Thus, according to the definitions given here, insula, thalamus, and parieto-occipital cortex are generators of the “alpha brain rhythm” while only parieto-occipital cortex contribute to the “EEG alpha rhythm”. We note that the analyses presented in this paper do not allow the distinction of whether a structure belonging to a rhythm generating system oscillates in that frequency range. It seems unlikely that joint EEG/fMRI recordings can resolve the extent of phasic, versus tonic, participation in a brain rhythm of a structure that does not produce a measurable EEG. In other words, there is still invisible information for an EEG/fMRI fusion analysis, namely, the fine temporal characteristics of those areas that are invisible in the scalp EEG. In future planned experiments, it may be possible to resolve this issue through conjoint fMRI and depth electrode studies.

The tri-PLS2 method introduced in this paper is an example of multimodal image fusion, which takes advantage of the spatial resolution of the fMRI, as well as the temporal resolution of the

EEG. This data analytic approach is capable of parsimoniously determining which EEG components are significant in the final analyses, and of revealing new features of the data by differentiating regions exposed within the fMRI data from those indicated solely through inverse solutions using the EEG. We are pursuing a number of improvements to enhance the integration of both types of data modalities by this method. In the first, we are developing a variant of tri-PLS2 that will estimate the spatial components, not on the scalp topography, as is done now, but instead directly in the source space. This would integrate source localization into the procedure rather than applying it as a postprocessing step for the topographies of the EEG atoms. Additionally, the autocorrelation of both the EEG and BOLD time series will be taken into account, whereas the model presented here ignores this information. Finally, it may well be that there are interactions between time, topography, and frequency spectrum that the current algorithm cannot account for.

#### Acknowledgments

The authors want to thank Dr. Eduardo Aubert from the Cuban Neuroscience Center for his important collaboration on handling MRI images used in this paper. Also, for their support and continuous comments on this work, we thank Nelson Trujillo, Lester Melie, and Ernesto Palmero from the Neurophysics Department of the Cuban Neuroscience Center. MSC is supported for this work under NIH DA15549 and DA13054.

Finally, we appreciate and thank Prof. Yoko Yamaguchi, head of the Laboratory for Dynamics of Emergent Intelligence of the RIKEN Brain Science Institute, for her hospitality and support for the successful conclusion of this work.

#### Appendix A. Tri-PLS2 algorithm

To calculate an atom of the tri-PLS2 model, we rewrite the model for dependent and independent variables taking only one atom,  $k$ , into account. Here the independent variable is the time-varying EEG spectrum, convolved previously with the hemodynamic response function, which is a three-way array  $\mathbf{S}$ . The dependent variable is the fMRI 2D matrix  $\mathbf{F}$ . The structural models then are

$$\hat{S}_{dwt} = a_{dk} b_{wk} c_{tk} \quad (\text{A.1})$$

and

$$\hat{F}_{st} = u_{sk} v_{tk}. \quad (\text{A.2})$$

The score vectors are those dependent on time (temporal signatures), i.e.,  $\mathbf{c}_k = (c_{1k}, \dots, c_{tk}, \dots, c_{N_t k})^T$  and  $\mathbf{v}_k = (v_{1k}, \dots, v_{tk}, \dots, v_{N_v k})^T$ ; the others are also called the weights (spatial and spectral signatures of the EEG, spatial signature of the fMRI). The indices  $t = 1, \dots, N_t$ ,  $w = 1, \dots, N_w$ ,  $d = 1, \dots, N_d$  and  $s = 1, \dots, N_s$  represent time, frequency, channels, and voxels, respectively. For given weight vectors, the least-squares solution for determining the score vectors are:

$$c_{tk} = \sum_{w=1}^{N_w} \sum_{d=1}^{N_d} S_{dwt} a_{dk} b_{wk} \quad (\text{A.3})$$

and

$$\mathbf{v}_{tk} = \sum_{s=1}^{N_s} F_{st} \mathbf{u}_{sk}. \quad (\text{A.4})$$

Our problem is to find a set of normalized weight vectors,  $\mathbf{a}_k$ ,  $\mathbf{b}_k$ , and  $\mathbf{u}_k$ , which produce score vectors,  $\mathbf{c}_k$  and  $\mathbf{v}_k$ , having maximal covariance. The objective function to be maximized is:

$$\max_{\mathbf{a}_k, \mathbf{b}_k, \mathbf{u}_k} \left[ \sum_{t=1}^{N_t} c_{tk} v_{tk} \mid c_{tk} = \sum_{w=1}^{N_w} \sum_{d=1}^{N_d} S_{dwt} a_{dk} b_{wk} \wedge v_{tk} = \sum_{s=1}^{N_s} F_{st} u_{sk} \right] \quad (\text{A.5})$$

For simplicity, the restriction of normalization on the weight vectors is not made explicit. Eq. (A.5) is not strictly correct because there is no correction for degrees of freedom, but as this correction is constant for a given atom, it will not affect the maximization. Eq. (A.5) also does not express the covariance if  $\mathbf{S}$  and  $\mathbf{F}$  have not been centered.

The next procedure is performed in two ways. First, Eq. (A.5) could be taken to:

$$\max_{\mathbf{a}_k, \mathbf{b}_k} \left[ \sum_{t=1}^{N_t} \sum_{w=1}^{N_w} \sum_{d=1}^{N_d} S_{dwt} v_{tk} a_{dk} b_{wk} \right] = \max_{\mathbf{a}_k, \mathbf{b}_k} \left[ \sum_{w=1}^{N_w} \sum_{d=1}^{N_d} z_{dwt} a_{dk} b_{wk} \right], \quad (\text{A.6})$$

where  $z_{dwt} = \sum_{t=1}^{N_t} S_{dwt} v_{tk}$  are the elements of an auxiliary matrix,  $\mathbf{Z}_k$ . If one writes Eq. (A.6) in matrix notation, the equation will become:

$$\max_{\mathbf{a}_k, \mathbf{b}_k} [\mathbf{b}_k^T \mathbf{Z}_k \mathbf{a}_k] \Rightarrow (\mathbf{b}_k, \lambda_k, \mathbf{a}_k) = \text{SVD}(\mathbf{Z}_k, 1). \quad (\text{A.7})$$

In other words, the weight vectors,  $\mathbf{a}_k$  and  $\mathbf{b}_k$  can be computed from the first component of a singular value decomposition of  $\mathbf{Z}_k$  [SVD( $\mathbf{Z}_k, 1$ )]. This follows directly from the properties of SVD.

Second, substituting in Eq. (A.5) the corresponding score vector for the dependent variable:

$$\max_{\mathbf{u}_k} \left[ \sum_{t=1}^{N_t} \sum_{s=1}^{N_s} F_{st} c_{tk} u_{sk} \right] = \max_{\mathbf{u}_k} \left[ \sum_{s=1}^{N_s} y_{sk} u_{sk} \right], \quad (\text{A.8})$$

where  $y_{sk} = \sum_{t=1}^{N_t} F_{st} c_{tk}$  are the elements of an auxiliary vector  $\mathbf{y}_k$ . Since  $\mathbf{u}_k$  is restricted to be normalized, the maximum value of the expression (A.8) is reached when  $\mathbf{u}_k$  is a unit vector in the same direction as  $\mathbf{y}_k$ . Therefore, the solution is:

$$\mathbf{u}_k = \frac{\mathbf{y}_k}{\|\mathbf{y}_k\|} = \frac{\mathbf{F}^T \mathbf{c}_k}{\|\mathbf{F}^T \mathbf{c}_k\|} \quad (\text{A.9})$$

On the other hand, through the models of the data sets given in Eqs. (A.1) and (A.2), the prediction model between  $\mathbf{S}$  and  $\mathbf{F}$  is found by using a regression model for the so-called inner relation (established for the loadings matrices, i.e., for all atoms at the same time):

$$\mathbf{V} = \mathbf{C}\mathbf{X} + \mathbf{E}_v.$$

This expression ensures that the maximum covariance restriction holds, and allows prediction of new samples of dependent variables. As the different atoms for score vectors are not always

orthogonal, all of these atoms must be taken into account in calculating regression coefficients. The regression thus leads to:

$$\mathbf{x}_k = (\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \mathbf{v}_k. \quad (\text{A.10})$$

Finally, we can summarize the algorithm as follows:

1. Center  $\mathbf{S}$  and  $\mathbf{F}$ .
2. Let  $\mathbf{v}_k$  equal a column in  $\mathbf{F}$ .
3. Atom  $k = 1$ .
4. Compute matrix  $\mathbf{Z}_k$  using  $\mathbf{S}$  and  $\mathbf{v}_k$ .
5. Determine  $\mathbf{a}_k$  and  $\mathbf{b}_k$  from Eq. (A.7).
6. Calculate  $\mathbf{c}_k$  from Eq. (A.3).
7. Compute  $\mathbf{u}_k$  from Eq. (A.9).
8. Compute  $\mathbf{v}_k$  from Eq. (A.4).
9. If the results converge, continue. Otherwise go to step 4.
10. Do the regression, finding  $\mathbf{x}_k$  from Eq. (A.10).
11.  $\mathbf{S}_t = \mathbf{S}_t - c_{tk} \mathbf{b}_k \mathbf{a}_k^T$  (for all  $t$ ) and  $\mathbf{F} = \mathbf{F} - \mathbf{C} \mathbf{x}_k \mathbf{u}_k^T$ .
12.  $k = k + 1$ . Repeat from 4 until  $\mathbf{F}$  is properly described.

## References

- Babiloni, F., Babiloni, C., Carducci, F., Angelone, L., Del Gratta, C., Romani, G.L., Rossini, P.M., Cincotti, F., 2001. Linear inverse estimation of cortical sources by using high resolution EEG and fMRI priors. *IJBEM* 3, 1.
- Baillet, S., Leahy, R.M., Singh, M., Shattuck, D.W., Mosher, J.C., 2001. Supplementary motor area activation preceding voluntary finger movements as evidenced by magnetoencephalography and fMRI. *IJBEM* 3, 1.
- Bro, R., 1996. Multi-way calibration. *Multi-linear PLS*. *J. Chemom.* 10, 47–61.
- Bro, R., 1998. *Multi-way Analysis in the Food Industry: Models, Algorithms and Applications*. PhD Thesis. University of Amsterdam (NL) and Royal Veterinary and Agricultural University (DK).
- Carroll, J.D., Chang, J., 1970. Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart–Young' decomposition. *Psychometrika* 35, 283–319.
- Casanova, R., Valdés-Sosa, P.A., García, F., Aubert, E., Riera, J., Korin, W., Lins, O., 2000. Frequency domain distributed inverse solutions. In: Aine, C.J., Okada, Y., Stroink, G., Swithenby, S.J., Wood, C.C. (Eds.), *Biomag 96, Proceedings of the 10th International Conference on Biomagnetism*. Springer-Verlag, New York (ISBN:0387989153).
- Chen, S., Donoho, D., Saunders, M., 2001. Atomic decomposition by basis pursuit. *SIAM Rev.* 43 (1), 129–159.
- Churchland, P.S., Sejnowski, T.J., 1988. Perspectives on cognitive neuroscience. *Science* 242 (4879), 741–745.
- Cohen, M.S., 1997. Parametric analysis of fMRI data using linear systems methods. *NeuroImage* 6 (2), 93–103.
- Dahlhaus, R., 1997. Fitting time series models to non-stationary processes. *Ann. Stat.* 25, 1–37.
- Davison, A.C., Hinkley, D.V., 1997. In: Gill, R., Ripley, B.D., Ross, S., Stein, M., Williams, D. (Eds.), *Bootstrap Methods and Their Application*. Cambridge Univ. Press, UK.
- de Jong, S., Phatak, A., 1997. Partial least squares regression. Recent advances in total least squares techniques and errors—in variables modeling. In: Van Huffel, E. (Ed.), *SIAM*, Philadelphia.
- Düzel, E., Habib, R., Schott, B., Schoenfeld, A., Lobaugh, N., McIntosh, A.R., Scholz, M., Heinze, H.J., 2003. A multivariate, spatiotemporal analysis of electromagnetic time–frequency data of recognition memory. *NeuroImage* 18, 185–197.

- Estienne, F., Matthijs, N., Massart, D.L., Ricoux, P., Leibovici, D., 2001. Multi-way modelling of high-dimensionality electroencephalographic data. *Chemom. Intell. Lab. Syst.* 58, 59–71.
- Field, A.S., Graupe, D., 1991. Topographic component (Parallel Factor) analysis of multichannel evoked potentials: practical issues in trilinear spatiotemporal decomposition. *Brain Topogr.* 3 (4), 407–423.
- Friston, K.J., Frith, C.D., Fletcher, P., Liddle, P.F., Frackowiak, R.S.J., 1996. Functional topography: multidimensional scaling and functional connectivity in the brain. *Cereb. Cortex* 6, 156–164.
- Fuchs, M., Wagner, M., Kohler, T., Wischman, H.A., 1999. Linear and nonlinear current density reconstructions. *J. Clin. Neurophysiol.* 16 (3), 267–295.
- Galán, L., Biscay, R., Rodríguez, J.L., Pérez Abalo, M.C., Rodríguez, R., 1997. Testing topographic differences between event related brain potentials by using non-parametric combinations of permutation tests (published erratum appears in *Electroencephalogr. Clin. Neurophysiol.* 107(1998 Nov)(5): 380–381). *Electroencephalogr. Clin. Neurophysiol.* 102 (3), 240–247.
- Goldman, R.I., Stern, J.M., Engel, J., Cohen, M.S., 2000. Acquiring simultaneous EEG and functional MRI. *Clin. Neurophysiol.* 111, 1974–1980.
- Goldman, R.I., Stern, J.M., Engel, J., Cohen, M.S., 2002. Simultaneous EEG and fMRI of the alpha rhythm. *NeuroReport* 13 (18), 2487–2492.
- Gonzalez-Andino, S.L., Blanke, O., Lantz, G., Thut, G., Grave de Peralta, R., 2001. The use of functional constraints for the neuroelectromagnetic inverse problem: alternatives and caveats. *IJBEM* 3, 1.
- Harshman, R.A., 1970. Foundations of the PARAFAC procedure: models and conditions for an ‘explanatory’ multi-modal factor analysis. *UCLA Work. Pap. Phon.* 16, 1–84.
- Horwitz, B., Poeppel, D., 2002. How can EEG/MEG and fMRI/PET data be combined? *Hum. Brain Mapp.* 17, 1–3.
- Ioannides, A.A., 1999. Problems associated with the combination of MEG and fMRI data: theoretical basis and results in practice. In: Yoshimoto, T., Kotani, M., Kuriki, S., Karibe, H., Nakasato, N. (Eds.), *Recent Advances in Biomagnetism*. Tohoku University Press, Sendai, pp. 133–136.
- Kiers, H.A.L., 1991. Hierarchical relations among three-way methods. *Psychometrika* 56, 449–470.
- Koenig, T., Martí-López, F., Valdés-Sosa, P.A., 2001. Topographic time–frequency decomposition of the EEG. *NeuroImage* 14, 383–390.
- Kruskal, J.B., 1976. More factors than subjects, test and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, 41.
- Kruskal, J.B., 1977. Three-way arrays: rank and uniqueness of trilinear decomposition with applications to arithmetic complexity and statistics. *Linear Algebra Appl.* 18, 95–138.
- Lobaugh, N.J., West, R., McIntosh, A.R., 2001. Spatiotemporal analysis of experimental of experimental differences in event-related potential data with partial least squares. *Psychophysiology* 38, 517–530.
- Martens, H., Naes, T., 1989. *Multivariate Calibration*. Wiley, Chichester.
- McIntosh, A.R., Bookstein, F.L., Haxby, J.V., Grady, C.L., 1996. Spatial pattern analysis of functional brain images using Partial Least Square. *NeuroImage* 3, 143–157.
- Miwakeichi, F., Martínez-Montes, E., Valdés-Sosa, P.A., Mizuhara, H., Nishiyama, N., Yamaguchi, Y., 2004. Decomposing EEG data into space–time–frequency components using parallel factor analysis. *NeuroImage* 22, 1035–1045.
- Möcks, J., 1988a. Decomposing event-related potentials: a new topographic components model. *Biol. Psychol.* 26, 199–215.
- Möcks, J., 1988b. Topographic components model for event-related potentials and some biophysical considerations. *IEEE Trans. Biomed. Eng.* 35, 482–484.
- Pascual-Marqui, R.D., 1999. Review of methods for solving the EEG inverse problem. *Int. J. Bioelectromagn.* 1 (1), 75–86.
- Pascual-Marqui, R.D., Valdés-Sosa, P.A., Alvarez, A., 1988. A parametric model for multichannel EEG spectra. *Int. J. Neurosci.* 40, 89–99.
- Pascual-Marqui, R.D., Michel, C.M., Lehmann, D., 1994. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *Int. J. Psychophysiol.* 18, 49–65.
- Priestley, M.B., 1965. Evolutionary spectra and non-stationary processes. *J. R. Stat. Soc., Ser. B Stat. Methodol.* 27, 204–237.
- Schneider, T., Neumaier, A., 2001. Algorithm 808: ARfit—A Matlab package for the estimation of parameter and eigenmodes of multivariate autoregressive models. *ACM Trans. Math. Softw.* 27 (1), 58–65.
- Singh, M., Patel, P., Al-Dayeh, L., 1998. fMRI of brain activity during alpha rhythm. *Int. Soc. Mag. Res. Med.* 3, 1493.
- Stähle, L., 1989. Aspects of analysis of three-way data. *Chemom. Intell. Lab. Syst.* 7, 95–100.
- Szava, S., Valdés-Sosa, P.A., Biscay, R., Galán, L., Bosch, J., Clark, I., Jiménez, J.C., 1994. High resolution quantitative EEG analysis. *Brain Topogr.* 6, 211–219.
- Tallon-Baudry, C., Bertrand, O., 1999. Oscillatory gamma activity in humans and its role in object representation. *Trends Cogn. Sci.* 3 (4), 151–162.
- Thomson, D.J., 1982. Spectrum estimation and harmonic analysis. *Proc. IEEE* 70, 1055–1096.
- Trujillo, N.J., Martínez, E., Melie, L., Valdés, P.A., 2001. A Symmetrical Bayesian Model for fMRI and EEG/MEG Neuroimage Fusion. *IJBEM* 3, 1.
- Valdés-Sosa, P.A., Bosch, J., Virués, T., Aubert, E., Fermín, E., González, E., 1998. EEG source frequency domain SPM. *NeuroImage* 7 (4), 636.

## **Estimating brain functional connectivity with sparse multivariate autoregression**



# Estimating brain functional connectivity with sparse multivariate autoregression

Pedro A. Valdés-Sosa\*, Jose M. Sánchez-Bornot, Agustín Lage-Castellanos, Mayrim Vega-Hernández, Jorge Bosch-Bayard, Lester Melie-García and Erick Canales-Rodríguez

*Cuban Neuroscience Center, Avenue 25, No. 15202 esquina 158 Cubanacan, PO Box 6412 Playa, Area Code 11600 Ciudad Habana, Cuba*

There is much current interest in identifying the anatomical and functional circuits that are the basis of the brain's computations, with hope that functional neuroimaging techniques will allow the *in vivo* study of these neural processes through the statistical analysis of the time-series they produce. Ideally, the use of techniques such as multivariate autoregressive (MAR) modelling should allow the identification of effective connectivity by combining graphical modelling methods with the concept of Granger causality. Unfortunately, current time-series methods perform well only for the case that the length of the time-series  $Nt$  is much larger than  $p$ , the number of brain sites studied, which is exactly the reverse of the situation in neuroimaging for which relatively short time-series are measured over thousands of voxels. Methods are introduced for dealing with this situation by using sparse MAR models. These can be estimated in a two-stage process involving (i) penalized regression and (ii) pruning of unlikely connections by means of the local false discovery rate developed by Efron. Extensive simulations were performed with idealized cortical networks having small world topologies and stable dynamics. These show that the detection efficiency of connections of the proposed procedure is quite high. Application of the method to real data was illustrated by the identification of neural circuitry related to emotional processing as measured by BOLD.

**Keywords:** functional connectivity; fMRI; variable selection; sparse multivariate autoregressive model; graphical model

## 1. INTRODUCTION

There is much current interest in identifying the anatomical and functional circuits that we believe are the basis of the brain's computations (Varela *et al.* 2001). Interest in neuroscience has shifted away from mapping sites of *activation*, towards identifying the *connectivity* that weave them together into dynamical systems (Lee *et al.* 2003; Bullmore *et al.* 2004). More importantly, the availability of functional neuroimaging techniques, such as fMRI, optical images, and EEG/MEG, opens hope for the *in vivo* study of these neural processes through the statistical analysis of the time-series they produce. Unfortunately, the complexity of our object of study far outstrips the amount of data we are able to measure. Activation studies already face the daunting problem of analysing large amounts of correlated variables, measured on comparatively few observational units. These problems escalate when all pairs of relations between variables are of interest—a situation that has led some to consider that the concept of connectivity itself is 'elusive' (Horwitz 2003).

A neural system is an instance of a complex network. A convenient representation is that of a graph (figure 1) defined by a set of nodes that represents observed or unobserved (latent) variables, a set of edges, that indicate relations between nodes, and a set of probability statements about these relations (Speed & Kiiveri 1986; Wermuth & Lauritzen 1990; Cowell *et al.* 1999; Jensen 2002; Jordan 2004). Graphs, with only undirected edges, have been extensively used in the analysis of covariance relations (Wermuth & Lauritzen 1990; Wermuth & Cox 1998, 2004), but do not attempt causal interpretations. Neuroimaging studies based on this type of model will identify what Friston has defined as 'functional connectivity' (Friston 1994). To apply graphical models to functional neuroimaging data, one must be aware of the additional specificity that they are vector-valued time-series, with  $\mathbf{y}_{t(p \times 1)} = \{y_{t,i}\}_{1 \leq i \leq p, 1 \leq t \leq Nt}$  the vector of observations at time  $t$ , observed at  $Nt$  time instants. The  $p$  components of the vector are sampled at different nodes or spatial points in the brain. There has been much recent work in combining graphical models with multiple time-series analysis. An excellent example of the use of undirected graphs in the frequency domain is Bach & Jordan (2004) with applications to fMRI functional connectivity in Salvador *et al.* (2005).

A different line of work is represented by Pearl (1998, 2000, 2003) and Spirtes *et al.* (1991, 1998, 2000), among others, who studied graphs with directed

\* Author for correspondence (peter@cneuro.edu.cu, pedro.valdes.sosa@gmail.com, peter\_valdes@hotmail.com, peter\_valdes@yahoo.com).

One contribution of 21 to a Theme Issue 'Multimodal neuroimaging of brain connectivity'.

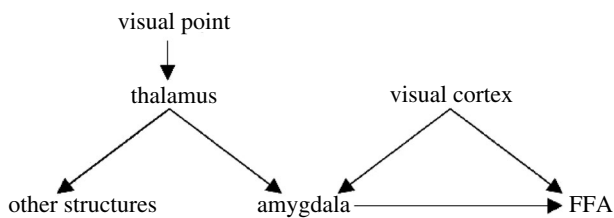


Figure 1. Directed graphical model of a (hypothetical) brain-causal network. Each node in the graph denotes a brain structure. An arrow between two nodes indicates that one structure (parent) exerts a causal influence on another node (child), a relation also known as ‘effective connectivity’. For functional images (EEG or fMRI), observations at each node are time-series. It should be noted that, optimally, time-series from all brain regions should be analysed simultaneously. Ignoring, for example, the amygdala might lead to erroneous conclusions about the influence of visual cortex on FFA, if only the latter were observed. A necessary (but not sufficient) condition for effective connectivity is that knowledge of activity in the parent improves prediction in the child (Granger causality). It is assumed that the set of directed links in real networks is *sparse* and therefore can be recovered by regression techniques that enforce this property.

edges that represent causal relations between variables. In the context of neuroimaging, searching for causality is what Friston terms the identification of effective connectivity. We will be concerned with this more ambitious type of modelling.

For functional neuroimages, the arrow of time may be used to help in the identification of causal relations. To be more specific, we model these time-series by means of a linear (stationary) multivariate autoregressive (MAR) model (Hamilton 1994; Harrison *et al.* 2003). While this type of model is very restrictive and brain-unrealistic, it will serve our purpose of developing methods for identifying connectivities in large complex neural networks for which the number of nodes  $p$  is very large compared with  $Nt$ . The general MAR model reads:

$$\mathbf{y}_t = \sum_{k=1}^{Nt} \mathbf{A}_k \mathbf{y}_{t-k} + \mathbf{e}_t \quad t = Nk + 1, \dots, Nt \quad (1.1)$$

The dynamics of the process modelled are determined by the matrices of autoregressive coefficients  $\mathbf{A}_{k(p \times p)} = \{a_{i,j}^k\}_{1 \leq i,j \leq p}$  that are defined for different time lags  $k$  and the spatial covariance matrix  $\Sigma_{(p \times p)}$  of  $\mathbf{e}_{t(p \times 1)}$ , the white-noise input process (innovations). MAR modelling has been widely applied in neuroscience research (Baccala & Sameshima 2001; Kaminski *et al.* 2001; Harrison *et al.* 2003).

Note that the coefficients  $a_{i,j}^k$  measure the influence that node  $j$  exerts on node  $i$  after  $k$  time instants. Knowing that  $a_{i,j}^k$  is non-zero is equivalent to establishing effective connectivity and is also closely related to the concept of Granger causality (Granger 1969; Kaminski *et al.* 2001; Goebel *et al.* 2003; Hesse *et al.* 2003; Valdés-Sosa 2004; Eichler 2005). The merge of causality analysis (Pearl 1998, 2000; Spirtes *et al.* 1991, 2000) with multi-time-series theory has originated graphical time-series modelling as exemplified in Brillinger *et al.* (1976); Dahlhaus (1997); Dahlhaus *et al.* (1997); Eichler (2004; 2005).

Unfortunately there is a problem with this approach when dealing with neuroimaging data: the brain is a network with extremely large  $p$ , in the order of hundreds of thousands. A ‘curse of complexity’ immediately arises. The total number of parameters to be estimated for model (1.1) is  $s = N_k \cdot p^2 + (p^2 + p)/2$ , a situation for which usual time-series methods break down. One approach to overcome this curse of complexity is to pre-select a small set of regions of interest (ROI), on the basis of prior knowledge. Statistical dependencies may then be assayed by standard methods of time-series modelling (Hamilton 1994) that in turn are specializations of multivariate regression analysis (Mardia *et al.* 1979). The real danger is the probable effect of spurious correlations induced by the other brain structures not included for study. Thus, the ideal would be to develop MAR models capable of dealing with large  $p$ .

An alternative to using ordinary multivariate regression techniques for model (1.1) is to attempt regression based on selection of variables. This could drastically reduce the number of edges in the network graph to be determined, effectively restricting our attention to networks with sparse connectivity. That this is a reasonable assumption is justified by studies of the numerical characteristics of network connectivity in anatomical brain databases (Sporns *et al.* 2000; Stephan *et al.* 2000; Hilgetag *et al.* 2002; Kotter & Stephan 2003; Sporns *et al.* 2004). The main objective of this paper is to develop methods for the identification of sparse connectivity patterns in neural systems. We expect this method to be scaled, eventually, to cope with hundreds or thousands of voxels. Explicitly, we propose to fit the model with sparsity constraints on  $\mathbf{A}_{k(p \times p)}$  and  $\Sigma_{(p \times p)}$ .

Researchers into causality (Scheines *et al.* 1998; Pearl 2000) have explored the use of regression by the oldest of variable selection techniques—stepwise selection for the identification of causal graphs. This is the basis of popular algorithms such as principal components embodied in programmes such as TETRAD. These techniques have been proposed for use in graphical time-series models by Demiralp & Hoover (2003). Unfortunately these techniques do not work well for large  $p/Nt$  ratios. A considerable improvement may be achieved by stochastic search variable selection (SSVS), which relies on Markov chain–Monte Carlo (MCMC) exploration of possible sparse networks (Dobra *et al.* 2004; Jones & West 2005). These approaches, however, are computationally very intensive and not practical for implementing a pipeline for neuroimage analysis.

A different approach has arisen in the data mining context, motivated to a great extent by the demands posed by analysis of micro-array data (West 2002; Efron *et al.* 2004; Hastie & Tibshirani 2004; Hastie *et al.* 2001). This involves extensive use of Bayesian regression modelling and variable selection, capable of dealing with the  $p \gg Nt$  situation. Of particular interest is recent work in the use of penalized regression methods for existing variable selection (Fan & Li 2001; Fan & Peng 2004) which unify nearly all variable selection techniques into an easy-to-implement iterative application of minimum norm or

ridge regression. These techniques have been shown to be useful for the identification of the topology of huge networks (Leng *et al.* 2004; Meinshausen & Bühlmann 2004).

Methods for variable selection may also be combined with procedures for the control of the false discovery rates (FDR) (Efron 2003, 2004, 2005) in situations where a large number of null hypothesis is expected to be true. Large  $p$  in this case becomes a strength instead of a weakness, because it allows the non-parametric estimation of the distribution of the null hypotheses to control false discoveries effectively.

In a previous paper, Valdés-Sosa (2004) introduced a Bayesian variant of MAR modelling that was designed for the situation in which the number of nodes far outnumbers the time instants ( $p \gg Nt$ ). This approach is, therefore, useful for the study of functional neuro-imaging data. However, that paper stopped short of proposing practical methods for variable selection. The present work introduces a combination of penalized regression with local FDR methods that are shown to achieve efficient detection of connections in simulated neural networks. The method is additionally shown to give plausible results with real fMRI data and is capable of being scaled to analyse large datasets.

It should be emphasized that in the context of functional imaging there are a number of techniques for estimating the effective connectivity, or edges, among the nodes of small pre-specified neuroanatomic graphs. These range from maximum likelihood techniques using linear and static models (e.g. structural equation modelling; McIntosh & Gonzalez-Lima 1994) to Bayesian inference on dynamic nonlinear graphical models (e.g. dynamic causal modelling; Friston *et al.* 2003). Almost universally, these approaches require the specification of a small number of nodes and, in some instances, a pre-specified sparsity structure, i.e. elimination of edges to denote conditional independence among some nodes. The contribution of this work is to enable the characterization of graphical models with hundreds of nodes using the short imaging time-series. Furthermore, the sparsity or conditional independence does not need to be specified *a priori* but is disclosed automatically by an iterative process. In short, we use the fact that the brain is sparsely connected as part of the solution, as opposed to treating it as a specification problem.

The structure of this paper is as follows. The subsequent section introduces a family of penalized regression techniques useful for identifying sparse effective connectivity patterns. The effectiveness of these methods for detecting the topology of large complex networks is explored in §2 by means of extensive simulations and is quantified by means of ROC measures. These methods are then applied together with local FDR techniques to evaluate real fMRI data. The paper concludes with a discussion of implications and possible extensions.

## 2. SPARSE MAR MODELS

We now describe a family of penalized regression models that will allow us to estimate sparse multivariate autoregressive (SMAR) models. In the following we

shall limit our presentation to first order SMAR models in which  $Nk=1$ . This will simplify the description of models and methods, allowing us to concentrate on conceptual issues. Previous studies (Martinez-Montes *et al.* 2004; Valdés-Sosa 2004) have shown that first order MAR models fit fMRI data well (as indicated by the model selection criteria such as GCV, AIC or BIC). However, it is clear that for other types of data such as EEG, more complex models are necessary. All expressions given below generalize to the more complete model. In fact, all software developed to implement the methods described has been designed to accommodate all model orders.

We first review classical MAR methods. For a first order MAR equation (1.1) simplifies to:

$$\mathbf{y}_t = \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{e}_t \quad t = 2, \dots, Nt \quad (2.1)$$

where  $\mathbf{e}_t$  is assumed to follow a multivariate Gaussian distribution  $N(\mathbf{0}, \Sigma)$ , with zero mean  $\mathbf{0}_{(p \times 1)}$  and precision matrix  $\Sigma_{(p \times p)}^{-1}$ .

This model can be recast as a multivariate regression:

$$\mathbf{Z} = \mathbf{X}\mathbf{B} + \mathbf{E} \quad \mathbf{E}_i \sim N(\mathbf{0}, \Sigma) \quad i = 1, \dots, m \quad (2.2)$$

where we define  $m = Nt - 1$  and introduce the notation:

$$\mathbf{Z}_{(m \times p)} = [\mathbf{y}_2, \dots, \mathbf{y}_t, \dots, \mathbf{y}_{Nt}]^T = [\mathbf{z}_1, \dots, \mathbf{z}_i, \dots, \mathbf{z}_p],$$

$$\mathbf{B}_{(p \times p)} = \mathbf{A}_1^T = [\beta_1, \dots, \beta_p],$$

$$\mathbf{X}_{(m \times p)} = [\mathbf{y}_1 \cdots \mathbf{y}_m]^T,$$

$$\mathbf{E}_{(m \times p)} = [\mathbf{e}_2, \dots, \mathbf{e}_t, \dots, \mathbf{e}_{Nt}]^T.$$

Usual time-series methods rely on maximum likelihood (ML) estimation of model (2.2), which is equivalent to finding:

$$\hat{\mathbf{B}} = \arg \min_{\mathbf{B}} \|(\mathbf{Z} - \mathbf{X}\mathbf{B})\|_{\Sigma}^2. \quad (2.3)$$

This has an explicit solution, the OLS estimator:

$$\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Z}. \quad (2.4)$$

It should be noted that the unrestricted ML estimator of the regression coefficients does not depend on the spatial covariance matrix of the innovations (Hamilton 1994). One can therefore carry out separate regression analyses for each node. In other words, it is possible to estimate separately each column  $\beta_i$  of  $\mathbf{B}$ :

$$\hat{\beta}_i = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{z}_i \quad i = 1, \dots, p, \quad (2.5)$$

where  $\mathbf{z}_i$  is the  $i$ -th column of  $\mathbf{Z}$ . It is to be emphasized that these definitions will work only if  $m \gg p$ . Additionally, it is also well known that OLS does not ensure sparse connectivity patterns for  $\mathbf{A}_1$ . We must therefore turn to regression methods specifically designed to ensure sparsity.

The first solution that comes to mind is to use the readily available stepwise variable selection methods. Such is the philosophy of TETRAD (Glymour *et al.* 1988; Spirtes *et al.* 1990). Unfortunately, stepwise methods

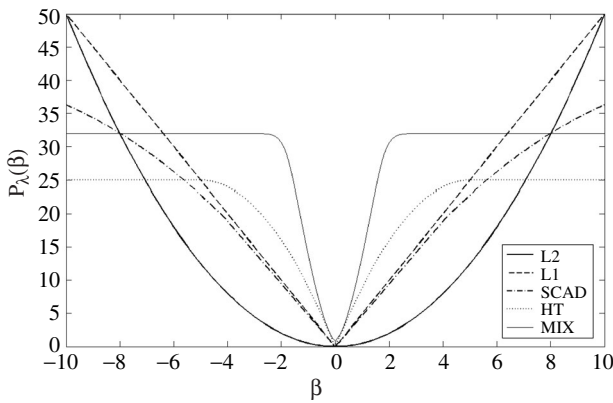


Figure 2. Penalization functions used for the iterative estimation of sparse causal relations. At each step of the iterative process, the regression coefficients of each node with all others are weighted according to their current size. Many coefficients are successively down-weighted and ultimately set to zero—effectively carrying out variable selection. *y*-Axis: weight according to current value of a regression coefficient  $\beta$  (*x*-axis). Each curve corresponds to a different type of penalization: heavy line, L2 norm (ridge regression); dashed, L1 norm (LASSO). Dotted, Hard-Threshold; dash-dot, SCAD; light line, mixture.

are not consistent (Hastie *et al.* 2001). This means that even increasing the sample size indefinitely ( $Nt \rightarrow \infty$ ) does not guarantee the selection of the correct set of non-zero coefficients. This result still holds even if all subsets of variables are exhaustively explored.

Procedures with better performance are those based on Bayesian methods in which assumptions about  $\mathbf{B}$  are combined with the likelihood by means of Bayes' theorem. A very popular method is stochastic search variable selection (SSVS) (George & McCulloch 1997; George 2000). SSVS is based on a hierarchical model in which the first stage is just the likelihood defined by equation (2.1), and the other stage assumes that the elements of  $\mathbf{B}$  ( $\beta$ ) are each sampled *a priori* from a mixture of two probability densities:  $p_0 f_{p_0}(\beta) + (1 - p_0) f_{p_1}(\beta)$ . The density  $f_{p_0}(\beta)$  is concentrated around zero, while  $f_{p_1}(\beta)$  has a larger variance. The decision of sampling from either is taken with binomial probabilities  $p_0$  and  $(1 - p_0)$ , respectively. When  $p_0$  is large, this means we expect the matrix  $\mathbf{B}$  to be very sparse. The model is explored using Monte Carlo–Markov chain techniques. This limits the application of this method to a rather small number of nodes  $p$  as analysed in Dobra *et al.* (2004), Dobra & West (2005) and Jones *et al.* (2005).

For this reason, we chose to explore other methods as alternatives to SSVS for variable selection, giving preference to those that were computationally more feasible. There has been much recent attention on different forms of penalized regression models. The simplest and best known of this family of methods is ridge regression (Hoerl & Kennard 1970), also known as quadratic regularization, which substitutes the argument (2.3) for the following one:

$$\hat{\mathbf{B}} = \arg \min_{\mathbf{B}} \|(\mathbf{Z} - \mathbf{X}\mathbf{B})\|_{\Sigma}^2 + \lambda^2 \|(\mathbf{P}\mathbf{B})\|^2. \quad (2.6)$$

Minimization of this functional leads to the estimator:

Table 1. Derivatives of penalty functions.

type of penalization	derivative
LASSO	$p'_{\lambda}(\theta) = \lambda \text{sign}(\theta)$
SCAD	$p'_{\lambda}(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(\alpha\lambda - \theta)_+}{(\alpha - 1)\lambda} I(\theta > \lambda) \right\}$
Hard-Threshold	$p'_{\lambda}( \theta ) = -2( \theta  - \lambda)_+$
ridge	$p'_{\lambda}(\theta) = 2\lambda\theta$
MIX	$p'_{\lambda}(\theta) = -\lambda \left[ \frac{p_0 f'_{p_0}(\theta) + p_1 f'_{p_1}(\theta)}{p_0 f_{p_0}(\theta) + p_1 f_{p_1}(\theta)} \right]$ where $f_p(\theta) = \frac{p^{1-(1/p)}}{2\sigma_p \Gamma(\frac{1}{p})} \exp\left(-\frac{1}{p} \frac{ x-x_0 ^p}{\sigma_p}\right)$ $\Gamma(\cdot)$ denotes the Gamma function

$$\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X} + \lambda^2 \mathbf{P}^T \mathbf{P})^{-1} \mathbf{X}^T \mathbf{Z}, \quad (2.7)$$

$\lambda$  being the regularization parameter which determines the amount of penalization enforced. There are very efficient algorithms based on the singular value decomposition for calculating these estimators as well as their standard errors. Forms of ridge regression have been recently applied (with  $\mathbf{P} = \mathbf{I}_p$ ) to analyse microarray data by West (2002) and (with  $\mathbf{P}$  a spatial Laplacian operator) to study fMRI time-series by Valdés-Sosa (2004). These papers showed the ability of this method to achieve stable and plausible estimates in the situation  $p \gg n$ . In the present paper, we explore the feasibility of using ridge regression as part of a technique for variable selection. It should be clear that ridge regression does not carry out variable selection *per se*. For this reason it is necessary to supplement this procedure with a method for deciding which coefficients of  $\hat{\mathbf{B}}$  are actually zero. This will be described in detail below.

Following ridge regression, a number of penalized regression techniques have been introduced in order to stabilize regressions and perform variable selection. All these methods can be expressed as the solution of the minimization of:

$$\hat{\beta} = \arg \min_{\beta} \|\mathbf{Z} - \mathbf{X}\beta\|^2 + \lambda^2 \sum_{j=1}^d p(|\beta_j|). \quad (2.8)$$

where  $p(|\beta_j|)$  is the penalty function applied to each component of the vector of regression coefficients  $\beta$ . The form of different penalty functions as a function of the current value of a regression coefficient  $\beta$  is shown in figure 2. It should be noted that the quadratic function is the ridge regression described above. Another type of penalty, perhaps one of the best known in the statistical learning literature, is the LASSO (Hastie *et al.* 2001), or L1 norm. This method has been recently implemented with great computational efficiency (Efron *et al.* 2004).

During the process of implementing algorithms for each type of penalty function, advantage was taken of the recent demonstration by Fan & Li (2001), Fan & Peng (2004), Hunter (2004) and Hunter & Lange (2004) that estimation of any one of many penalized regressions can be carried out by iterative application of ridge regression:

$$\hat{\beta}_i^{k+1} = (\mathbf{X}^T \mathbf{X} + \lambda^2 \mathbf{D}(\hat{\beta}_i^k)) \mathbf{X}^T \mathbf{z}_i \quad i = 1, \dots, p. \quad (2.9)$$

where  $\mathbf{D}(\hat{\beta}_i^k)$ , a diagonal matrix is defined by  $\mathbf{D}(\theta) =$

$\text{diag}(p'_\lambda(\theta_k)/|\theta_k|)$   $k = 1, \dots, p$  and  $p'_\lambda(\theta)$  is the derivative of the penalty function being evaluated.

The algorithm described by Fan & Peng (2004) unifies a large number of penalized regression techniques. These are summarized in table 1, in which the derivatives of the penalty functions are provided.

The reason that this algorithm works may be inferred from figure 2. At each step of the iterative process, the regression coefficients of each node with all others are weighted according to their current size. Many coefficients are successively down-weighted and ultimately set to zero—effectively carrying out variable selection in the case of the LASSO, Hard-Threshold and SCAD penalization. It must be emphasized that the number of variables set to zero in any of the methods described will depend on the value of the regularization parameter  $\lambda$  with higher values selecting fewer variables. In this paper, the value of the tuning parameter  $\lambda$  was selected to minimize the generalized crossvalidation criterion (GCV).

The penalizations explored in this article for variable selection are:

- (i) ridge: the L2 norm;
- (ii) LASSO: the L1 norm;
- (iii) Hard-Thresholding;
- (iv) SCAD: smoothly clipped absolute deviation penalty of Fan & Li (2001); and
- (v) MIX: mixture penalty.

It came as a pleasant surprise to us during the programming of the variable selection algorithms, that the SSVS of George & McCulloch (1997) can also be expressed as a penalized regression with penalty  $-\ln(p_o f_{p_o}(\beta) + (1 - p_o) f_{p_1}(\beta))$ . We therefore added to the comparisons this ‘quick and dirty’ implementation of SSVS as the MIX criteria which also carries out automatic variable selection.

The specific implementation of penalized regression used in this article is that of the maximization–minorization (MM) algorithm (Hunter 2004; Hunter & Lange 2004), which exploits an optimization technique that extends the central idea of EM algorithms to situations not necessarily involving missing data, nor even ML estimation. This new algorithm retains virtues of the Newton–Raphson algorithm. All algorithms were implemented in MATLAB 7.0 and will be made available on the website of this journal (see Electronic Appendix).

Additionally, the iterative estimation algorithm allows us to compute the covariance matrix of the resulting regression coefficient via a ‘sandwich formula’. This allows the estimation of standard errors for different contrasts of interest. For example, these standard errors were used to define a  $t$  statistic for each autoregressive coefficient to test its presence, or to calculate confidence intervals for different contrasts.

### 3. PERFORMANCE OF PENALIZED REGRESSION METHODS WITH SIMULATED DATA

#### (a) Description of simulations

In order to measure the performance of different penalized regression methods for estimating SMAR

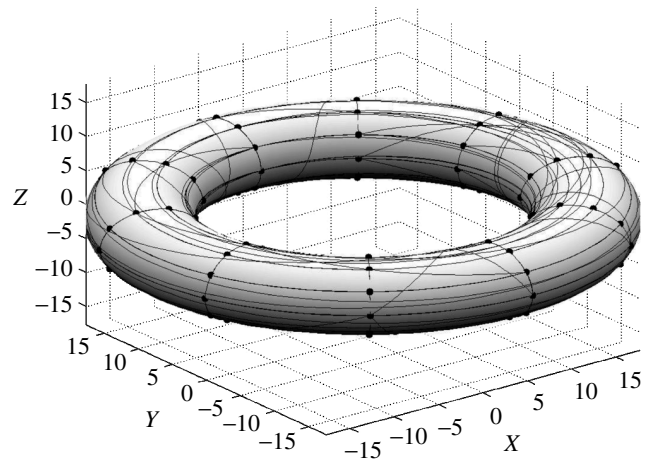


Figure 3. Idealized cortical models used to test regression methods for the identification of sparse graphs were simulated by a ‘small world’ network topology. Nodes resided on a two-dimensional grid on the surface of a torus, thus imposing periodic boundary conditions in the plane. For each simulation, a set of directed connections was first formed with a distribution crafted to induce the ‘small world effect’. The strengths of the connections between parents and children were sampled from a Gaussian distribution. Directed links are shown on the surface of the torus for one sample network.

models, a number of simulations were carried out. For this purpose, a universe of idealized cortical models was defined based on the concept of ‘small world topology’ (Watts & Strogatz 1998; Albert & Barabasi 2002; Jirsa 2004; Sporns *et al.* 2004; Sporns & Zwi 2004; Sporns 2005).

The simulated ‘cortex’ was defined as a set of nodes comprising a two-dimensional grid on the surface of a torus (figure 3). This geometry was chosen to avoid special boundary conditions since the network is periodic in the plane in both dimensions. For each simulation a set of directed connections was formed randomly. Following Sporns & Zwi (2004), the existence of a directed connection between any nodes  $i$  and  $j$  was sampled from a binomial distribution with probabilities  $p_{ij}$ . These probabilities were in turn sampled from a mixture density:

$$p_{ij} = \pi_{ij} \exp\left(\frac{r_{ij}^2}{\alpha^2}\right) + (1 - \pi_{ij})\gamma.$$

The Gaussian component of the mixture (depending on distance) will produce short-range connections and induce high clustering among nodes. The uniform component of the mixture ensures the presence of long-range connections which induce short-path lengths between any pair of nodes in the network. The parameters of the mixture ( $\alpha$ ,  $\gamma$ ) were tuned by hand to produce a ‘small world’ effect, which was in practice, possible with only a small proportion of uniformly distributed connections. The directed links for one sample network are shown on the surface of the torus in figure 3.

A more detailed view of a sample small-world network is shown in figure 4 which shows in (a) the two-dimensional view of the links between nodes and in (b), their connectivity matrix. Once the connectivity matrix of the network was defined, the strengths of the connections between parents

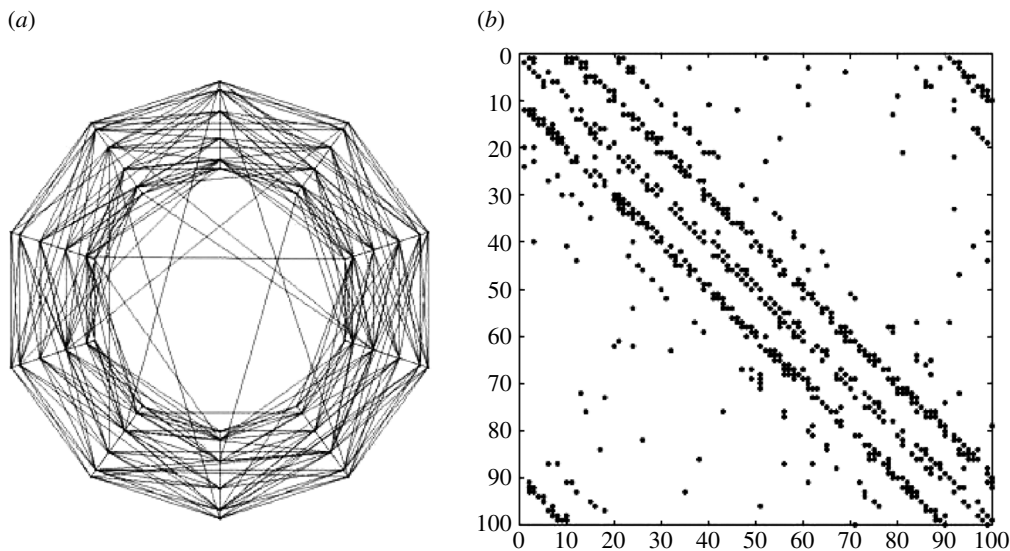


Figure 4. Connectivity structure of the simulated cortical network shown in figure 3. This type of small-world network has a high probability of connections between geographical neighbours and a small proportion of larger range connections. The network mean connectivity was: 6.23; the scaled clustering: 0.87; the scaled length: 0.19. (a) Two-dimensional view of the links between nodes. (b) Connectivity (0–1) matrix in with a row for each node and non-zero elements for its children.

and children were sampled from a Gaussian distribution truncated around zero with a variable threshold  $\tau$ . With higher  $\tau$ , only stronger connections were allowed, thus increasing the ‘signal to noise ratio’ for the detection of network connections. The resulting matrix of (auto)-regressive coefficients  $\mathbf{A}_1$  of the network has the same sparsity structure as that of the connectivity matrix. Those  $\mathbf{A}_1$  with singular values greater than one were rejected from the simulation, since our purpose was to study stable SMAR models.

Simulated fMRI time-series were generated by the first order SMAR model (2.1) with the connectivity matrix obtained as described above. A random starting state was selected, and then a ‘burning in’ period of several thousand samples was first generated and discarded to avoid system transients. Subsequent samples were retained for the analyses presented below. The result of this process, a typical fMRI simulation is shown in figure 5.

Simulations with different types of innovations  $\mathbf{e}_t$  were carried out. They differed in the type of inverse covariance matrices from which they were generated. Three variants of connectivity patterns for the spatial covariance  $\Sigma$  of the innovations were used to simulate fMRI time-series. Shown in figure 6 are the connectivity matrices for the precisions  $\Sigma^{-1}$  (a) spatial independence with a diagonal precision matrix, (b) nearest-neighbour dependency with partial autocorrelations existing only between nodes close to each other, (c) nearest-neighbour topology with an additional ‘master’ node linked to all other nodes in the network.

### (b) Comparison of methods

It must be remembered that the purpose of the simulations was to generate time-series from which the network topology of the idealized cortical network could be estimated. As is usual in the evaluation of diagnostic methods, a number of indices were calculated to evaluate the performance of different penalized

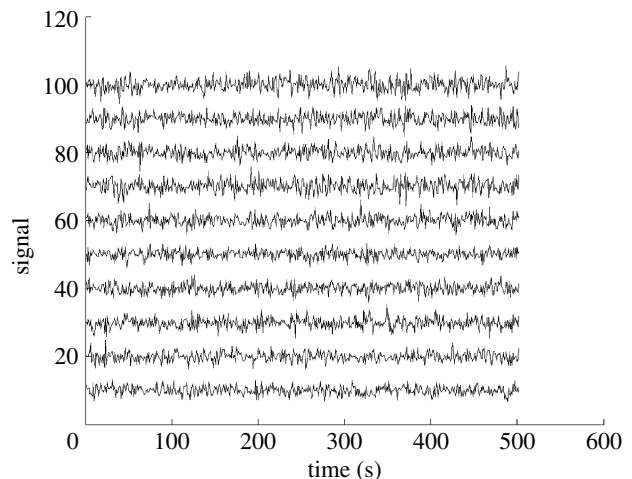


Figure 5. Simulated fMRI time-series generated by a first order multivariate autoregressive model  $\mathbf{y}_t = \mathbf{A}_1 \mathbf{y}_{t-1} + \mathbf{e}_t$ , the autoregressive matrix being sampled as described in figures 3 and 4. The innovations  $\mathbf{e}_t$  (noise input) were sampled from a Gaussian distribution with a prescribed inverse covariance matrix  $\Sigma^{-1}$  as described in figure 6. Y-axis: simulated BOLD signal, x-axis: time. The effect of different observed lengths of time-series ( $N$ ) on the detection of connections was studied.

regression techniques. For reference purposes, the definition of these indices is summarized in table 2.

The actual sensitivity and specificity of each regression method depends, of course, on the threshold selected to reject the null hypothesis for the  $t$  statistic of each regression coefficient. Overall performance for each regression method under different conditions was measured by means of their receiver operating characteristic (ROC) curves which are, as is well known, the representation of the tradeoffs between sensitivity (Sn) and specificity (Sp) (table 2). The plot shows false alarm rate ( $1 - \text{Sp}$ ) on the x-axis and detection rate (Sn) on the y-axis. ROC curves are further summarized by their areas, which we shall call for brevity the ‘detection efficiency’. In all comparisons, at least 25 simulated

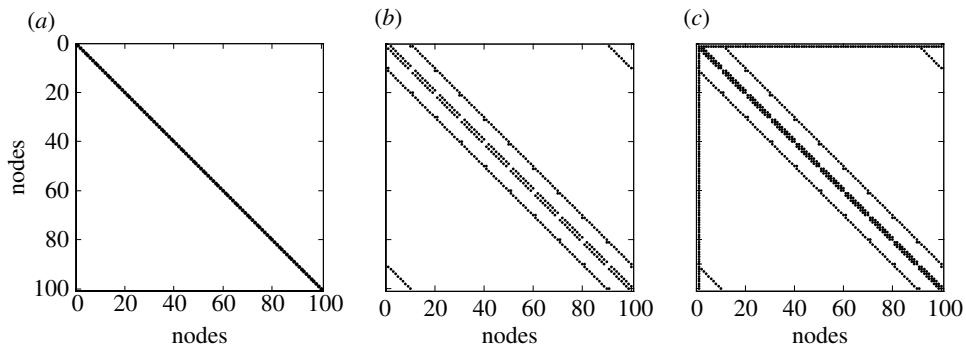


Figure 6. Connectivity matrices for the precisions  $\Sigma^{-1}$ . Three situations were explored: (a) spatial independence with a diagonal precision matrix, (b) nearest-neighbour dependency with partial autocorrelations existing only between nodes close to each other, (c) nearest-neighbour topology with a ‘master’ node linked to all other nodes in the network.

fMRI series were generated. For each comparison, each method was represented by its worst case scenario, the ROC curve with the lowest detection efficiency for all 25 replications. A typical example of ROC curves is shown in figure 7, which corresponds to ridge regression applied to a simulated network with  $p=100$  nodes and a recorded length of  $Nt=200$  time points. The dark line corresponds to a simulated fMRI generated with spatially independent noise, as well as with a high signal to noise ratio. The ROC curve is well above the diagonal line that would be the result with a random detection procedure.

From the whole set of simulations a number of findings can be summarized.

In the first place, the detection efficiency in all simulations was well above the chance level, validating the hypothesis that penalized regression techniques are useful for the detection of connectivity topologies in complex networks. The difference between penalization techniques was rather disappointing, as summarized in figure 8 which shows that all methods are roughly equivalent with respect to detection efficiency. Exceptions are the hard threshold penalty which performs slightly worse than the others and ridge regression that performs slightly better. In view of the ease with which ridge regression is computed, there seems to be no point in using more complicated techniques. For this reason, from now onwards, unless explicitly stated, all results presented and discussed correspond to ridge regression.

With regard to the  $p/Nt$  ratio, figure 8 shows the detection efficiency as a function of  $Nt$  for a fixed number of nodes ( $p=100$ ). All methods perform equally well when the number of nodes is small with regard to the number of time points. Efficiencies decrease uniformly when the number of data points decreases but are well above chance levels even for  $p=4Nt$ .

Detection efficiency depends monotonically on the  $S/N$  ratio connection strength. Figure 9 shows that even with networks with small connection strengths relative to the system noise, good detection efficiencies are possible (LASSO penalization).

Strong spatial correlations in the innovations tended to diminish the detection efficiency for  $\mathcal{A}_1$  with respect to the uncorrelated case. The worse performance is with innovations generated from precision matrices with strong structure and a master driving node. The

Table 2. Definition of quantities used for assessing the methods network reconstruction.

quantity	definition
number of true edges	TP + FN
number of zero-edges	TN + FP
significant edges	TP + FP
detection rate	TP/(TP + FN)
false alarm rate	FP/(TN + FP)

thin line in figure 7 corresponds to a time-series generated with both spatially correlated innovations (nearest-neighbour topology), as well as with a low signal to noise ratio. Note the interaction of both factors that produce marked decreases of detection efficiency when compared with the situation denoted by the thick line (high  $S/N$  and no spatial correlation).

For the real fMRI experiments, we must select a threshold for rejecting the null hypothesis. This involves multiple comparisons for a large number of autoregressive coefficients. The simulations gave us the opportunity of checking the usefulness for this purpose of the FDR procedure introduced by Benjamini & Hochberg (1995). Given a set of  $p$  hypotheses, out of which an unknown number  $p_0$  are true, the FDR method identifies the hypotheses to be rejected, while keeping the expected value of the ratio of the number of false rejections to the total number of rejections below  $q$ , a user-specified control value. In the present paper we use a modification of this procedure, the ‘local’ FDR (which we shall denote as ‘fdr’ in lower case) as developed by Efron (2003, 2004, 2005). Multiple tests are modelled as being sampled from the mixture of two densities given by  $f_{(z)} = p_0 f_{0(z)} + p_1 f_{1(z)}$ , which are estimated with non-parametric methods. An R program `LOCfDR` is available from the CRAN website for this calculation. The fdr procedure was used to analyse the same data used to generate figure 7. Figure 10 shows the results of applying `locfdr` which estimates the  $t$  statistics for all regression coefficients as the mixture of two of the null and alternative densities. Figure 11 shows the fdr curve produced which allows the selection of a threshold with a given local false-positive rate. Looking back to figure 7, the dashed line shows the performance of the local fdr thresholds calculated *without* knowledge of the true

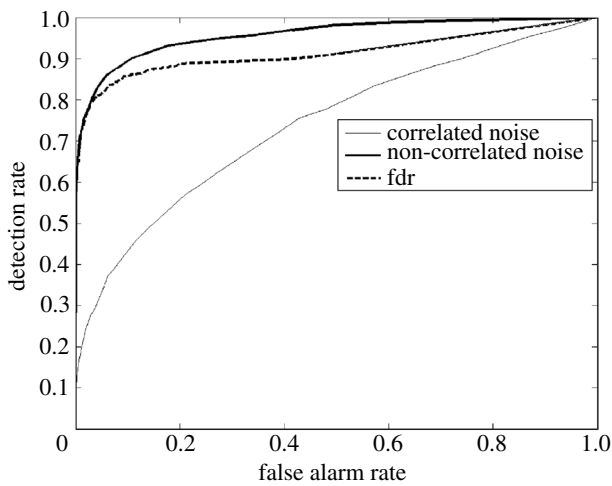


Figure 7. Efficiency of ridge regression for the detection of causal connections in simulated fMRI from a network with  $p=100$  nodes and a recorded length of  $Nt=200$  time points, as measured by receiver operating curves (ROC).  $y$ -Axis: probability of detection of true connections,  $x$ -axis: probability of false detections. The dark line corresponds to an fMRI generated with spatially independent noise as well as with a high signal to noise ratio. The thin line corresponds to a time-series generated with spatially correlated noise (nearest neighbour), as well as with a low signal to noise ratio. Note the decreases of detection efficiency with these factors. The dashed line shows the performance of the local false discovery rate thresholds calculated *without* knowledge of the true topology of the network. Note the excellent correspondence at low false-positive rates.

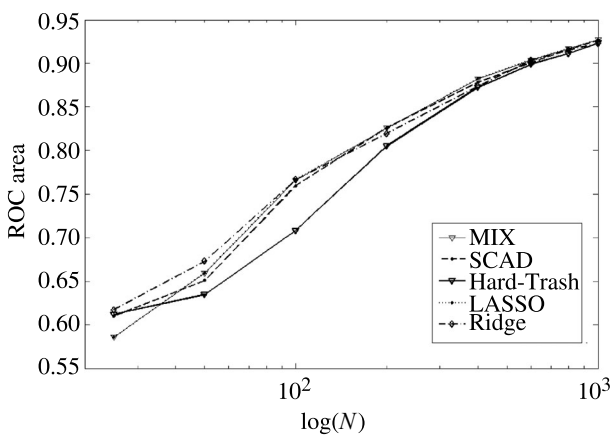


Figure 8. Effect of the ratio of network size ( $p$ ) to temporal sample size ( $Nt$ ) on the detection efficiency for different penalized regression methods. The number of nodes in the network was kept at  $p=100$ .  $y$ -Axis: area under ROC curve.  $x$ -Axis: sample size ( $N$ ). Though efficiency decreases with smaller sample sizes, all methods perform well above chance even for  $p=4N$ . Ridge regression dominates the other methods for  $p=N$  with no significant differences at other  $p/Nt$  ratios

topology of the network. Note the excellent correspondence between the fdr and the ROC curve at low false-positive rates.

#### 4. ANALYSIS OF FMRI DATA

A combination of ridge regression and local FDR was used to analyse fMRI data recorded during a face processing experiment. No attempt was made to reach

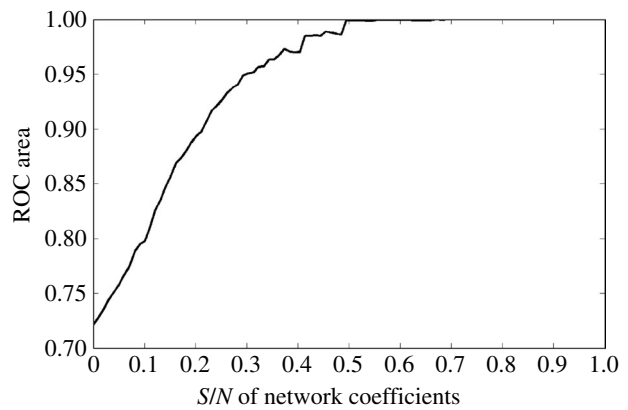


Figure 9. Effect of signal to noise ratio of network connectivity generation on efficiency of detection by LASSO.  $y$ -Axis: area under the ROC,  $x$ -axis: signal to noise ratio.

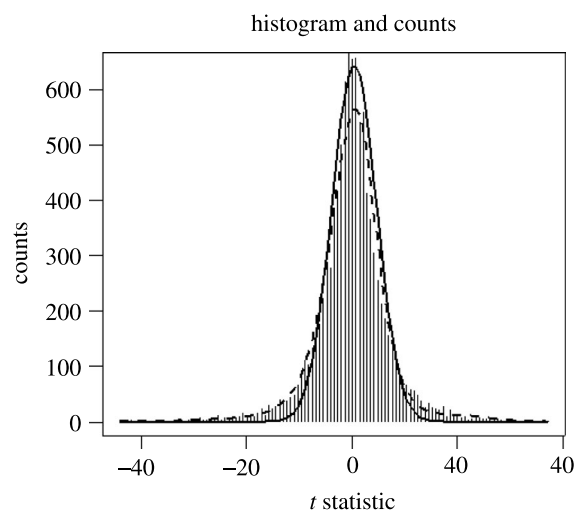


Figure 10. The local FDR (fdr) is ideal for the detection of sparse connections. If there are few connections, then testing for links between all nodes should lead to a sample of test statistics for which the null hypothesis predominates. The distribution of the statistics can therefore be modelled as a mixture of the density of null hypothesis with that of the alternative hypothesis. These are separated by non-parametric density estimation as shown in this figure, in which the thick line denotes the estimated null distribution and the thin one the estimated alternative distribution for the ridge regression example shown in figure 7 (thick line).  $y$ -Axis: counts,  $x$ -axis: values of the  $t$  statistics for estimated regression coefficients.

exhaustive substantive conclusions about the experiment analysed, since the purpose of this exercise was only to demonstrate the feasibility of working with the new methods. The experimental paradigm consisted of the presentation of faces of both men and women under the following conditions:

Condition 1: static faces with fearful expressions (SFF);

Condition 2: neutral faces (with no emotional content), (NF);

Condition 3: dynamic fear faces (in this condition faces are morphed from neutral emotional content to fear; DFF).

The subject was asked to count the number of faces that belonged to women. Stimuli were presented in a



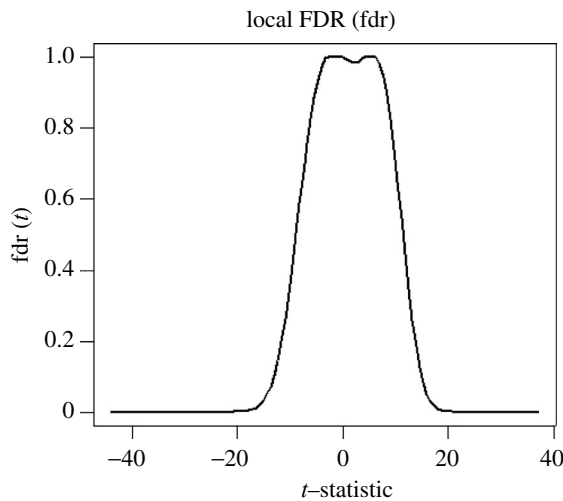


Figure 11. The local false discovery of the ridge regression example of figure 7. *y*-Axis: *fdr*, *x*-axis: *t* statistic for estimated regression coefficients.

block design with the following order: SFF—NF—DFF. Each block lasted 40 s and was repeated six times. The experiment duration was 720 s = 12 min. The duration of each stimulus was 1 s for each condition. Stimuli presentation and synchronization to the MR scanner was performed using COGENT modelling software v.2.3 (<http://cogent.psyc.bbk.ac.uk/>; figure 12).

Images were acquired using a 1.5 T Symphony Scanner, Siemens, Erlangen, Germany. Functional images were acquired using a T2\* weighted echo planar sequence in 25 oblique slices (interleave acquisition). The EPI sequence was defined by: TE = 60 ms, TR = 4000 ms, flip angle: 90°, FOV = 224 mm, slice thickness: 3.5 mm, acquisition matrix = 64 × 64. The number of scans recorded was 185. The first five scans were rejected for the analysis because of T1 saturation effect. A high resolution anatomical image acquisition was also acquired using a T1 MPRAGE sequence (TE = 3.93 ms/TR = 3000 ms), voxel size = 1 × 1 × 1 mm<sup>3</sup>, FOV = 256 mm. Matrix size = 256 × 256.

The fMRI data were first analysed using the STATISTICAL PARAMETRIC Mapping Software package SPM2 ([www.fil.ion.ucl.ac.uk/spm/software/spm2/](http://www.fil.ion.ucl.ac.uk/spm/software/spm2/)). Preprocessing with SPM was restricted to the following steps: (i) slice time correction (using trilinear interpolation); (ii) motion correction; (iii) unwarping. No temporal smoothing was used. As a preliminary check, using standard SPM procedures for the comparison of conditions it was possible to show activation of fusiform face area (FFA) as well as involvement of limbic structures to the presentation of fearful faces.

Inspection of the fMRI time-series for all fMRI voxels revealed a rhythmic artefact, synchronous for all voxels that was eliminated by suppression of the first pair of singular vectors in the SVD decomposition of the raw data matrix. In order to reduce the spatial dimensions of the data, the subject's MRI time was segmented into 116 different structures using an automated procedure and based on the macroscopic anatomical parcellation of the MNI MRI single-subject brain used by Tzourio-Mazoyer *et al.* (2002).

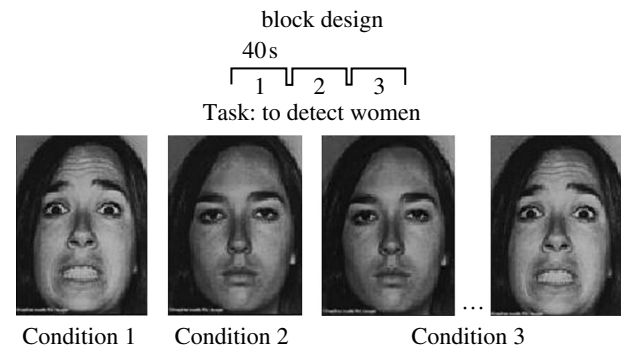


Figure 12. fMRI acquisition: the experimental paradigm consisted of visual stimuli presented under three conditions. Condition 1, static fearful faces, (SFF); Condition 2, neutral faces (with no emotional content), (NF); Condition 3, dynamic fearful faces (in this condition faces are morphed from neutral emotional content to fear; DFF). A general linear model was posited that included not only a different mean level  $\mu_C$  vector, but also a different autoregressive matrix  $A_1^C$  for each condition *C*. Thus, the model explores changes across voxels not only of mean level of activity but also of connectivity patterns.

The fMRI time-series data were spatially averaged over these ROI to yield 116 time-series.

For the analysis of these data, model (2.1) was expanded to:

$$y_t = d_t + \mu_t^C + A_1^C y_{t-k} + e_t \quad t = 2, \dots, N, \quad (4.1)$$

where  $d_t$  is a drift term estimated by a second-order polynomial defined over the whole experiment,  $\mu_C$  is the mean level for conditions and  $A_1^C$  the condition-dependent autoregressive matrices. Thus, the model explores changes across voxels, not only of mean level of activity, but also of connectivity patterns. We decided to compare conditions SFF and DFF (fearful faces). The model was fitted by means of ridge regression (with no regularization on the drift and condition mean effects). *t* Statistics were computed for the relevant contrasts.

Figure 13 shows the tomography of the *t* statistics contrasting the average of the fearful face means ( $(\mu_{SFF} + \mu_{DFF})/2$ ) with that of neutral faces  $\mu_{NF}$ . The map is thresholded using the local FDR (*fdr*) as explained above with  $q = 0.01$ . Note the activation of the FFA area which was very similar to that obtained with the analysis carried out with SPM2.

A similar analysis was carried out with the connectivity matrices (figure 14). The contrast compared the pooled estimate of fearful faces  $(A_1^{SFF} + A_1^{DFF})/2$  to that of neutral faces ( $A_1^{NF}$ ). Graphs are constructed with only those edges which fell above the *fdr* threshold for the *t* statistics of the contrast. Both (a) and (b) of figure 14 show the same data with a more schematic and a more realistic rendering, respectively. It is interesting to note the involvement of brain structures involved in processing emotional stimuli. Absent are connections to FFA which have approximately the same level in all face conditions.

## 5. DISCUSSION

This paper proposes a method for identifying large scale functional connectivity patterns from relatively short

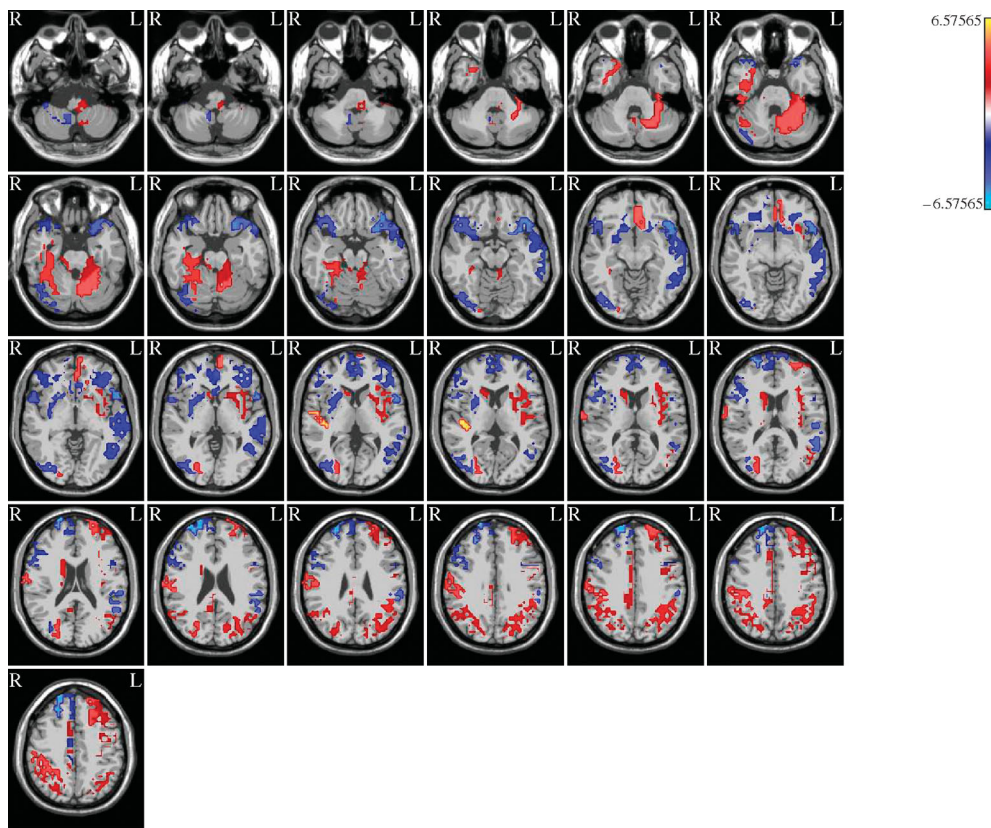


Figure 13. Tomography of  $t$  statistics contrasting fearful face means  $(\mu_{\text{SFF}} + \mu_{\text{DFF}})/2$  with that of neutral faces  $\mu_{\text{NF}}$ .  $t$ -Values are obtained by Bayesian ridge regression and thresholded using the local FDR (fdr) as explained in figures 10 and 11. Note the activation of the FFA which was very similar to that obtained with the SPM package.

time-series of functional neuroimages. The method is based on estimating SMAR models by a two-stage process that first applies penalized regression (Fan & Peng 2004), and is then supplemented by pruning of unlikely connections by use of the local FDR procedure developed by Efron (2003). The methods are demonstrated to perform well in identifying complex patterns of network connectivity by means of simulations on an idealized small world cortical network. These simulations also show that the simplest of the methods, ridge regression, performs as well as more sophisticated and recent techniques. This does not rule out that the performance of other penalized techniques might be improved, for example, by a better estimate of the regularization parameter, just to mention one possibility. Of particular interest is the complete exploration, not carried out in the present project owing to time constraints, of the mixture penalties that provide a bridge between SSVS (George & McCulloch 1997) and penalized regression techniques.

The simulations also highlight an important area for improvement. The detection efficiency of penalized regression decreases with unobserved correlations between the inputs of the system which in graphical models correspond to unobserved latent variables. This is in agreement with theoretical insights provided by statistical analyses of causality (Pearl 1998), as well as being part of the accumulated experience of time-series analysis in the neurosciences (Kaminski *et al.* 2001). Part of the problem is the relative unreliability of estimating very large dimensional covariance matrices. Inspection of

Table 3. Effect on detection efficiency of different spatial correlation patterns of the innovations for a network with  $p=100$  and  $N_t=60$ .

(The two columns correspond to the detection efficiencies for estimates that do not take into consideration  $\Sigma^{-1}$  and those that do.)

$\Sigma^{-1}$	detection efficiency for $A_1$ estimated alone	detection efficiency for $A_1$ estimated with information about $\Sigma^{-1}$
diagonal	0.8001	0.8012
nearest neighbour	0.7873	0.7880
nearest neighbour with master node	0.6747	0.6298

table 3 shows that estimation and use of the covariance matrix of the innovations does not improve the detection efficiency for autoregressive coefficients.

The assumption of sparsity of neural connections has been supported by quantitative studies of databases of neural connections (Hilgetag *et al.* 2002). Sparseness is a central concept of modern statistical learning (Gribonval *et al.* 2005), but had not been applied, to our knowledge, to the estimation of MAR models. This general requirement for sparsity may be combined in the future with the information provided by fibre tractography methods based on diffusion MRI.

The simulations presented and the real fMRI example analysed comprised 100 and 116 time-series,

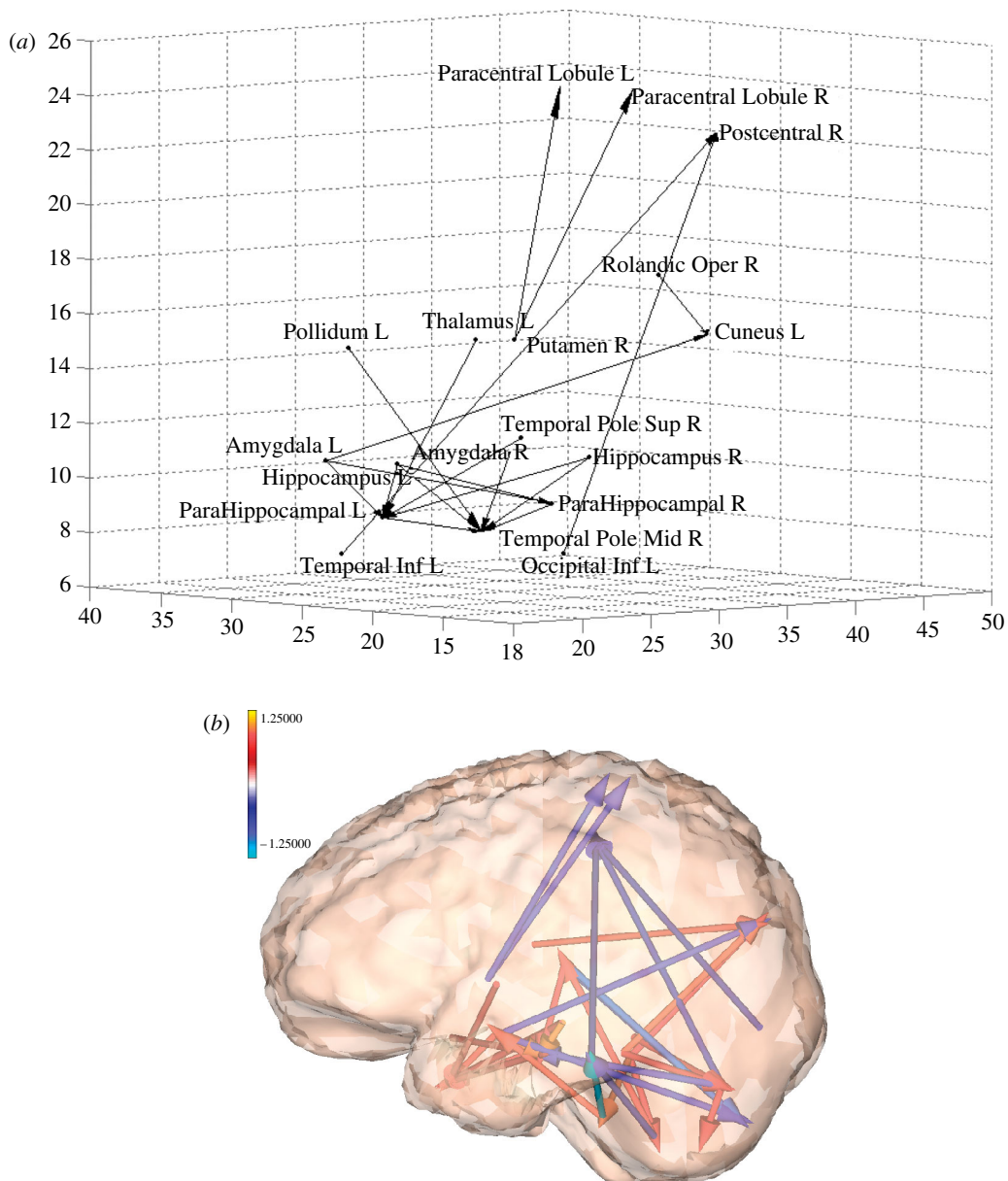


Figure 14. (a) Graph of connections that change with appearance of fearful expression. Obtained by element wise comparison of the autoregressive matrices of fearful faces ( $\mathcal{A}_1^{\text{SFF}} + \mathcal{A}_1^{\text{DFE}}/2$ ) as compared with that of neutral faces ( $\mathcal{A}_1^{\text{NF}}$ ). Only those connections above the *fd*r threshold are shown. Note involvement of areas related to emotional responses. (b) Three-dimensional rendering of the connectivity patterns shown in (a).

respectively. Although falling short of the spatial dimensionality of functional neuroimages, they represent an order of magnitude increase in the size of problem than those that are solvable standard time-series techniques. The methods and software developed have been tested to be scalable for the analysis of hundreds of thousands of voxels.

For the sake of simplicity, the SMAR has been posited to be linear, stationary and to involve only lags of the first order. It is relatively straightforward to generalize this formalism to the analysis of more complex situations. Such extensions have already been carried out for the small  $p$  case for non-stationary time-series analysis (Hesse *et al.* 2003) and for non-linear processes (Freiwald *et al.* 1999). Work is currently in progress to apply sparse restrictions in order to address more realistic assumptions when modelling functional neuroimages.

While it is true that nothing can substitute for the lack of data, the next best thing, if the data are scarce, is not to use it in estimating things that are probably not there.

The authors thank Mitchell Valdés-Sosa, Maria A. Bobes-León, Nelson Trujillo Barreto and Lorna García-Pentón for providing the experimental data analysed in this paper, as well as for valuable insights and support.

## REFERENCES

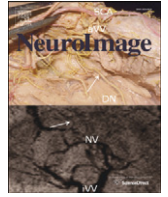
- Albert, R. & Barabasi, A. L. 2002 Statistical mechanics of complex networks. *Rev. Modern Phys.* **74**, 47–97.
- Baccala, L. A. & Sameshima, K. 2001 Partial directed coherence: a new concept in neural structure determination. *Biol. Cybern.* **84**, 463–474.
- Bach, F. R. & Jordan, M. I. 2004 Learning graphical models for stationary time series. *IEEE Trans. Signal Proc.* **52**, 2189–2199.

- Benjamini, Y. & Hochberg, Y. 1995 Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodological* **57**, 289–300.
- Brillinger, D. R., Bryant, H. L. & Segundo, J. P. 1976 Identification of synaptic interactions. *Biol. Cybern.* **22**, 213–228.
- Bullmore, E., Harrison, L., Lee, L., Mechelli, A. & Friston, K. 2004 Brain connectivity workshop, Cambridge UK, May 2003. *Neuroinformatics* **2**, 123–125.
- Cowell, R. G., Dawid, P. A., Lauritzen, S. L. & Spiegelhalter, D. J. 1999 *Probabilistic networks and expert systems*. New York: Springer.
- Dahlhaus, R. 1997 Fitting time series models to nonstationary processes. *Ann. Stat.* **25**, 1–37.
- Dahlhaus, R., Eichler, M. & Sandkuhler, J. 1997 Identification of synaptic connections in neural ensembles by graphical models. *J. Neurosci. Methods* **77**, 93–107.
- Demiralp, S. & Hoover, K. D. 2003 Searching for the causal structure of a vector autoregression. *Oxford Bull. Econ. Stat.* **65**, 745–767.
- Dobra, A. & West, M. 2005 HdBCS—Bayesian covariance selection. (See <http://ftp.isds.duke.edu/WorkingPapers/04-23.pdf>.)
- Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G. A. & West, M. 2004 Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90**, 196–212.
- Efron, B. 2003 Robbins, empirical Bayes and microarrays. *Ann. Stat.* **31**, 366–378.
- Efron, B. 2004 Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc.* **99**, 96–104.
- Efron, B. 2005 Bayesians, frequentists, and physicists. (See <http://www-stat.stanford.edu/~brad/papers/physics.pdf>.)
- Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. 2004 Least angle regression. *Ann. Stat.* **32**, 407–451.
- Eichler, M. 2004 *Causal inference with graphical time series models*. *Brain Connectivity Workshop Havana, April 26–29* p. 1
- Eichler, M. 2005 A graphical approach for evaluating effective connectivity in neural systems. *Phil. Trans. R. Soc. B* **360**, 953–967. (doi:10.1098/rstb.2005.1641.)
- Fan, J. Q. & Li, R. Z. 2001 Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 1348–1360.
- Fan, J. Q. & Peng, H. 2004 Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.* **32**, 928–961.
- Freiwald, W. A., Valdés, P., Bosch, J., Biscay, R., Jimenez, J. C., Rodriguez, L. M., Rodriguez, V., Kreiter, A. K. & Singer, W. 1999 Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. *J. Neurosci. Methods* **94**, 105–119.
- Friston, K. J. 1994 Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapp.* **2**, 56–78.
- Friston, K. J., Harrison, L. & Penny, W. 2003 Dynamic causal modeling. *Neuroimage* **19**, 1273–1302.
- George, E. I. 2000 The variable selection problem. *J. Am. Stat. Assoc.* **95**, 1304–1308.
- George, E. I. & McCulloch, R. E. 1997 Approaches for Bayesian variable selection. *Stat. Sinica* **7**, 339–373.
- Glymour, C., Scheines, R., Spirtes, P. & Kelly, K. 1988 Tetrad—discovering causal-structure. *Multivariate Behav. Res.* **23**, 279–280.
- Goebel, R., Roebroeck, A., Kim, D. S. & Formisano, E. 2003 Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. *Magn. Reson. Imaging* **21**, 1251–1261.
- Granger, C. W. J. 1969 Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 414.
- Gribonval, R., Figueras i Ventura, R. M. & Vandergheynst, P. 2005 A simple test to check the optimality of sparse signal approximations. (See [http://lts1pc19.epfl.ch/repository/Gribonval2005\\_1167.pdf](http://lts1pc19.epfl.ch/repository/Gribonval2005_1167.pdf).)
- Hamilton, J. D. 1994 *Time series analysis*. Princeton, NJ: Princeton University Press.
- Harrison, L., Penny, W. D. & Friston, K. 2003 Multivariate autoregressive modeling of fMRI time series. *NeuroImage* **19**, 1477–1491.
- Hastie, T. & Tibshirani, R. 2004 Efficient quadratic regularization for expression arrays. *Biostatistics* **5**, 329–340.
- Hastie, T., Tibshirani, R. & Friedman, J. 2001 *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.
- Hesse, W., Moller, E., Arnold, M. & Schack, B. 2003 The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies. *J. Neurosci. Methods* **124**, 27–44.
- Hilgetag, C., Kotter, R. & Stephan, K. E. 2002 Computational methods for the analysis of brain connectivity. In *Computational neuroanatomy* (ed. G. A. Ascoli). Totowa, NJ: Humana Press.
- Hoerl, A. E. & Kennard, W. R. 1970 Ridge regression—biased estimation for nonorthogonal problems. *Technometrics* **12**, 55.
- Horwitz, B. 2003 The elusive concept of brain connectivity. *NeuroImage* **19**, 466–470.
- Hunter, D. R. 2004 MM algorithms for generalized Bradley–Terry models. *Ann. Stat.* **32**, 384–406.
- Hunter, D. R. & Lange, K. 2004 A tutorial on MM algorithms. *Am. Stat.* **58**, 30–37.
- Jensen, F. R. 2002 *Bayesian networks and decision graphs*. New York: Springer.
- Jirsa, V. K. 2004 Connectivity and dynamics of neural information processing. *Neuroinformatics* **2**, 183–204.
- Jones, B., West, M. 2005 Covariance decomposition in undirected Gaussian graphical models. (See <http://ftp.isds.duke.edu/WorkingPapers/04-15.pdf>.)
- Jones, B., Carvalho, C., Dobra, A., Hans, Ch. 2005 Experiments in stochastic computation for high-dimensional graphical models. *Technical Report 2004-1(papers)*.
- Jordan, M. I. 2004 Graphical models. *Stat. Sci.* **19**, 140–155.
- Kaminski, M., Ding, M. Z., Truccolo, W. A. & Bressler, S. L. 2001 Evaluating causal relations in neural systems: granger causality, directed transfer function and statistical assessment of significance. *Biol. Cybern.* **85**, 145–157.
- Kotter, R. & Stephan, M. E. 2003 Network participation indices: characterizing component roles for information processing in neural networks. *Neural Netw.* **16**, 1261–1275.
- Lee, L., Harrison, L. M. & Mechelli, A. 2003 A report of the functional connectivity workshop, Dusseldorf 2002. *NeuroImage* **19**, 457–465.
- Leng, C., Lin, Y. & Whaba, G. 2004 A note on the Lasso and related procedures in model selection. (See <http://www.stat.wisc.edu/~wahba/ftp1/tr1091rxx.pdf>.)
- Mardia, K. V., Kent, J. T. & Bibby, J. M. 1979 *Multivariate analysis*. London: Academic Press.
- Martinez-Montes, E., Valdés-Sosa, P., Miwakeichi, F., Goldman, R. & Cohen, M. 2004 Concurrent EEG/fMRI analysis by multi-way partial least squares. *NeuroImage* **22**, 1023–1034.
- McIntosh, A. R. & Gonzalez-Lima, F. 1994 Structural equation modeling and its applications to network analysis in functional brain imaging. *Hum. Brain Mapp.* **2**, 2–22.

- Meinshausen, N. & Bühlmann, P. 2004 Consistent neighborhood selection for sparse high-dimensional graphs with the Lasso ([http://stat.ethz.ch/research/research\\_reports/2004/123](http://stat.ethz.ch/research/research_reports/2004/123).)
- Pearl, J. 1998 Graphs, causality, and structural equation models. *Sociol. Methods Res.* **27**, 226–284.
- Pearl, J. 2000 *Causality*. Cambridge: Cambridge University Press.
- Pearl, J. 2003 Statistics and causal inference: a review. *Test* **12**, 281–318.
- Salvador, R., Suckling, J., Schwarzbauer, C. & Bullmore, E. 2005 Undirected graphs of frequency dependent functional connectivity in whole brain networks. *Phil. Trans. R. Soc. B* **360**, 937–946. (doi:10.1098/rstb.2005.1645.)
- Scheines, R., Spirtes, P., Glymour, C., Meek, C. & Richardson, T. 1998 The TETRAD project: constraint based aids to causal model specification. *Multivariate Behav. Res.* **33**, 65–117.
- Speed, T. P. & Kiiveri, H. T. 1986 Gaussian markov distributions over finite graphs. *Ann. Stat.* **14**, 138–150.
- Spirtes, P., Scheines, R. & Glymour, C. 1990 Simulation studies of the reliability of computer-aided model-specification using the tetrad-ii, eqs, and lisrel Programs. *Sociol. Methods Res.* **19**, 3–66.
- Spirtes, P., Glymour, C. & Scheines, R. 1991 From probability to causality. *Phil. Stud.* **64**, 1–36.
- Spirtes, P., Richardson, T., Meek, C., Scheines, R. & Glymour, C. 1998 Using path diagrams as a structural equation modeling tool. *Sociol. Methods Res.* **27**, 182–225.
- Spirtes, P., Glymour, C. & Scheines, R. 2000 *Causation, prediction, and search*. Cambridge: The MIT Press.
- Sporns, O. 2005 Complex neural dynamics. (See [http://www.indiana.edu/~cortex/cd2002\\_draft.pdf](http://www.indiana.edu/~cortex/cd2002_draft.pdf).)
- Sporns, O. & Zwi, J. D. 2004 The small world of the cerebral cortex. *Neuroinformatics* **2**, 145–162.
- Sporns, O., Toning, G. & Edelman, G. M. 2000 Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cereb. Cortex* **10**, 127–141.
- Sporns, O., Chialvo, D. R., Kaiser, M. & Hilgetag, C. C. 2004 Organization, development and function of complex brain networks. *Trends Cogn. Sci.* **8**, 418–425.
- Stephan, K. E., Hilgetag, C. C., Burns, G. A. P.C., O'Neill, M. A., Young, M. P. & Kotter, R. 2000 Computational analysis of functional connectivity between areas of primate cerebral cortex. *Phil. Trans. R. Soc. B* **355**, 111–126. (doi:10.1098/rstb.2000.0552.)
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B. & Joliot, M. 2002 Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15**, 273–289.
- Valdes-Sosa, P. A. 2004 Spatio-temporal autoregressive models defined over brain manifolds. *Neuroinformatics* **2**, 239–250.
- Varela, F., Lachaux, J. P., Rodriguez, E. & Martinerie, J. 2001 The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* **2**, 229–239.
- Watts, D. J. 1998 S H Strogatz, Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442.
- Wermuth, N. & Cox, D. R. 1998 On association models defined over independence graphs. *Bernoulli* **4**, 477–495.
- Wermuth, N. & Cox, D. R. 2004 Joint response graphs and separation induced by triangular systems. *J. R. Stat. Soc. B Stat. Method.* **66**, 687–717.
- Wermuth & Lauritzen, S. L. 1990 On substantive research hypotheses, conditional-independence graphs and graphical chain models. *J. R. Stat. Soc. B Methodological.* **52**, 21–50.
- West, M. 2002 Bayesian factor regression models in the “large p, small n” paradigm. (See <http://ftp.isds.duke.edu/WorkingPapers/02-12.pdf>.)

The supplementary Electronic Appendix is available at <http://dx.doi.org/10.1098/rstb.2005.1654> or via <http://www.journals.royalsoc.ac.uk>.

## **Effective connectivity: influence, causality and biophysical modeling**



## Comments and Controversies

## Effective connectivity: Influence, causality and biophysical modeling

Pedro A. Valdes-Sosa<sup>a,\*</sup>, Alard Roebroeck<sup>b</sup>, Jean Daunizeau<sup>c,d</sup>, Karl Friston<sup>c</sup><sup>a</sup> Cuban Neuroscience Center, Ave 25 #15202 esquina 158, Cubanacan, Playa, Cuba<sup>b</sup> Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, The Netherlands<sup>c</sup> The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen Square, London, WC1N 3BG, UK<sup>d</sup> Laboratory for Social and Neural Systems Research, Institute of Empirical Research in Economics, University of Zurich, Zurich, Switzerland

## ARTICLE INFO

## Article history:

Received 22 September 2010

Revised 15 March 2011

Accepted 23 March 2011

Available online 6 April 2011

## Keywords:

Granger Causality

Effective connectivity

Dynamic Causal Modeling

EEG

fMRI

## ABSTRACT

This is the final paper in a Comments and Controversies series dedicated to “The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution”. We argue that discovering effective connectivity depends critically on state-space models with biophysically informed observation and state equations. These models have to be endowed with priors on unknown parameters and afford checks for model Identifiability. We consider the similarities and differences among Dynamic Causal Modeling, Granger Causal Modeling and other approaches. We establish links between past and current statistical causal modeling, in terms of Bayesian dependency graphs and Wiener–Akaike–Granger–Schweder influence measures. We show that some of the challenges faced in this field have promising solutions and speculate on future developments.

© 2011 Elsevier Inc. All rights reserved.

## Introduction

Following an empirical evaluation of effective connectivity measurements (David et al., 2008) and a primer on its implications (Friston, 2009a), the Comments and Controversy (C&C) exchange, initiated by Roebroeck et al. (2011b-this issue) and continued by David (2011-this issue), Friston (2011b-this issue), and Roebroeck et al. (2011a-this issue), has provided a lively and constructive discussion on the relative merits of two current techniques, Granger Causal Modeling (GCM)<sup>1</sup> and Dynamic Causal Modeling (DCM), for detecting effective connectivity using EEG/MEG and fMRI time series. The core papers of the C&C have been complemented by authoritative contributions (Bressler and Seth, 2011-this issue; Daunizeau et al., 2011a-this issue; Marinazzo et al., 2011-this issue) that clarify the state of the art for each approach.

This final paper in the series attempts to summarize the main points discussed and elaborate a conceptual framework for the analysis of effective connectivity (Figs. 1 and 2). Inferring effective connectivity comprises the successive steps of model specification, model identification and model (causal) inference (see Fig. 1). These steps are common to DCM, GCM and indeed any evidence-based inference. We will look at the choices made at each stage to clarify current areas of agreement and disagreement, of successes and shortcomings.

This entails a selective review of key issues and lines of work. Although an important area, we will not consider models that are just used to measure statistical associations (i.e. functional connectivity). In other words, we limit our focus to discovering effective connectivity (Friston, 2009a); that is causal relations between neural systems. Importantly, we hope to establish a clear terminology to eschew purely semantic discussions, and perhaps dispel some confusion in this regard. While preparing this material, we were struck with how easy it is to recapitulate heated arguments in other fields (such as econometrics), which were resolved several decades ago. We are also mindful of the importance of referring to prior work, to avoid repeating past mistakes<sup>2</sup> and to identify where more work is needed to address specific problems in the neurosciences.

We shall emphasize several times in this paper that causality is an epistemological concept that can be particularly difficult to capture with equations. This is because one's intuitive understanding of causality becomes inherently constrained whenever one tries to model it. In brief, one can think of causality in at least two distinct ways:

- Temporal precedence, i.e.: causes precede their consequences;
- Physical influence (control), i.e.: changing causes changes their consequences.

\* Corresponding author at: Ave 25 #15202 esquina 158, Cubanacan, Playa, Apartado 6648 Habana 6 CP 10600, Cuba.

E-mail address: [peter@cneuro.edu.cu](mailto:peter@cneuro.edu.cu) (P.A. Valdes-Sosa).

<sup>1</sup> Note that GCM is also used as an acronym for Granger Causal Mapping (Roebroeck et al., 2005). To avoid confusion we shall use GCM mapping or the abbreviation GCMMap.

<sup>2</sup> “Progress, far from consisting in change, depends on retentiveness. When change is absolute there remains nothing to improve and no direction is set for possible improvement: and when experience is not retained, as among savages, infancy is perpetual. Those who cannot remember the past are condemned to repeat it.” George Santayana, *The Life of Reason, Volume 1, 1905*.

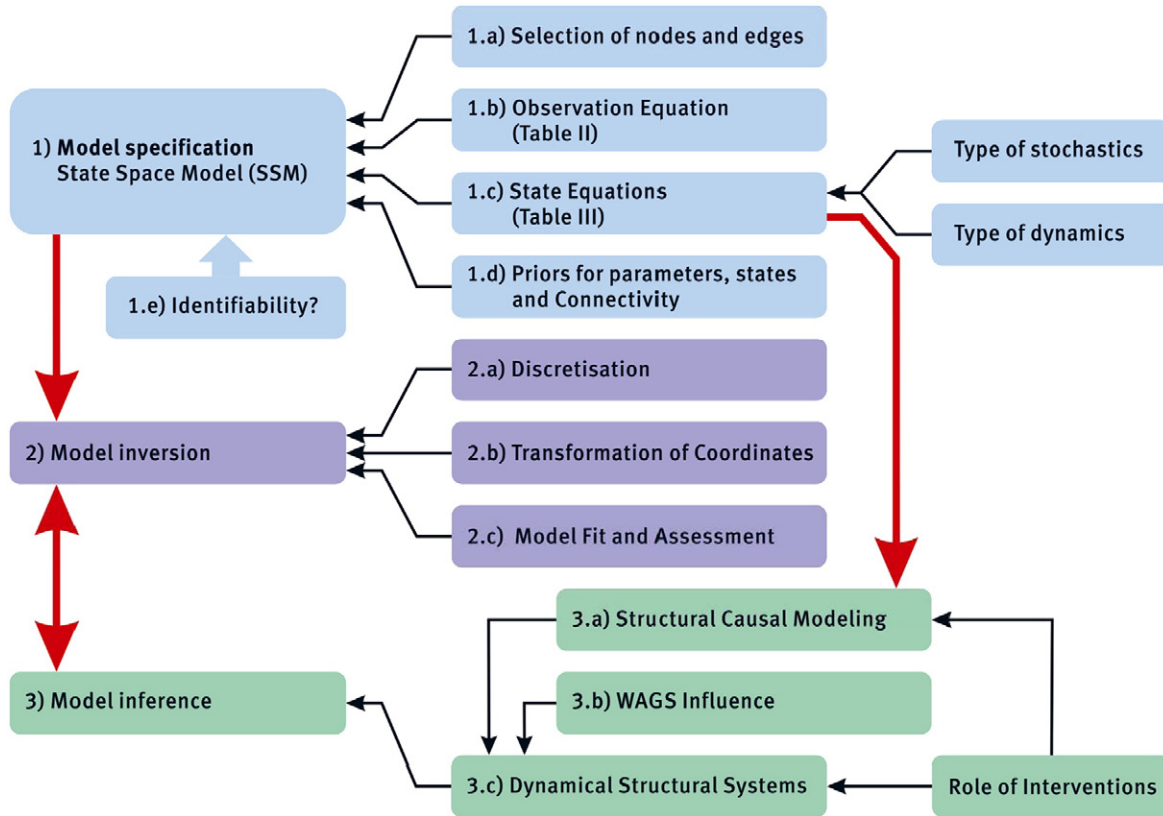


Fig. 1. Overview of causal modeling in Neuroimaging. Overall view of conceptual framework for defining and detecting effective connectivity in Neuroimaging studies.

This distinction is important, since it is the basis for any statistical detection of causal influence. In the context of brain connectivity, identifying causal relationships between two regions in the brain thus depends upon whether one tests for improvement in predictive capacity between temporally distinct neural events or one assesses the distal effect of (experimentally controlled) interventions.

Temporal precedence is the basis for Granger-like (what we call WAGS influence, see [WAGS influence](#) section) inferences about causality. In its simplest form, the idea is the following: A WAGS-causes B if one reduces the uncertainty about the future of B given the

past of A. Statistical tests of WAGS-causality thus rely upon information theoretic measures of predictability (of B given A).

In contradistinction, physical influence speaks to the notion of intervention and control, which has been formalized using a probabilistic framework called causal calculus (Pearl, 2000) ([Structural causal modeling: graphical models and Bayes-Nets](#) section). Observing (or estimating) activity at a network node potentially provides information about its effects at remote nodes. However, physically acting upon (e.g., fixing) this activity effectively removes any other physical influence this node receives. This means that inferences based on the effects of an intervention are somewhat different in nature from those based on purely observational effects. Generally speaking, inference on structural causality rests on modeling the effects of (controlled) experimental manipulations of the system, c.f. the popular quote ‘no causes in, no causes out’ (Cartwright, 2007). As we shall see later, these two approaches can be combined ([Dynamic structural causal modeling](#) section).

The structure of the paper is as follows. We first review the types of models used for studying effective connectivity. We then touch briefly on the methods used to invert and make inferences about these models. We then provide a brief summary of modern statistical causal modeling, list some current approaches in the literature and discuss their relevance to brain imaging. Finally, we list outstanding issues that could be addressed and state our conclusions.

**Model specification**

*State-space models of effective connectivity*

From the C&C discussion, there seems to be a consensus that discovering effective connectivity in Neuroimaging is essentially a comparison of generative models based on state-space models (SSM) of controllable (i.e., “causal” in a control theory sense) biophysical processes that have hidden neural states and possibly exogenous

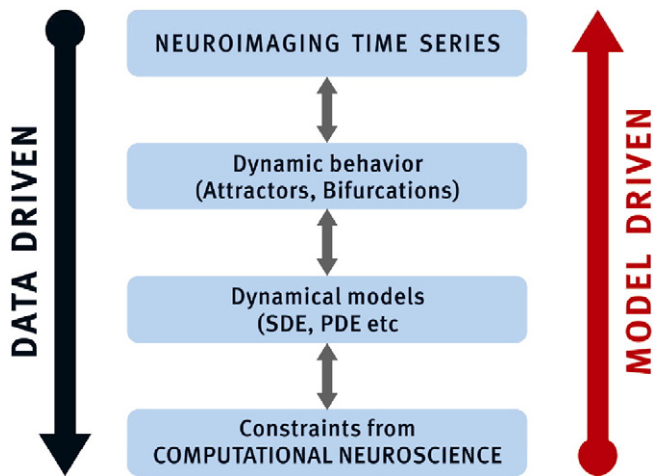


Fig. 2. Data and model driven approaches to causal modeling. Data driven approaches look for nonparametric models that not only fit the data but also describe important dynamical properties. They complement hypothesis driven approaches that are not only constrained by having to explain dynamical behavior but also provide links to computational models of brain function.



input. While having a long history in engineering (Candy, 2006; Kailath, 1980), SSM was only introduced recently for inference on hidden neural states (Valdes-Sosa et al., 1999; Valdes-Sosa et al., 1996; Valdés-Sosa et al., 2009a). For a comprehensive review of SSM and its application in Neuroscience see the forthcoming book (Ozaki, 2011).

Neural states describe the activity of a set of “nodes” that comprise a graph, the purpose of causal discovery being the identification of active links (edges or connections) in the graph. The nodes can be associated with neural populations at different levels; most commonly at the macroscopic (whole brain areas) or mesoscopic (sub-areas to cortical columns) level. These state-space models have unknown parameters (e.g., effective connectivity) and hyperparameters (e.g., the amplitude of random fluctuations). The specific model, states, parameters, hyperparameters and observables chosen determines the type of analysis and the nature of the final inference about causality. These choices are summarized in Fig. 1 (Step 1).

Given a set of observations or brain measurements, the first problem is: which data features are relevant for detecting causal influences? The most efficient way to address this question is to specify a *generative model*, i.e. a set of equations that quantify how observed data are affected by the presence of causal links. Put simply, this model translates the assumption of (i) *temporal precedence* or (ii) *physical influence* into how data should appear, given that (i) or (ii) is true. By presenting the data to generative models, model comparison can then be used to decide whether some causal link is likely to be present (by comparing models with and without that link). We now turn to the specification of generative models, in the form of a SSM.

#### Nodes and random variables

The first things we consider are the basic units or nodes, among which one wants to find causal links. These are usually modeled as macroscopic, coarse grained, ensembles of neurons, whose activity is summarized by a time varying state vector  $x_r(t)$  or  $x(r, t)$ :  $r \in R$ . For example  $x(t)$  could be instantaneous (ensemble average) post-synaptic membrane depolarization or pre-synaptic firing rate of neurons. The set  $R$  of nodes is usually taken as a small number of neural masses corresponding to pre-selected regions of interest (ROI) as is typical in both DCM and GCM. However, there has been recent interest in making  $R$  a continuous manifold (i.e. the cortex) that is approximated by a very high dimensional representation at the voxel level. We denote the complete set of random variables associated with each node as  $X = \{X_i, X_i\}$  whose joint distribution is described using a generative model.  $X_i$  is the set of nodes without node  $i$  and  $p(x) \triangleq p(X=x)$ .

#### The observation equation

Any model always includes an explicit or implicit observation equation that generally varies with the imaging modality. This equation specifies how hidden (neuronal) states  $x_r(t)$  produce observable data  $y_q(t_k)$ :  $q \in Q$ . This is the sensor data sampled at discrete time points  $t_k = k\Delta$ :

$$y_q(t_k) = g(x_r, t) + e(t_k) : r \in R_r, t \in [t_k, t_{k-1}] \quad (1)$$

for  $k=1 \dots K$ . It is important to note that observations at a given sensor  $q$  only reflect neural states from a subset of brain sites, modified by the function  $g$  over a time interval determined by the sampling period  $\Delta t$  and corrupted by instrumental noise  $e(t_k)$ . When the sampling period is not considered explicitly, the observations are denoted by  $y_q(k)$ . In most cases, this mapping does not need to be dynamic since there is no physical feedback from observed data to brain processes. In this special case, the observation equation reduces to an instantaneous transformation:  $Y(t) = \tilde{g}(X(t))$ , where  $\tilde{g}$  is

derived from  $g$  and any retarded (past) hidden states have been absorbed in  $X(t)$  (e.g., to model hemodynamic convolutions).

A selected collection of observation equations used in Neuroimaging is provided in Table 1. The observation equation is sometimes simplified by assuming that observed data is a direct measurement of neural states (with negligible error). While this might be an acceptable assumption for invasive electrophysiological recordings, it is inappropriate in many other situations: for example, much of the activity in the brain is reflected in the EEG/MEG via the lead field with a resultant spatial smearing. For the BOLD signal, the C&C articles have discussed exhaustively the need to account for temporal smearing produced by the hemodynamic response function (HRF) when analyzing BOLD responses. This is important for fMRI because the sampling period is quite large with respect to the time course of neural events (we shall elaborate on this below).

Instrumental or sensor noise can seriously affect the results of causal analyses. One simple approach to causal modeling is to take the observation equation out of the picture by inverting the observation equation (i.e., mapping from data to hidden states). The estimated states can then be used for determining effective connectivity. This approach has been taken both for the EEG (Supp et al., 2007) and fMRI (David et al., 2008). However, this is suboptimal because it assumes that the causal modeling of hidden states is conditionally independent of the mapping from data. This is generally not the case (e.g., non-identifiability between observation and evolution processes described below). The optimal statistical procedure is to invert the complete generative model, including the observation and state equations modeling the evolution of hidden states. This properly accommodates conditional dependencies between parameters of the observer and state equations. A nice example of this is DCM for EEG and MEG, in which a SSM of coupled neuronal sources and a conventional electromagnetic forward model are inverted together. This means the parameters describing the spatial deployment of sources (e.g., dipole orientation and location) are optimized in relation to parameters controlling the effective connectivity among hidden sources. This sort of combined estimation has been described for simple noise models (Table 1-#2 by Nalatore et al. (2007)). For fMRI, DCM models the hemodynamic response with hidden physiological states like blood flow and volume and then uses a nonlinear observer function to generate BOLD responses (Table 2-#4). Early applications of GCM did not model the HRF but in recent years a number of papers have included explicit observation models in GCM (Ge et al., 2009; Havlicek et al., 2010), which have even incorporated the full nonlinear HRF model used in DCM (Havlicek et al., 2009; Havlicek et al., 2011).

#### The state equation

The evolution of the neuronal states is specified by the dynamical equations:

$$x_r(t) = f\left(x_{r' \in R_r}(\tau), u(\tau), \xi_{r' \in R_r}(\tau)\right) : \tau \in (t, t-t_0]. \quad (2)$$

This equation<sup>3</sup> expresses,  $x_r(t)$ , the state vector of node  $r$  at time  $t$  as a function of:

- the states of nodes  $x_{r'}(\tau)$ :  $r' \in R_r \subseteq R$
- exogenous inputs,  $u(\tau)$  and a
- stochastic process  $\xi_{r'}(\tau)$ .

Note that the dependence of the current states at node  $r$  may be contingent on values of other variables from an arbitrary past from  $t-t_0$  to just before  $t$ . The time dependence of Eq. (2) is important because it allows to model feedback processes within the network.

<sup>3</sup> We use the following conventions for intervals, [a,b) indicates that the left endpoint is included but not the right one and that b precedes a.

**Table 1**  
 Observation equations. Examples of observation equations used for causal modeling of effective connectivity in the recent literature. Abbreviations: discrete (D), continuous (C), white noise (WN). Note for Models #5 and #6 the observation equation is considered as all the equations except for the (neural) state equations. Strictly speaking, the observer function is just the first equality (because the subsequent equations of motion are part of the state equation); however, we have presented the equations like this so that one can compare instantaneous observation equations that are a function of hidden states, convolution operators or a set of differential equations that take hidden neuronal states as their inputs.

Model	Observation equation	Measurement	Space	Time	Equation type	Kind of stochastic process
1 None (Bressler and Seth, 2010)	$y(r, k) = x(r, k)$	EEG/fMRI	D	D	Identity	none
2 Added noise (Nalatore et al., 2007)	$y(r, k) = x(r, k) + e(r, t)$	fMRI	D	D	Linear regression	WN
3 Spatial smearing (Riera et al., 2006)	$y(q, t) = \int_{r \in R} k(r, r') x(r', t) dr' + e(r, t)$	EEG/MEG	D	C	Volterra integral equation with noise	none
4 Convolution with linear HRF (Glover, 1999)	$y(r, k) = \int_{-\infty}^{t-k} h(\tau) x(r, t-\tau) d\tau + e(r, k)$	fMRI	D	C	Temporal convolution	WN
5 Nonlinear HRF function (Friston et al., 2000)	$y_t = V_0(a_1(1-q_t) - a_2(1-v_t))$ $\dot{v}_t = \frac{1}{\tau_0} (f_t - v_t^{1/\alpha})$ $\dot{q}_t = \frac{1}{\tau_0} \left( \frac{f_t(1 - (1-E_0)^{1/f_t})}{E_0} - \frac{q_t}{v_t^{1-1/\alpha}} \right)$ $\dot{s}_t = \varepsilon u_t - \frac{1}{\tau_s} s_t - \frac{1}{\tau_f} (f_t - 1)$ $f_t = s_t$	fMRI	C	C	Nonlinear differential equation	none
6 Nonlinear HRF function (Valdes-Sosa et al., 2009a)	$\begin{cases} \dot{g}_e(t) = s_e(t) \\ \dot{s}_e(t) = \frac{a_e}{\tau_e} (u_e(t - \delta_e) - 1) - \frac{2}{\tau_e} s_e(t) - \frac{1}{\tau_e^2} (g_e(t) - 1) \\ \dot{g}_i(t) = s_i(t) \\ \dot{s}_i(t) = \frac{a_i}{\tau_i} (u_i(t - \delta_i) - 1) - \frac{2}{\tau_i} s_i(t) - \frac{1}{\tau_i^2} (g_i(t) - 1) \end{cases}$ $x = \frac{1}{1 + e^{-c(g_e(t) - d)}}$ $\begin{cases} \dot{f}(t) = s_f(t) \\ \dot{s}_f(t) = \varepsilon (u_e(t - \delta_f) - 1) - \frac{s_f(t)}{\tau_s} - \frac{f(t) - 1}{\tau_f} \end{cases}$ $m_i(t) = g_i(t), m_e(t) = \frac{2-x}{2-x_0} g_e(t), m(t) = \frac{\gamma m_e(t) + m_i(t)}{\gamma + 1}$ $\begin{cases} \dot{v}(t) = \frac{1}{\tau_0} (f(t) - f_{out}(v, t)) \\ \dot{q}(t) = \frac{1}{\tau_0} \left( m(t) - f_{out}(v, t) \frac{q(t)}{v(t)} \right), f_{out}(v, t) = v^{\frac{1}{\alpha}} \end{cases}$ $y(t) = V_0(a_1(1-q) - a_2(1-v))$	EEG/fMRI	C	C	Nonlinear random differential algebraic equation	none

Many specific forms have been proposed for Eq. (2); some examples are listed in Table 2, which is just a selection to illustrate different points discussed below. Some types of equations, to our knowledge, have not been yet used for the analysis of effective connectivity. Several general observations emerge from these examples:

*Discrete versus continuous time modeling:* The state equations for GCM have been for the most part discrete time recurrence models (Bressler and Seth, 2010). Those for DCM are based on continuous time models (differential equations) (Friston, 2009a). The latter have advantages in dealing with the problem of temporal aggregation and sub-sampling as we shall see below. In fact, DCM is distinguished from general SSM by the fact it is based on differential equations of one sort or another.

*Discrete versus continuous spatial modeling:* GCM has been applied to continuous space (neural fields) though limited to discrete time (Galka et al., 2004; Valdes-Sosa, 2004). DCM has mainly been developed for discrete-space (ROIs) and, as mentioned above, continuous time. State space models that are continuous in space and time have recently been looked at in the context of neural field equations (Daunizeau et al., 2009c; Galka et al., 2008).

*Type of equation:* GCM has been predominantly based on linear stochastic recurrence (autoregressive) models (Bressler and Seth, 2010). DCM on the other hand has popularized the use of deterministic ordinary differential equations (ODE). These range from simple bilinear forms for fMRI that accommodate interactions between the input and the state variables (Friston, 2009a) to complicated nonlinear equations describing the ensemble dynamics of neural mass models. In their most comprehensive form, these models can be formulated as Hierarchical Dynamical Models (HDM) (Friston, 2008a,b). HDM uses nonlinear random differential equations and static nonlinearities, which can be deployed hierarchically to reproduce most known parametric models. However, as noted in the C&C, GCM is not limited to linear models. GCM mapping (Roebroek et al., 2005) uses an (implicit) bilinear model, because the Autoregressive coefficients depend on the stimulus; this bilinearity is explicit in GCM on manifolds (Valdés-Sosa et al., 2005) GCM has also been extended to cover nonlinear state-equations (Freiwald et al., 1999; Marinazzo et al., 2011). The type of models used as state equations are very varied (and are sometimes equivalent). One can find (for discrete spatial nodes) recurrence equations, ordinary differential equations, and (for neural fields) differential-integral and partial differential equations.

**Table 2**

State equations. Examples of the state equations used in the recent literature for causal modeling of effective connectivity. Abbreviations: C (continuous), D (discrete), WN (white noise).

Model	State equation	Space	Time	Equation type	Stochastic process
Linear GCM (Bressler and Seth, 2010)	$x(r, k) = \sum_{r'=1}^{N_r} \sum_{l=1}^T a_l(r, r') x(r', k-l) + \xi(r, k)$	D	D	Linear multivariate linear autoregressive (VAR)	WN
2 Nonlinear GCM (Freiwald et al., 1999)	$x(r, k) = \sum_{r'=1}^{N_r} \sum_{l=1}^T a[l, r, r'; x(r', k-l)] x(r', k-l) + \xi(r, k)$	D	D	Nonlinear nonparametric VAR (NNp_MVAR)	WN
3 Linear bivariate GCM mapping (Roebroeck et al., 2005)	$\begin{bmatrix} x(r, k) \\ x(ROI, k) \end{bmatrix} = \sum_{r'=1}^{N_I} \begin{bmatrix} a_l(r, r) & a_l(r, ROI) \\ a_l(ROI, r) & a_l(ROI, ROI) \end{bmatrix} \begin{bmatrix} x(r, k-l) \\ x(ROI, k-l) \end{bmatrix} + \begin{bmatrix} \xi(r, k) \\ \xi(ROI, k) \end{bmatrix}$ $\forall r \in R \quad x(ROI, k) = \int_{r \in R} x(r, k) dr$	D	D	VAR since $a_l(r, r')$ Implicitly bi-linear changes with state. (GCMMap)	WN
4 Linear GCM on spatial manifold (Valdés et al., 2006)	$x(r, k) = \sum_{l=1}^{N_I} \int_{r \in R} a_l(r, r') x(r', k-l) dr' + \xi(r, k)$	C	D	Implicitly bi-linear VAR as in 3	WN
5 Nonlinear DCM (Stephan et al., 2008)	$\dot{x}(r, t) = \sum_{r'=1}^{N_x} a(r, r') x(r', t) + \sum_{i=1}^{N_u} u(i, t) \sum_{r'=1}^{N_x} b(r, r') x(r', t) + \sum_{r'=1}^{N_x} \sum_{r''=1}^{N_x} d(r, r', r'') x(r', t) x(r'', t) + \sum_{i=1}^{N_u} c(r, i) u(i, t)$	D	D	Differential equation bilinear in both states and inputs (DE)	None
6 Neural mass model (Valdes et al., 1999)	$\dot{x}(r, t) = f(x(r, t)) + \xi(r, t)$	C	C	Ito stochastic differential (SDE)	WN as formal derivative of Brownian motion
7 Hierarchical dynamic causal model (Friston, 2008a,b)	$\dot{x}(r, t) = f(x(r, t), u(t)) + \xi(r, t)$	D	C	General nonlinear (HDM)	Analytic, non-Markovian
8 Neural field (Jirsa et al., 2002)	$\left(\frac{\partial^2}{\partial t^2} + 2\omega_0 \frac{\partial}{\partial t} + \omega_0^2 - v^2 \nabla^2\right)^{3/2} x(r, t) = \left(\omega_0^3 + \omega_0^2 \frac{\partial}{\partial t}\right) S[x(r, t) + \xi(r, t)]$	C	C	Stochastic fractional partial differential (StPDE)	WN
9 Modified neural field (P. A. Valdes-Sosa et al., 2009a)	$\ddot{x}(r, t) = f(\dot{x}(r, t), x(r, t)) + S(z(r, t)) + \xi(r, t)$ $z(r, t) = \int_R a(r, r') x(r', \tau(r, r')) dr'$ $\tau(r, r') = t - \frac{ r-r' }{v}$	C	C	Random differential-algebraic-equation (RDE)	General

To underscore the variety of forms for effective connectivity, we note entry #8 in Table 2 which boasts a fractional differential operator! Fractional operators arise in the context of neural fields in more than one dimension; they result from the Fourier transform of a synaptic connection density that is a continuous function of physical distance. However, the ensuing fractional differential operators are usually replaced by ordinary (partial) differential operators, when numerically solving the neural wave propagation equation given in Table 2; see Bojak and Liley (2010) and Coombes et al. (2007) for the so-called 'long wavelength approximation'.

Among other things, it can be important to include time delays in the state equation; this is usually avoided when possible to keep the numerics simple (delay differential equations are infinite dimensional) and are generally considered unnecessary for fMRI. However, delays are crucial when modeling electromagnentic data, since they can have a profound effect on systems dynamics (Brandt et al., 2007). For example, delayed excitatory connections can have an inhibitory instantaneous effect. In fact starting with Jansen and Rit (1995) it has been common practice to include time delays. This can be implemented within the framework of ODEs; David et al. (2006) describe an ODE approximation to delayed differential equations in the context of DCM for EEG and MEG. An example of the potential richness of model structures is found in Valdes-Sosa et al. (2009a) in a neural field forward model for EEG/fMRI fusion, which includes anatomical connections and delays as algebraic constraints. This approach (of including algebraic constraints) affords the possibility of building complex models from nonlinear components, using simple interconnection rules—something that has been developed for control theory (Shampine and Gahinet, 2006). Note that algebraic constraints may be added to any of the aforementioned forms of state equation.

*Type of stochastics:* for GCM-type modeling with discrete-time models, Gaussian White Noise (GWN) is usually assumed for the random fluctuations (state noise) or driving forces (innovations) for the SSM and poses no special difficulties. However in continuous time the problem becomes more intricate. A popular approach is to treat the innovation as nowhere differentiable but continuous Gaussian White Noise (the “derivative” of Brownian motion (i.e., a Wiener process)). When added to ordinary differential equations we obtain “stochastic differential equations” (SDE) as described in Medvegyev (2007) and used for connectivity analysis of neural masses in Riera et al., (2007a,b), Riera et al. (2006). Wiener noise is also central to the theory of Stochastic Partial Differential Equations (SPDE) (Holden et al., 1996), which may play a similar role in neural field theory as SDEs have played for neural masses (Shardlow, 2003).

Despite the historical predominance of the classical SDE formulation in econometrics (and SSM generally), we wish to emphasize the following developments, which may take us (in the biological sciences) in a different direction:

1. The first is the development of a theory for “random differential equations” (RDE) (Jentzen and Kloeden, 2009). Here randomness is not limited to additive Gaussian white noise because the parameters of the state equations are treated as stochastic. RDE are treated as deterministic ODE, in the spirit of Sussmann (1977), an approach usable to great advantage in extensive neural mass modeling (Valdes-Sosa et al., 2009a) that is implicitly a neural field.
2. The second development, also motivated by dissatisfaction with classical SDE was introduced in Friston and Daunizeau (2008). In that paper, it was argued that DCMs should be based on stochastic processes, whose sample paths are infinitely differentiable—in other words, analytic and non-Markovian.

Though overlooked in the heyday of SDE theory, this type of process was described very early by Belyaev (1959).<sup>4</sup> In fact any band-limited stochastic process is an example of an analytic random process; a stochastic process with a spectrum that decreases sharply with frequency, has long memory, and is non-Markovian (Łuczka, 2005). The connection between analytic stochastic processes and RDE can be found in Calbo et al. (2010). An interesting point here is that for the process to be analytic its successive derivatives must have finite variances, as explained in Friston and Daunizeau (2008). This leads to the generalization of classical SSM into generalized coordinates of motion that model high-order temporal derivatives explicitly. As pointed out in Friston (2008a,b), it is possible to cast an RDE as a SDE by truncating the temporal derivatives at some suitably high order (see also Carbonell et al., 2007). However, this is not necessary because the theory and numerics for RDEs in generalized coordinates are simpler than for the equivalent SDE (and avoid the unwieldy calculus of Markovian formulations, due to Ito and Stratonovich).

3. The third development is the recognition that non-Markovian processes may be essential for neurobiological modeling. This has been studied for some time in physics (Łuczka, 2005) but has only recently been pointed out by Friston (2008a,b) in a neuroscience setting. In fact, Faugeras et al. (2009) provide a constructive mean-field analysis of multi-population neural networks with random synaptic weights and stochastic inputs that exhibits, as a main characteristic, the emergence of non-Markovian stochastics.
4. Finally the fourth development is the emergence of neural field models (Coombes, 2010; Deco et al., 2008), which not only poses much larger scale problems but also the use of integral equations, differential–integral equations, and partial differential equations which have yet to be exploited by DCM or GCM.

*Biophysical versus non-parametric motivation:* As discussed above, there is an ever increasing use of biophysically motivated neural mass and field state equations and, in principle, these are preferred when possible because they bring biophysical constraints to bear on model inversion and inference. When carrying out exploratory analyses with very large SSM, it may be acceptable to use simple linear or bilinear models as long as basic aspects of modeling are not omitted.

*Further generalizations:* We want to end this subsection by mentioning that there is a wealth of theory and numerics for other stochastic (point) processes (Aalen and Frigessi, 2007; Commenges and Gégout-Petit, 2009) that have not yet been, to our knowledge, treated formally in Neuroimaging. Spike trains, interictal-spikes, and random short-timed external stimuli may be treated as point processes and can be analyzed in a unified framework with the more familiar continuous time series. This theory even encompasses mixtures of slow wave and spike trains. Causal modeling depends very specifically on the temporal and spatial scales chosen and the implicit level of granularity chosen to characterize functional brain architectures. For example, if we were to study the interaction of two neural masses and model the propagation of activity between them in detail, we would have to make use of the PDE that describes the propagation of nerve impulses. If we eschew this level of detail, we may just model the fact that afferent activity arrives at a neural mass with a conduction delay and use delay differential equations. In short, the specification of the appropriate SSM depends on the spatial and temporal scale that one is analyzing. For example, in concurrent EEG/fMRI analysis of resting state oscillations

(Martínez-Montes et al., 2004) the temporal scale of interesting phenomena (fluctuations of the EEG spectrum) is such that one may convolve the EEG signal and do away with the observation equation! This is exactly the opposite of the deconvolution approach mentioned above. The purpose of Tables 1 and 2 is to highlight the variety of forms that both state and observation equations can take; for example, in Table 2–#6 key differential equations are transformed into differential algebraic equations to great computational advantage (Valdes-Sosa et al., 2009a).

#### Specification of priors

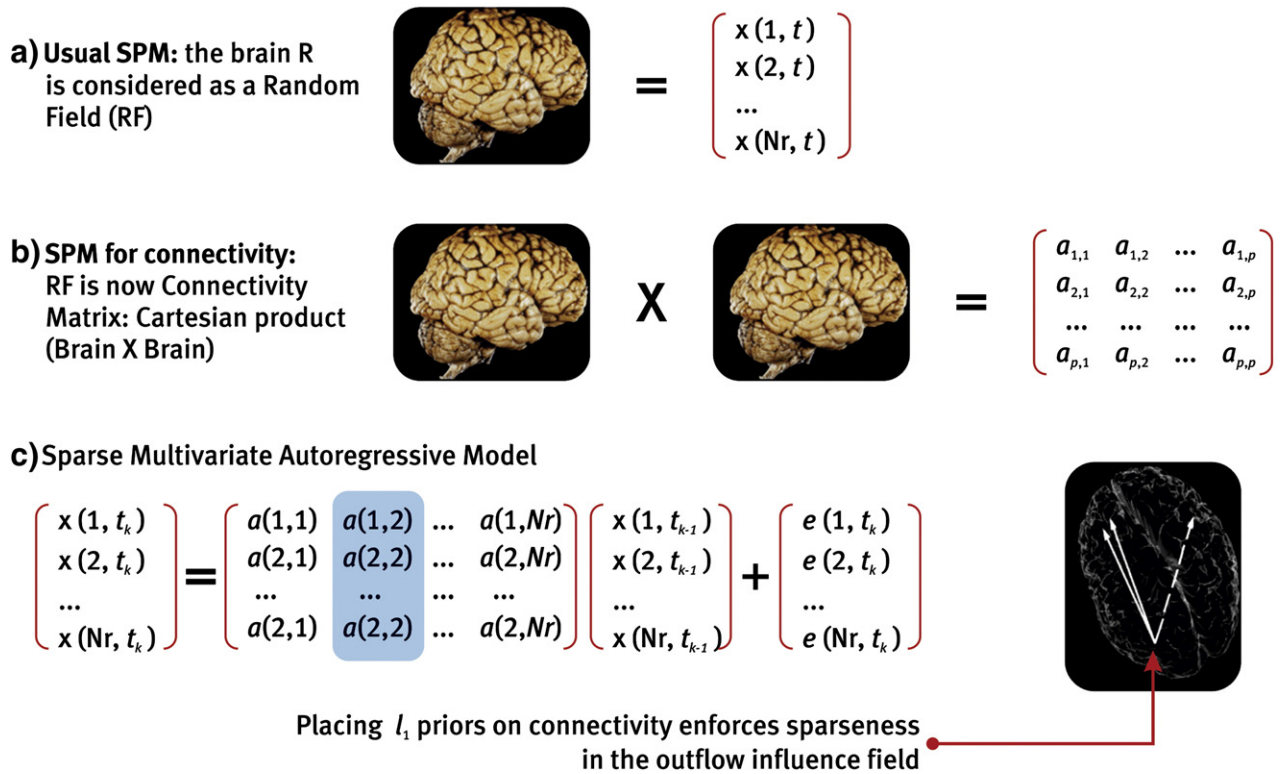
It is safe to say that the Neuroimaging (and perhaps generally) modeling can be cast as Bayesian inference. This is just a euphemism for saying that inference rests on probability theory. The two key aspects of Bayesian inference we will appeal to in this article are (i) the importance of prior beliefs that form an explicit part of the generative model; and (ii) the central role of Bayesian model evidence in optimizing (comparing and selecting) models to test hypotheses. In terms of priors, it was very clear in an early state space model for EEG connectivity (Valdes-Sosa et al., 1996) that without prior assumptions about the spatial and temporal properties of the EEG, it was not possible to even attempt source reconstruction. Indeed the whole literature on ill-posed inverse problems rests on regularization that can be cast in terms of prior beliefs.

In the SSM formulation, priors may be placed upon parameters in the observation and state equations, and the states themselves (e.g., through priors on the higher-order motion of states or state-noise). Sometimes, it may be necessary to place priors on the priors (hyperpriors) to control model complexity. There has been an increasing use of priors in fMRI research, as clearly formulated in the DCM and HDM framework (Friston, 2008a,b). In connectivity analyses, in addition to the usual use of priors to constrain the range of parameters quantitatively; formal or structural priors are crucial for switching off subsets of connections to form different (alternative) models of observed data. Effectively, this specifies the model in terms of its adjacency matrix, which defines allowable connections or conditional dependencies among nodes. Conditional independence (absence of an edge or anti-edge) is easy to specify by using a prior expectation of zero and with zero variance. This is an explicit part of model specification in DCM and is implicit in Granger tests of autoregressive models, with and without a particular autoregression coefficient.

Crucially, formal priors are not restricted to the parameters of a model; they can also be applied to the form of the prior density over parameters. These can be regarded as formal hyperpriors. An important example here is the prior belief that connections are distributed sparsely (with lots of small or absent connections and a small number of strong connections). This sort of hyperprior can be implemented by assuming the prior over parameters is sparse. A nice example of this can be found in Valdés-Sosa (2004), Valdés-Sosa et al. (2005, 2006), and Sánchez-Bornot et al. (2008).

The essential features of their model are shown in Fig. 3. The authors analyzed slow fluctuations in resting state EEG. In this situation, convolving these electrophysiological fluctuations with a HRF affords (convolved) EEG and BOLD signals on the same time scale, permitting lag-based inference. An example is presented in Fig. 4, which shows the results of GCM Mapping for 579 ROIs from an EEG inverse solution and concurrent BOLD signals. The EEG sources were obtained via a time resolved VARETA inverse solution (Bosch-Bayard et al., 2001) at the peak of the alpha rhythm. The graphs present the result of inverting a (first order) multivariate vector autoregression model, where a sparse  $l_1$  norm penalty was imposed on the parameters (coefficient matrix). The implications of these results will be further discussed in Conclusion and suggestions for further work section below.

<sup>4</sup> With suggestion by A.N. Kolmogorov.



**Fig. 3.** Bayesian inference on the connectivity matrix as a random field. a) Causal modeling in Neuroimaging has concentrated on inference on neural states  $x(r, t) \in \mathbb{R}$  defined on a subset of nodes in the brain. However, spatial priors can be used to extend models into the spatial domain (cf., minimum norm priors over current source densities in EEG/MEG inverse problems). b) In connectivity analysis, attention shifts to the AR (connectivity) matrix (or function)  $a(r, r')$ , where the ordered pairs  $(r, r')$  belong to the Cartesian product  $R \times R$ . For this type of inference, priors are now placed on the connectivity matrix. c) Sparse multivariate autoregression obtains by penalizing the columns of a full multivariate autoregressive model (Valdés-Sosa et al., 2005) thus forcing the columns of the connectivity matrix to be sparse. The columns of the connectivity matrix are the “outfields” that map each voxel to the rest of the brain. This is an example of using sparse (spatial) hyperpriors to regularize a very difficult inverse problem in causal modeling.

*Model comparison and Identifiability*

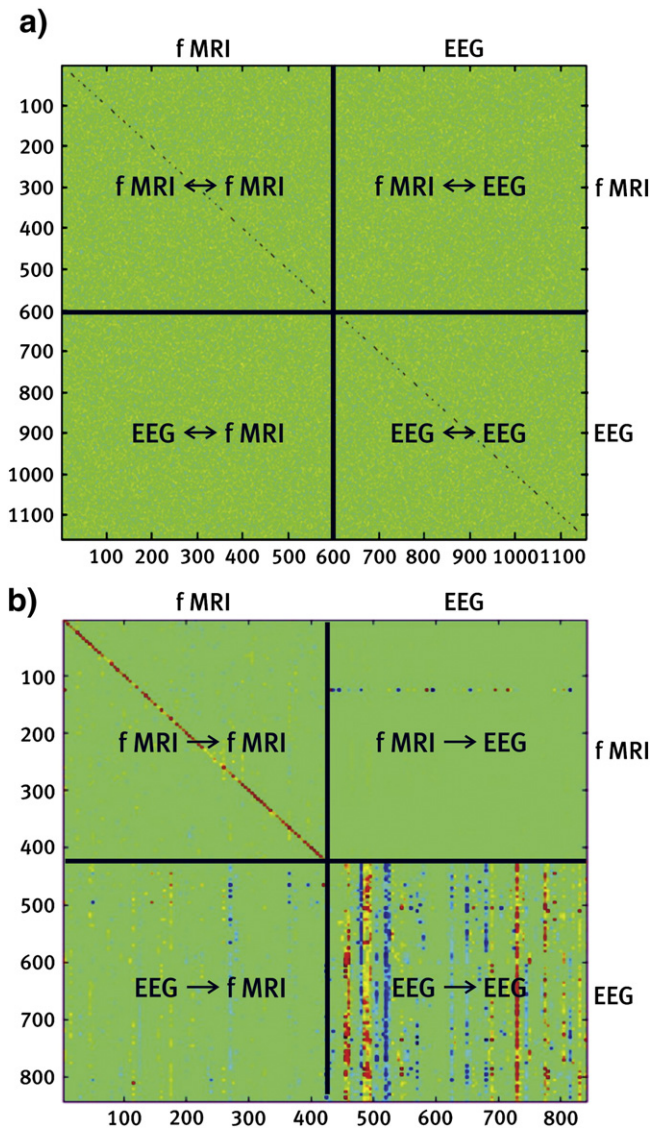
As we have seen, the SSMs considered for EEG and fMRI analysis are becoming increasingly complex, with greater spatial or temporal coverage and improved biological realism. A fundamental question arises: *Are these models identifiable?* That is to say, are all states and parameters uniquely determined by a given set of data? This is a basic issue for all inverse problems, and indeed we are faced with a dynamical inverse problem of the greatest importance. For example, recent discussions about whether lag information can be derived from the fMRI signal (in spite of heavy smoothing by the HRF and the subsequent sub sampling) can be understood in terms of the identifiability of delays in the corresponding SSM. It is striking that, in spite of much classical work on the Identifiability of SSMs (see for example Ljung and Glad, 1994), a systematic treatment of identification has not been performed for Neuroimaging models (but see below). An example of the type of problem encountered is the complaint that a model with many neural masses and different configurations or parameter values can produce traces that “look the same as an observed response”.

Identifiability has been addressed in bioinformatics, where much theory for nonlinear SSM has been developed (Anguelova and Wennberg, 2010; August and Papachristodoulou, 2009). Of particular note is DAISY, a computer algebra system for checking nonlinear SSM Identifiability (Saccomani et al., 2010). Another framework for modeling and fitting systems defined by differential equations in bioinformatics is “Potters Wheel” (Maiwald and Timmer, 2008), which uses a profile likelihood approach (Raue et al., 2009) to explore “practical identifiability” in addition to structural (theoretical) Identifiability. So why has Neuroimaging not developed similar schemes?

In fact, it has. In a Bayesian setting the issue of model (and parameter) identifiability is resolved though Bayesian model com-

parison. If two models generate exactly the same data with the same number of parameters (complexity), then their evidence will be identical. This means there is no evidence for one model over the other and they cannot be distinguished. We will refer a lot to model evidence in what follows: model evidence is simply the probability of the data given the model. It is the marginal likelihood that obtains from marginalizing the likelihood over unknown model parameters. This is useful to remember because it means the likelihood of a model (the probability of data given a model and its parameters) is a special case of model evidence that results when we ignore uncertainty about the parameters. In the same way, classical likelihood ratio tests of two models are special cases of Bayes Factors used in Bayesian model comparison. In this context, identifiability is a particular aspect of model comparison. Identifiability mandates that changing a component of a model changes the model evidence. This is the basic idea behind the profile likelihood approach (Raue et al., 2009), which is based on the profile of the evidence for models with different parameter values. There are other examples that can be regarded as special cases of model comparison; for example, the Kullback–Leibler information criterion proposed for model identification (Chen et al., 2009). The evidence can be decomposed into an accuracy and complexity term (see Penny et al., 2004). Interestingly, the complexity term is the Kullback–Leibler divergence between the posterior and prior densities over parameters. This means that in the absence of informative priors, model evidence reduces to accuracy; and identifiability reduces to a (nontrivial) change in the accuracy or fit when changing a model or parameter.

The Bayes–Net literature (see below) has dealt with the problem of Identifiability for graphical causal models at its inception (Spirtes et al., 2000). It can be shown that a given data set can be compatible not with a single causal model but with an equivalence class of models



**Fig. 4.** Sparse multivariate autoregression of concurrent EEG/fMRI recordings. Intra and inter modality connectivity matrix for a concurrent EEG/fMRI recordings. The data analyzed here were the time courses of the average activity in 579 ROI: for BOLD (first half of data vector) and EEG power at the alpha peak. A first-order sparse multivariate autoregressive model was fitted with an  $l_1$  norm (hyper) prior on the coefficient matrix. The t-statistics of the autoregression coefficients were used for display. The color bar is scaled to the largest absolute value of the matrix, where green codes for zero. a) the innovation covariance matrix reflecting the absence of contemporaneous influences: b) t-statistics for the lag 1 AR coefficients.

(that all have the same evidence). The implications of this for Neuroimaging have been considered in Ramsey et al. (2010). From this discussion, it becomes clear that the ability to measure model evidence (or some proxy) is absolutely essential to make sensible

**Table 3**  
 Classification of observation and state equations used in Neuroimaging state-space models. Generic models lack specific biophysical constraints but are widely applicable. Biophysically informed models are hypothesis driven and may afford more efficient inference (if correct). The term parametric refers to models with a small enough parameter set to be identifiable without additional priors but that may yield biased estimators. Nonparametric models are richly parameterized and therefore require prior distributions to be estimable but are generally unbiased.

	Observation model		State model	
	Parametric	Non-parametric	Parametric	Non-parametric
Generic	Linear canonical HRF (Glover, 1999)	Linear spline HRF (Marrelec et al., 2003)	GCM (Bressler and Seth, 2010) Switching VAR (Smith et al., 2010a) bilinear discrete DCM (Penny et al., 2005)	GCM (Roebroeck et al., 2005)
Biophysically informed	DCM nonlinear HRF (Friston et al., 2000)	-	Neural mass models (Valdes et al., 1999) Biophysical DCM (Moran et al., 2008)	Neural fields (Daunizeau et al., 2009c)

inferences about models or architectures generating observed data. This is at the heart of evidence-based inference and DCM.

*Summary*

State space models for Neuroimaging come in an ever increasing variety of forms (Tables 1 and 2). It is useful to classify the types of models used in terms of their observation and state equations, as in Table 3. Here, we see a distinction between models that are fairly generic (in that they are not based on biophysical assumptions) and those that correspond to biologically informed models. The canonical HRF model is an example of generic HRF. Conventional GCM is based on a generic model for neural states: the VAR model and has been extended to switching VAR and bilinear models, the latter used in some forms of DCM. Being generic is, at the same time, a strength and weakness; biophysical models allow much more precise and informed inference—but only if the model is right or can be optimized in terms of its evidence. We have also seen the key role that model evidence plays in both making causal inferences by comparing models and (implicitly) establishing their identifiability. The evidence for a model depends on both accuracy and complexity and the complexity of the model depends on its priors.

Another distinction between models is their complexity (e.g., number of parameters they call on). It is clear that without prior beliefs, one cannot estimate more parameters than the degrees of freedom in the data available. However, modern statistical learning has gone beyond low dimensional parametric models to embrace non-parametric models with very high dimensional parameter spaces. The effective number of degrees of freedom is controlled by the use of priors. DCM has been concerned mainly with hypothesis driven parametric models, as has conventional GCM. However, nonparametric models, such as smoothness priors in the time domain have been used to estimate the HRF (Marrelec et al., 2003). Another example is the use of spatial priors to estimate the connectivity matrix in GCM (Valdes-Sosa, 2004). Finally, when choosing a State Space model, it is useful to appreciate that there are two agendas when trying to understanding the connectivity of complex systems:

1. A data driven exploratory (discovery) approach that tries to scan the largest model space possible, identifying robust phenomena or candidates that will serve as constraints for more detailed modeling. This type of approach generally uses nonparametric or simply parameterized models for knowledge discovery. Prior knowledge is generally nonspecific (e.g., connections are sparse) but relatively non-restrictive.
2. A model driven confirmatory approach that is based on specific hypothesis driven models that incorporate as much biophysical prior knowledge as possible. Generally, the priors entail specific hypothesis about connectivity that can be resolved using model comparison.

These two approaches are shown in Fig. 2 (modified from Valdés-Sosa et al., 1999). In both cases, modeling is constrained by the data, by biophysical plausibility and ultimately the ability to establish links with computational models (hypotheses) of information processing

in the brain. Table 3 shows that at one extreme the model-driven approach is epitomized by Generic Nonparametric Models. Here, modeling efforts are constrained by data and the attempt to disclose emergent behavior, attractors and bifurcations (Breakspear et al., 2006) that can be checked against biophysically motivated models. An example of this approach is searching the complete brain times brain connectivity space (Fig. 3) with GCM mapping (Valdes-Sosa, 2004; Roebroek et al., 2005). At the other end we have the parametric and biophysically informed approach that DCM has emphasized (Chen et al., 2008). Having said this, as evidenced by this paper and companion papers, there is convergence of the two approaches, with a gradual blurring of the boundaries between DCM and GCM.

### Model inversions and inference

In this section, we look at the problem of model identification or inversion; namely, estimating the states and parameters of a particular model. It can be confusing when there is discussion of a new model that claims to be different from previous models, when it is actually the same model but with a different inversion or estimation scheme. We will try to clarify the distinction between models and highlight their points of contact when possible. Our main focus here will be on different formulations of SSM and how these formulations affect model inversion.

#### Discrete or continuous time?

One (almost) always works with discretely sampled data. When the model is itself discrete, then the only issue is matching the sampling times of the model predictions and the data predicted. However, when starting from a continuous time model, one has to model explicitly the mapping to discrete time.

Mapping continuous time predictions to discrete samples is a well-known topic in engineering and (probably from the early 50s) has been solved by linearization of the ODEs and integration over discrete time steps; a method known as the Exponential Euler method for reasons we shall see below: see Minchev and Wright (2005) for a historical review. For a recent review, with pointers to engineering toolboxes, see Garnier and Wang (2008).

One of the most exciting developments in the 60s, in econometrics was the development of explicit methods for estimating continuous models from sampled data, initiated by Bergstrom (1966).<sup>5</sup> His idea was essentially the following. Consider 3 time series  $X_1(t)$ ,  $X_2(t)$ , and  $X_3(t)$  where we know the values at time  $t$ :

$$\begin{pmatrix} dX_1(t) \\ dX_2(t) \\ dX_3(t) \end{pmatrix} = A \begin{pmatrix} X_1(t) \\ X_2(t) \\ X_3(t) \end{pmatrix} dt + \sum^{1/2} dB(t). \quad (3)$$

Then the explicit integration<sup>6</sup> over the interval  $[t + \Delta t, t]$  is

$$\begin{pmatrix} X_1(t + \Delta t) \\ X_2(t + \Delta t) \\ X_3(t + \Delta t) \end{pmatrix} = \exp(A\Delta t) \begin{pmatrix} X_1(t) \\ X_2(t) \\ X_3(t) \end{pmatrix} + e(t + \Delta t) \quad (4)$$

$$e(t + \Delta t) = \int_0^{\Delta t} \exp(sA) \sum^{1/2} dB(t-s)$$

$$\Sigma_{discrete} = \int_0^{\Delta t} \exp(sA) \Sigma \exp(sA^T) ds$$

$$e(t + \Delta t) \sim \mathcal{N}(0, \Sigma_{discrete}).$$

<sup>5</sup> Who, in fact, did this not for SDE (ODE driven by Brownian noise) but for linear ODE driven by random measures, as reviewed in Bergstrom (1984).

<sup>6</sup> Note, once again, that we use the convention  $[t + \Delta t, t]$  for the time interval that goes from  $t$  in the past to  $t + \Delta t$  in the present; while not the conventional usage this will make later notation clearer.

The noise of the discrete process now has the covariance matrix  $\Sigma_{discrete}$ . It is immediately evident from the equation above that the lag zero covariance matrix  $\Sigma_{discrete}$  will show contemporaneous covariance even if the continuous covariance matrix  $\Sigma$  is diagonal. In other words, the discrete noise becomes correlated over the three time-series (e.g., channels). This is because the random fluctuations ‘persist’ through their influence on the motion of the states. Rather than considering this a disadvantage Bergstrom (1984) and Phillips (1974) initiated a line of work studying the estimation of continuous time Autoregressive models (Mccrorie and Chambers, 2006), and continuous time Autoregressive Moving Average Models (Chambers and Thornton, 2009). This approach tries to use both lag information (the AR part) and zero-lag covariance information to identify the underlying linear model.

The extension of the above methods to nonlinear stochastic systems was proposed by Ozaki (1992) and has been extensively developed in recent years, as reviewed in Valdes-Sosa et al. (2009a). Consider a nonlinear system of the form:

$$dX(t) = f(X(t))dt + \sum^{1/2} dB(t)$$

$$X(t) = \begin{pmatrix} X_1(t) \\ X_2(t) \\ X_3(t) \end{pmatrix}. \quad (5)$$

The essential assumption in local linearization (LL) of this nonlinear system is to consider the Jacobian matrix  $A = \partial f / \partial X$  as constant over the time period,  $[t + \Delta t, t]$ . This Jacobian plays the same role as the matrix of autoregression coefficient in the linear systems above. Integration over this interval follows as above, with the solution:

$$X(t + \Delta t) = X(t) + A^{-1}(\exp(A\Delta t) - I)f(X(t)) + e(t + \Delta t)^7 \quad (6)$$

where  $I$  is the identity matrix. This solution is locally linear but crucially it changes with the state at the beginning of each integration interval; this is how it accommodates nonlinearity (i.e., a state-dependent autoregression matrix). As above, the discretised noise shows instantaneous correlations. Examples of inverting nonlinear continuous time neural models using this procedure are described in Valdes-Sosa et al. (1999), Riera et al. (2007b), Friston and Daunizeau (2008), Marreiros et al. (2009), Stephan et al. (2008), and Daunizeau et al. (2009b). Local linearization of this sort is used in all DCMs, including those formulated in generalized coordinates of motion.

There are several well-known technical issues regarding continuous model inversion:

1. The econometrics literature has been very much concerned with identifiability in continuous time models—an issue raised by one of us in the C&C series (Friston, 2009b) due to the non-uniqueness of the inverse mapping of the matrix exponential operator (matrix logarithm) for large sampling periods  $\Delta t$ . This is not a problem for DCM, which parameterizes the state-equation directly in terms of the connectivity  $A$ . However, autoregressive models (AR) try to estimate  $A = \exp(A\Delta t)$  directly, which requires a mapping  $A = \frac{1}{\Delta t} \ln(A)$  to get back to the underlying connectivity. Phillips noted in the 70s that  $A$  is not necessarily invertible, unless one is sampling at twice the highest frequency of the underlying signal (the Nyquist frequency) (Phillips, 1973); in other words, unless one samples quickly, in relation to the fluctuations in hidden states. In econometrics, there are several papers that study the conditions in which under-sampled systems can avoid an implicit aliasing problem (Hansen and Sargent, 1983; Mccrorie and Chambers,

<sup>7</sup> Note integration should not be computed this way since it is numerically unstable, especially when the Jacobian is poorly conditioned. A list of robust and fast procedures is reviewed in Valdes-Sosa et al. (2009a).

2006; Mccrorie, 2003). This is not a problem for electrophysiological models because sampling is fast relative to the underlying neuronal dynamics. However, for fMRI this is not the case and AR models provide connectivity estimates,  $A = \frac{1}{\Delta t} \ln(A) \in \mathbb{C}^{N \times N}$  that are not necessarily unique (a phenomenon known as “aliasing” as discussed below). We will return to this problem in the next section, when considering the mediation of local (direct) and global (indirect) influences over time. Although this “missing time” problem precludes inference about coupling between neuronal states that fluctuate quickly in relation to hemodynamics, one can use AR models to make inferences about slow neuronal fluctuations based on fMRI (e.g., the amplitude modulation of certain frequencies; see Fig. 4). Optimal sampling for AR models has been studied extensively in the engineering literature—the essential point being that sampling should not be below or even much above the optimal choice that matches the natural frequencies (time-constants) of the hidden states (Astrom, 1969; Larsson et al., 2006).

- When the sampling period  $\Delta t$  is sufficiently small, the AR model is approximately true. What is small? We found very few practical recommendations, with the exception of Sargan (1974), who uses heuristic arguments and Taylor expansions to suggest that a sampling frequency 1.5 times faster than the Nyquist frequency allows the use of a bilinear (or Tustin) approximation in (two stage non-recursive) autoregression procedures. As shown in the references cited above, it might be necessary to sample at several times the Nyquist frequency to use AR models directly. However, an interesting “Catch 22” emerges for AR models: The aliasing problem mandates fast sampling, but fast sampling violates Markovian (e.g., Gaussian noise) assumptions, if the true innovations are real (analytic) fluctuations.
- A different (and a more complicated) issue concerns the identifiability of models of neural activity actually occurring at rates much higher than the sampling rates of fMRI, even when a DCM is parameterized in terms of neuronal coupling. This is an inverse problem that depends on prior assumptions. There are lessons to be learned from the EEG literature here: Linear deconvolution methods for inferring neural activity from EEG proposed by Glover (1999) and Valdes-Sosa et al. (2009a) correspond to a temporal version of the minimum norm and LORETA spatial inverse solutions respectively. Riera et al. (2007a) and Riera et al. (2006), proposed a nonlinear deconvolution method. In fact, every standard SPM analysis of fMRI data is effectively a deconvolution, where the stimulus function (that is convolved with an assumed HRF) provides a generative model whose inversion corresponds to deconvolution. In the present context, the stimulus function provides the prior expectations about neuronal activity and the assumed HRF places priors on the ensuing hemodynamics. In short, model inversion or deconvolution depends on priors. The extent to which identifiability will limit inferences about neuronal coupling rests on whether the data supports evidence for different models of neuronal activity. We already know that there is sufficient information in fMRI time series to resolve DCMs with different neuronal connectivity architectures (through Bayesian model comparison), provided we use simple bilinear models. The issue here is whether we can make these models more realistic (cf., the neural mass models used for EEG) and still adjudicate among them, using model evidences: When models are too complex for their data, their evidence falls and model selection (identification) fails. This is an unresolved issue.

As one can see from these points, the issue of inference from discretised data depends on the fundamental frequencies of fluctuations in hidden states, data sampling rate, the model, and the prior information we bring to the inferential problem. When writing these lines, we were reminded of the dictum, prevalent in the first years of

EEG source modeling, that one could “only estimate a number of dipoles that was less than or equal to a sixth of the number of electrodes”. Bayesian modeling has not increased the amount of information in data but it has given us a principled framework to optimize generative or forward models (i.e., priors) in terms of their complexity, by choosing priors that maximize model evidence. This has enabled advances in distributed source modeling and the elaboration of better constraints (Valdés-Sosa et al., 2009b). One might anticipate the same advances in causal modeling over the next few years.

#### *Time, frequency or generalized coordinates?*

A last point to mention is that (prior to model inversion) it may be convenient to transform the time domain data to a different coordinate system, to facilitate computations or achieve a theoretical objective. In particular transformation to the frequency domain has proved quite useful.

- This was proposed first for generic linear models in both continuous and discrete time (Robinson, 1991). More recently a nonparametric frequency domain approach has been proposed for Granger Causality (Dhamala et al., 2008).
- A recent stream of EEG/MEG effective connectivity modeling has been introduced by Nolte et al. (2008), Marzetti et al. (2008), Nolte et al. (2009), and Nolte et al. (2006) with the realization that time (phase) delays are reflected in the imaginary part of the EEG/MEG cross-spectra, whereas the real part contains contemporaneous contributions due to volume conduction.
- Linearised versions of nonlinear DCMs have also been transformed successfully to the frequency domain (Moran et al., 2008; Robinson et al., 2008).

As noted above Friston (2008a,b) has proposed a transformation to generalized coordinates, inspired by their success in physics. This involves representing the motion of the system by means of an infinite sequence of derivatives. The truncation of this sequence provides a summary of the time-series, in much the same way that a Fourier transform provides a series of Fourier coefficients. In classical time series analysis, the truncation is based on frequencies of interest. In generalized coordinates, the truncation is based on the smoothness of the time series. This use of generalized coordinates in causal modeling is predicated on the assumption that real stochastic processes are analytic (Belyaev, 1959).

#### *Model inversion and inference*

There are many inversion schemes to estimate the states, parameters and hyperparameters of a model. Some of the most commonly used are variants of the Kalman Filter, Monte-Carlo methods and variational methods (see e.g., Daunizeau et al., 2009b for a variational Bayesian scheme). As reviewed in Valdes-Sosa et al. (2009a) the main challenges are how to scale the numerics of these schemes for more realistic and extensive modeling. The one thing all these schemes have in common is that they (implicitly or explicitly) optimize model parameters with respect to model evidence. In this sense model inversion and inference on models *per se* share a common objective; namely to maximize the evidence for a model.

Selecting or optimizing a model for effective connectivity ultimately rests on model evidence used in model comparison or averaging. The familiar tests for GCM (i.e. Dickey–Fuller test) are based on likelihood comparisons. As noted above, the likelihood (the probability of the data given a model and its parameters) is the same as model evidence (the probability of the data given a model), if we ignore uncertainty about the model parameters. However, the models considered in this paper, that include qualitative prior beliefs call for measures of goodness that balance accuracy (expected log-likelihood)



with model complexity. All of these measures (AIC, BIC, GCV, and variational free energy) are approximations to the model evidence (Friston, 2008a,b). Model evidence furnishes the measure used for the final probabilistic inference about a causal architecture (i.e., causal inference). Clearly, to carry out model comparison one must have an adequate set of candidates. Model Diagnostics are useful heuristics in this context that ensure that the correct models have been chosen for comparison. An interesting example that can be used to perform a detailed check of the adequacy of models is to assess the spatial and temporal whiteness of the residual innovation of the model which is illustrated in (Galka et al., 2004). More generally, the specification and exploration of model sets (spaces) probably represents one of the greatest challenges that lie ahead in this area.

### Summary

In summary, we have reviewed the distinction between autoregression (AR) models and models formulated in continuous time (DCM). We have touched upon the important role of local linearisation in mapping from continuous dynamics of hidden states to discrete data samples and the implications for sampling under AR models. In terms of model inversion and selection, we have highlighted the underlying role played by model evidence and have cast most of the cores issues in model identifiability and selection in terms of Bayesian model comparison. This subsumes questions about the complexity of models that can be supported by fMRI data; through to ultimate inferences about causality, in terms of which causal model has the greatest evidence. This section concludes our review of pragmatic issues and advances in the causal modeling of effective connectivity. We now turn to more conceptual issues and try to link the causal modeling for Neuroimaging described in this section to classical constructs that have dominated the theoretical literature over the past few decades.

### Statistical causal modeling

In this section, we review some key approaches to statistical causality. At one level, these approaches have had relatively little impact on recent developments in causal modeling in Neuroimaging, largely because they based on classical Markovian (and linear) models or ignore dynamics completely. However, this field contains some deep ideas and we include this section in the hope that it will illuminate some of the outstanding problems we face when modeling brain connectivity. Furthermore, it may be the case that bringing together classical temporal precedence treatments with structural causal modeling will finesse these problems and inspire theoreticians to tackle the special issues that attend the analysis of biological time series.

### Philosophical background

Defining, discovering and exploiting causal relations have a long and enduring history (Bunge, 2009). Examples of current philosophical debates about causality can be found in Woodward (Woodward, 2003) and Cartwright (2007). An important concept, stressed by Woodward, is that a cause is something that “makes things happen”. Cartwright, on the other hand (Cartwright, 2007), argues for the need to separate the definition, discovery and use of causes; stresses the pluralism of the concept of cause and argues for the use of “thick causal concepts”. An example of what she calls a “thin causal claim” would be that “activity in the retina causes activity in V1”—represented as a directed arrow from one structure to the other. Instead, it might be more useful to say that the Retina is mapped via a complex logarithmic transform to V1 (Schwartz, 1977). A “thick causal” explanation tries to explain how information is actually transmitted. For a different perspective see Glymour (2009). It may be that both thin and thick causal concepts are useful when characterizing complex systems.

Despite philosophical disagreements about the study of causality, there seems to be a consensus that causal modeling is a legitimate statistical enterprise (Cox and Wermuth, 2004; Frosini, 2006; Pearl, 2003). One can clearly differentiate two current streams of statistical causal modeling; one based on Bayesian dependency graphs or graphical models which has been labeled as “Structural Causal Modeling” by White and Lu (2010). The other, apparently unrelated, approach rests on some variant of Granger Causality for which we prefer the terms WAGS influence<sup>8</sup> for reasons stated below. WAGS influence modeling appeals to an improved predictability of one time series by another. We will describe these two streams of modeling, which leads us to anticipate their combination in a third line of work, called Dynamic Structural Systems (White and Lu, 2010).

### Structural causal modeling: graphical models and Bayes-Nets

Structural Causal Modeling originated with Structural Equation Modeling (SEM) (Wright, 1921) and is characterized by the use of graphical models, in which direct causal links are encoded by directed edges in the graph (Lauritzen, 1996; Pearl, 2000; Spirtes et al., 2000). Ideally these edges can be given a mechanistic interpretation (Machamer et al., 2000). Using these graphs, statistical procedures then discover the best model (graph) given the data (Pearl, 2000; 2003; Spirtes et al., 2000). As explained in the previous section, the “best” model has the highest evidence. There may be many models with the same evidence; in this case, the statistical search produces an equivalence class of models with the same explanatory power. With regard to effective connectivity, the multiplicity of possibly equivalent models has been highlighted by Ramsey et al. (2010).

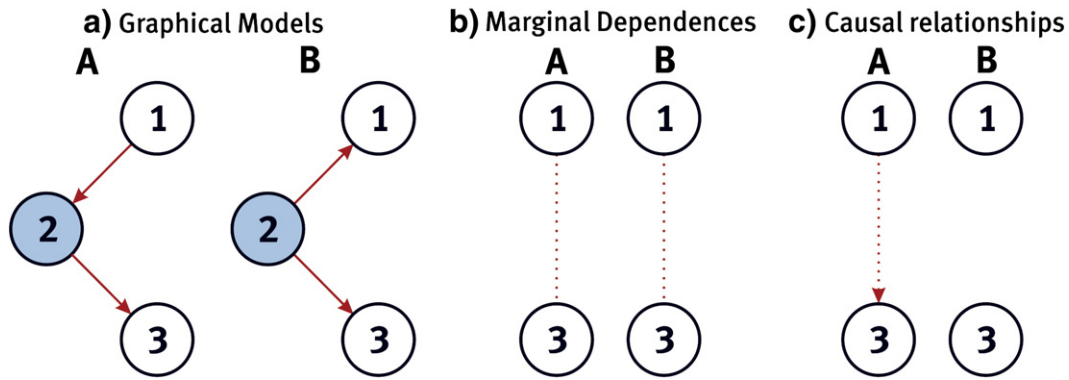
This line of work has furnished Statistical Causal Modeling with a rigorous foundation and specific graphical procedures such as the “Back-door” and “Front-door” criteria, to decide whether a given causal model explains observational data. Here, causal architectures are encoded by the structure of the graph. In fMRI studies these methods have been applied by Ramsey et al. (2010) to estimate directionality in several steps, first looking for “unshielded colliders” (paths of the form  $A \rightarrow B \leftarrow C$ ) and then finding out what further dependencies are implied by these colliders. We now summarize Structural Causal Modeling, as presented by Pearl (2000).

One of the key concepts in Pearl’s causal calculus is *interventional probabilities*, which he denotes  $p(x_i | do(X_i = x_i))$  or more simply  $p(x_i | do(x_i))$ , which are distinct from conditional probabilities  $p(x_i | X_i = x_i)$ . Pearl highlights the difference between the *action*  $do(X_i = x_i)$  and the *observation*  $X_i = x_i$ . Note that observing  $X_i = x_i$  provides information both about the children and parents of  $X_i$  in a directed acyclic graph (DAG<sup>9</sup>). However, whatever relationship existed between  $X_i$  and its parents prior to action, this relationship is no longer in effect when we perform the action  $do(X_i = x_i)$ .  $X_i$  is held fixed by the action  $do(X_i = x_i)$ , and therefore cannot be influenced. Thus, inferences based on evaluating  $do(X_i = x_i)$  are different in nature from the usual conditional inference. Interventional probabilities are calculated via a *truncated factorization*; i.e. by conditioning on a “mutilated graph”, with the edges (links) from the parents of  $X_i$  removed:

$$p(x_i | do(x_i)) = \prod_{j \neq i} p(x_j | pa_j) = \frac{p(x)}{p(x_i | pa_i)}. \quad (7)$$

<sup>8</sup> It might be preferable to use a more precise term “predictability” instead of influence.

<sup>9</sup> DAG = Directed Acyclic Graph. The word ‘graph’ refers to the mapping from the set of (factorized) joint probability densities over  $X$  and the actual directed acyclic graph that represents the set of conditional independencies implicit in the factorization of the joint pdf  $p(x)$ .



**Fig. 5.** The missing region problem. a) Two typical graphical models including a hidden node (node 2). b) Marginal dependence relationships implied by the causal structure depicted in (a), after marginalizing over the hidden node 2; the same moral graph can be derived from directed (causal) graphs A and B. c) Causal relationships implied by the causal structure depicted in (a), after marginalizing over the hidden node 2. Note that these are perfectly consistent with the moral graph in (b), depicting (non causal) statistical dependencies between nodes 1 and 3, which are the same for both A and B.

Here,  $pa_j$  denotes the set of all the parents of the  $j$ th node in the graph and  $p(x)$  is the full joint distribution. Such interventional probabilities exhibit two properties:

$$\begin{cases} P1 : p(x_i | do(pa_i)) = p(x_i | pa_i) \\ P2 : p(x_i | do(pa_i), do(s)) = p(x_i | do(pa_i)) \end{cases} \quad (8)$$

for all  $i$  and for every subset  $S$  of variables disjoint of  $\{X_i, PA_i\}$ . Property 1 renders every parent set  $PA_i$  exogenous relative to its child  $X_i$ , ensuring that the conditional  $p(x_i | pa_i)$  probability coincides with the effect (on  $X_i$ ) of setting  $PA_i$  to  $pa_i$  by external control. Property 2 expresses the notion of invariance: once we control its direct causes  $PA_i$ , no other interventions will affect the probability of  $X_i$ . These properties allow us to evaluate the (probabilistic) effect of interventions from the definition of the joint density  $p(x)$  associated with the pre-intervention graph.

This treatment of interventions provides a semantics for notions such as “causal effects” or “causal influence”. For example, to see whether a variable  $X_i$  has a causal influence on  $X_j$ , we compute (using the truncated factorization in Eq. (7)) the marginal distribution of  $X_j$  under the actions  $do(X_i = x_i)$  and check whether that distribution is sensitive to  $x_i$ . It is easy to see that only descendants of  $X_i$  can be influenced by  $X_i$ ; deleting the factor  $p(x_i | pa_i)$  from the joint distribution turns  $X_i$  into a root node<sup>10</sup> in the mutilated graph. This can be contrasted with (undirected) probabilistic dependencies that can be deduced from the factorization of the joint distribution *per se*. These dependencies can be thought as (non-causal and non-directed) correlations among measured variables that can be predicted on the basis of the structure of the network.

In the context of brain connectivity, the measures of interventional and conditional probabilities map onto the notions of effective connectivity and functional connectivity respectively. Let us consider two typical situations that arise in the context of the missing region problem. These are summarized in Fig. 5.

Consider Fig. 5a. In situation A, node 1 influences node 2, which influences node 3. That is, the causal effect of 1 on 3 is mediated by 2. The joint distribution of the graphical causal model can be factorized as  $p_A(x) = p(x_3 | x_2) p(x_2 | x_1) p(x_1)$ . In situation B, both 1 and 3 have a common cause: node 2 influences both 1 and 3. The joint distribution of this graphical causal model can then be factorized as:  $p_B(x) = p(x_1 | x_2) p(x_3 | x_2) p(x_2)$ . It is easy to prove that in both cases (A and B), 1 and 3 are conditionally independent given 2; i.e.,  $p(x_1, x_3 | x_2) = p(x_1 | x_2) p(x_3 | x_2)$ . This means that observing node 1 (respectively 3) does not convey additional information about 3 (respectively 1), once we know

<sup>10</sup> A root node is a node without parents. It is marginally independent of all other variables in a DAG, except its descendants.

2. Furthermore, note that 1 and 3 are actually *marginally dependent*; i.e.,  $p(x_1, x_3) = \int p(x) dx_2 \neq p(x_1) p(x_3)$ . This means that whatever value  $X_2$  might take,  $X_1$  and  $X_3$  will be correlated. Deriving the marginal independencies from the DAG produces an undirected graph (see, e.g., Fig. 5b). This undirected graph is called a *moral graph* and its derivation is called the *moralization* of the DAG. For example, moralizing the DAG A produces a fully connected moral graph.

In brief, both situations (A and B) are similar in terms of their statistical dependencies. In both situations, functional connectivity methods would recover the conditional independence of nodes 1 and 3 if node 2 was observed, and their marginal dependence if it is not (see Fig. 5b). However, the situations in A and B are actually very different in terms of the causal relations between 1 and 3. This can be seen using the interventional probabilities defined above: let us derive the interventional probabilities expressing the causal influence of node 1 onto node 3 (and reciprocally) in situation A:

$$\begin{aligned} p_A(x_3 | do(\tilde{x}_1)) &= \int p_A(x_2, x_3 | do(\tilde{x}_1)) dx_2 \\ &= \int p(x_3 | x_2) p(x_2 | \tilde{x}_1) dx_2 \\ &= p(x_3 | \tilde{x}_1) \end{aligned} \quad (9)$$

$$\begin{aligned} p_A(x_1 | do(\tilde{x}_3)) &= \int p_A(x_1, x_2 | do(\tilde{x}_3)) dx_2 \\ &= p(x_1) \int p(x_2 | x_1) dx_2 \\ &= p(x_1). \end{aligned} \quad (10)$$

Eq. (8) simply says that the likelihood of any value that  $x_3$  might take is dependent upon the value  $\tilde{x}_1$  that we have fixed for  $x_1$  (by intervention). In contradistinction, Eq. (9) says that the likelihood of any value that  $x_1$  might take is independent of  $x_3$ . This means that node 1 has a causal influence on node 3, i.e. there is a directed (mediated through 2) causal link from 1 to 3. The situation is quite different in B:

$$\begin{aligned} p_B(x_3 | do(\tilde{x}_1)) &= \int p_B(x_2, x_3 | do(\tilde{x}_1)) dx_2 \\ &= \int p(x_3 | x_2) p(x_2) dx_2 \\ &= p(x_3) \\ p_B(x_1 | do(\tilde{x}_3)) &= \int p_B(x_1, x_2 | do(\tilde{x}_3)) dx_2 \\ &= \int p(x_1 | x_2) p(x_2) dx_2 \\ &= p(x_1). \end{aligned} \quad (11)$$

This shows that nodes 1 and 3 are not influenced by intervention on the other. This means that here, there is no causal link between 1 and 3. This is summarized in Fig. 5c, which depicts the corresponding ‘effective’ causal graphs, having marginalized over node 2.

Causal calculus provides a simple but principled perspective on the “missing region” problem. It shows that effective connectivity analysis

can, in certain cases, address a subset of brain regions (subgraph), leaving aside potential variables (e.g., brain regions) that might influence the system of interest. The example above makes the precise confines of this statement clear: one must be able to perform interventional actions on source and target variables. Given that the principal ‘value-setting’ interventions available to us in cognitive neuroscience are experimental stimulus manipulations, our capacity for such interventions are generally limited to the primary sensory cortices. Intervention beyond sensorimotor cortex is much more difficult; although one could employ techniques such as transcranial magnetic stimulation (TMS) to perturb activity in superficial cortical areas. However, the perturbation in TMS is unnatural and known to induce compensatory changes throughout the brain rather than well-defined effects in down-stream areas.

The same undirected graph can be derived from the moralization of a set of DAGs (c.f. from Figs. 5a and b). This set contains a (potentially infinite) number of elements, and is referred to as the *equivalent class*. As stated by Pearl, the identification of causal (i.e., interventional) probabilities from observational data requires additional assumptions or constraints (see also Ramsey et al., 2010). Pearl mentions two such critical assumptions: (i) *minimality* and (ii) *structural stability*. Minimality appeals to complexity minimization, when maximizing model evidence (c.f., Occam's razor). In brief, among a set of causal models that would explain the observed data, one must choose the simplest (e.g., the one with the fewest parameters). Structural stability (also coined ‘faithfulness’) is a related requirement that is motivated from the fact that an absence of causal relationships is inferred from an observed absence of correlation. Therefore, if no association is observed, it is unlikely to be due to the particular instantiation of a given model for which this independence would be predicted (see below). Rather, it is more likely to be explained in terms of a model that would predict, for any parameter setting, the observed absence of correlation. This clearly speaks to the convergent application, mentioned above, of data driven exploratory approaches that scan the largest model space possible for correlations to be explained and a model driven confirmatory approach that appeal to structural stability: Within a Bayesian setting, we usually specify a prior distribution  $p(\theta|m)$  over model parameters, which are usually assumed to be independent. This is justified when the parameters represent mechanisms that are free to change independently of one another—that is, when the system is structurally stable. In other terms, the use of such prior favors structurally stable models. In most cases, stability and minimality are sufficient conditions for solving the structure discovery inverse problem in the context of observational data. If this is not sufficient to reduce the cardinality of the equivalent class, one has to resort to experimental interventions.<sup>11</sup> Within the context of Neuroimaging, this would involve controlling the system by optimizing the experimental design in terms of the psychophysical properties of the stimuli and/or through direct biophysical stimulation (e.g., transcranial magnetic stimulation – TMS – or deep brain stimulation—DBS).

### Summary

The causal calculus based on graphical models has some important connections to the distinction between functional and effective connectivity and provides an elegant framework in which one can deal with interventions. However, it is limited in two respects. First, it is restricted to discovering conditional independencies in *directed acyclic graphs*. This is a problem because the brain is a directed *cyclic* graph—every brain region is reciprocally connected (at least polysynaptically) and every computational theory of brain function rests

on some form of reciprocal or reentrant message passing. Second, the calculus ignores time: Pearl argues that what he calls a ‘causal model’ should rest upon *functional relationships* between variables, an example of which is structural equation modeling (SEM). However, these functional relationships cannot deal with (cyclic) feedback loops. In fact, DCM was invented to address these limitations, after evaluating structural causal modeling for fMRI time-series. This is why it was called *dynamic* causal modeling to distinguish it from *structural* causal modeling (Friston et al., 2003). Indeed, Pearl (2000) argues in favor of dynamic causal models, when attempting to identify what physicists call hysteresis effects, whereby the causal influence depends upon the history of the system. Interestingly, the DAG limitation can be finessed by considering dynamics and temporal precedence within structural causal modeling. This is because the arrow of time turns directed cyclic graphs into directed acyclic graphs, when the nodes are deployed over successive time points. This leads us to an examination of prediction-based measures of functional relations.

### WAGS influence

The second stream of statistical causal modeling is based on the premise that a cause must precede and increase the predictability of its consequence. This type of reasoning can be traced back at least to Hume (Triacca, 2007) and is particularly popular in time series analysis. Formally, it was originally proposed (in an abstract form) by Wiener (1956) (see Appendix A) and introduced into data analysis by Granger (1963). Granger emphasized that increased predictability is a necessary but not sufficient condition for a causal relation to exist. In fact, Granger distinguished between true causal relations and “prima facie” causal relations (Granger, 1988); the former only to be inferred in the presence of “knowledge of the state of the whole universe”. When discussing “prima facie causes” we recommend the use of the neutral term “influence” in agreement with other authors (Commenges & Gégout-Petit, 2009; Gégout-Petit & Commenges, 2010). Additionally, it should be pointed out that around the same time as Granger's work, Akaike (1968), and Schweder (1970) introduced similar concepts of influence, prompting us to refer to “WAGS influence modeling” (for Wiener–Akaike–Granger–Schweder). This is a generalization of a proposal by Aalen (1987) and Aalen and Frigessi (2007) who were among the first to point out the connections between the Granger and Shweder concepts.

An unfortunate misconception in Neuroimaging identifies WAGS influence modeling (WAGS for short) with just one of the specific proposals (among others) dealt with by Granger; namely, the discrete-time linear Vector Autoregressive Model (VAR). This simple model has proven to be a useful tool in many fields, including Neuroimaging—the latter work well documented in Bressler and Seth (2010). However, this restricted viewpoint overlooks the fact that WAGS has dealt with a much broader class of systems:

1. Classical textbooks, such as Lutkepohl (2005), show how WAGS can applied VAR models, infinite order VAR, impulse response functions, Vector Autoregressive Moving Average models (VARMA), etc.
2. There are a number of nonlinear WAGS methods that have been proposed for analyzing directed effective connectivity (Freiwald et al., 1999, Solo, 2008; Gouieroux et al., 1987; Marinazzo et al., 2011; Kalitzin et al., 2007)
3. Early in the econometrics literature, causal modeling was extended to linear and nonlinear random differential equations in continuous time (Bergstrom, 1988). These initial efforts have been successively generalized (Aalen, 1987; Commenges & Gégout-Petit, 2009; Comte & Renault, 1996; Florens & Fougere, 1996; Gill & Petrović, 1987; Gégout-Petit & Commenges, 2010; Mykland, 1986; Petrović & Stanojević, 2010) to more inclusive types of dynamical systems.
4. Schweder (1970) describes WAGS concepts for counting processes in continuous, time which has enjoyed applications in Survival

<sup>11</sup> For example, the back- and front-door criteria (Pearl, 2000) can be used to optimize the intervention.

Analysis—a formalism that could well be used to model interactions expressed in neural spike train data.

We now give an intuitive explanation of some of these definitions (the interested reader can refer to the technical literature for more rigorous treatments). Let us again consider triples of (possibly vector) time series  $X_1(t), X_2(t), X_3(t)$ , where we want to know if time series  $X_1(t)$  is influenced by time series  $X_2(t)$  conditional on  $X_3(t)$ . This last variable can be considered as any time series to be controlled for (if we were omniscient, the “entire universe”!). Let  $X[a, b] = \{X(t) | t \in [a, b]\}$  denote the history of a time series in the discrete or continuous time interval  $[a, b]$ . There are several types of influence. One distinction is based on what part of the present or future of  $X_1(t)$  can be predicted by the past or present of  $X_2(\tau) \tau < t$ . This leads to the following classification:

- If  $X_2(\tau): \tau < t$ , can influence any future value of  $X_1(s)$  for  $s > t$ , then it is a global influence.
- If  $X_2(\tau) \tau < t$ , can influence  $X_1(t)$  it is a local influence.
- If  $X_2(\tau) \tau = t$  can influence  $X_1(t)$  it is a contemporaneous influence.

Another distinction is whether one predicts the whole probability distribution (**strong** influence) or only given moments (**weak** influence). These two classifications give rise to six types of influence as schematized in Fig. 6 and Table 4 and 5. Briefly, the formal definitions are as follows.

$X_1(t)$  is strongly, conditionally, and globally independent of  $X_2(t)$  given  $X_3(t)$  (not SCGi), if

$$P(X_1(\infty, t] | X_1(t, -\infty], X_2(t, -\infty], X_3(t, -\infty]) = P(X_1(\infty, t] | X_1(t, -\infty], X_3(t, -\infty]). \tag{12}$$

When this condition does not hold we say  $X_2(t)$  strongly, conditionally, and globally influences (SCGi)  $X_1(t)$  given  $X_3(t)$ . Note that the whole future of  $X_1$  is included (hence the term “global”). And the whole past of all time series is considered. This means these definitions accommodate non-Markovian processes (for Markovian processes, we only consider the previous time point). Furthermore, these definitions do not depend on an assumption of linearity or any given functional form (and are therefore applicable to any of the state equations in Table 2). Note also that this definition is appropriate for point processes, discrete and continuous time series, even for categorical (qualitative valued) time series. The only problem with this formulation is that it calls on the whole probability distribution and therefore its practical assessment requires the use of measures such as mutual information.

$X_1(t)$  is weakly, conditionally and globally independent of  $X_2(t)$  given  $X_3(t)$  (not WCGi), if

$$E[X_1(\infty, t] | X_1(\infty, t], X_2(t, -\infty], X_3(t, -\infty)] = E[X_1(\infty, t] | X_1(t, -\infty], X_3(t, -\infty)]. \tag{13}$$

If this condition does not hold we say  $X_2(t)$  weakly, conditionally and globally influences (WCGi)  $X_1(t)$  given  $X_3(t)$ . This concept extends to any number of moments (such as the variance of the process). There are a number of relations between these concepts: not SCGi implies not WCGi for all its moments and the converse is true for influences (WCGi implies SCGi), but we shall not go into details here; see Florens and Mouchart (1985), Florens (2003), Florens and Fougere (1996), and Florens and Mouchart (1982).

Global influence refers to influence at any time in the future. If we want to capture the idea of immediate influence we use the local

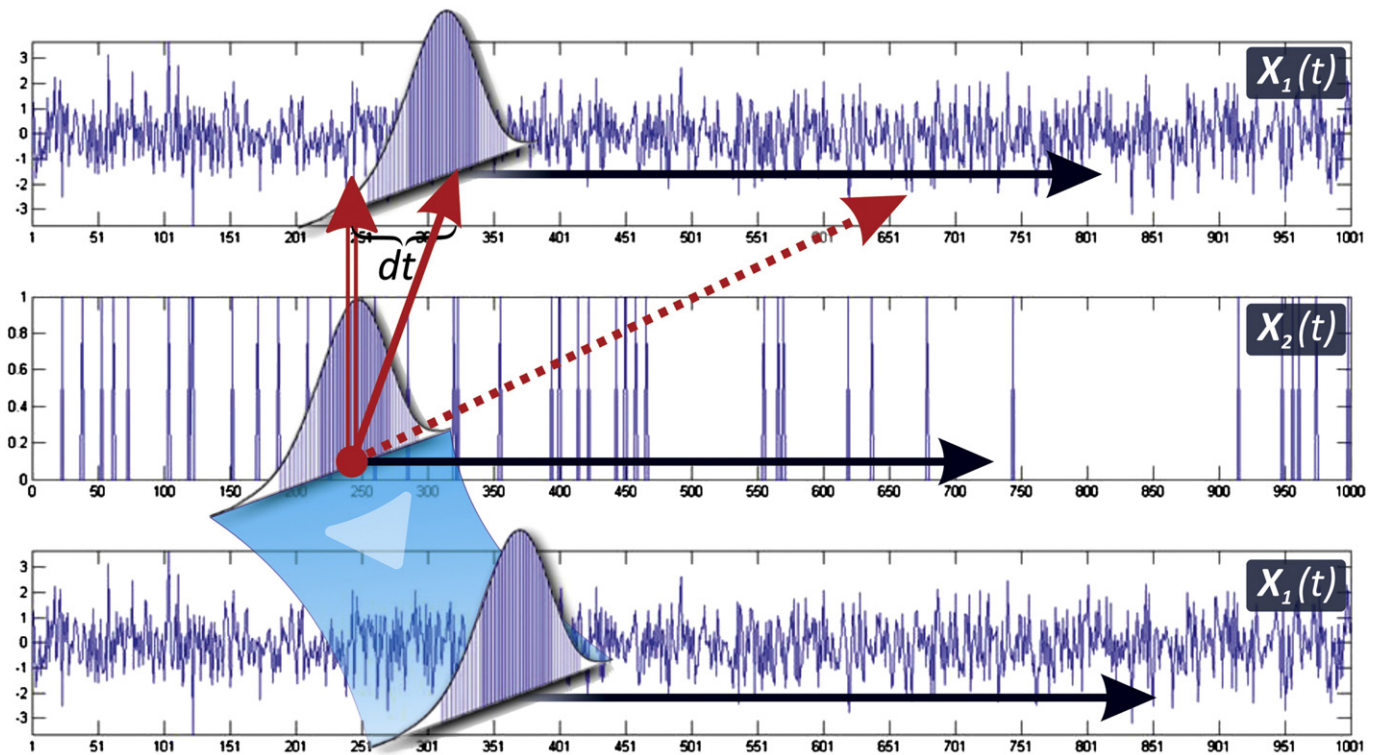


Fig. 6. Wiener-Akaike-Granger-Schweder (WAGS) Influences. This figure illustrates the different types of WAGS influence measures. In the middle  $X_2(t)$  a continuous time point process, which may be influencing the differentiable continuous time process  $X_1(t)$  (top and bottom). This process may have local influence (full arrows), which indicate predictability in the immediate future ( $dt$ ), or global influence (dashed arrow) at any set of future times. If predictability pertains to the whole probability distribution, this is a strong influence (bottom), and a weak influence (top) if predictability is limited to the moments (e.g., expectation) of this distribution.

**Table 4**  
Conditional Independence relations.

	Strong (Probability Distribution)	Weak (Expectation)
Global (for all horizons)	Strongly, Conditionally, Globally, independence (not SCGi)	Weakly, Conditionally, Globally, independence (not WCGi)
Local (Immediate future)	Strongly, Conditionally, Locally, independence (not SCLi)	Weakly, Conditionally, Globally, independence (not WCLi)
Contemporaneous	Strongly, Conditionally, Contemporaneously, independence (not SCCi)	Weakly, Conditionally, Contemporaneously, independence (not WCCi)

**Table 5**  
Types of Influence defined by absence of the corresponding independences in Table 4.

	Strong (Probability Distribution)	Weak (Expectation)
Global ( for all horizons)	Strongly, Conditionally, Globally, influence (SCGi) - Strong Granger or Sims influence	Weakly, Conditionally, Globally, influence (WCGi) - Weak Granger or Sims influence
Local (Immediate future)	Strongly, Conditionally, Locally, influence (SCLi) - Influence (Possibly indirect)	Weakly, Conditionally, Globally, influence (WCLi) - Direct Influence
Contemporaneous	Strongly, Conditionally, Contemporaneously, influence (SCCi)	Weakly, Conditionally, Contemporaneously, influence (WCCi)

concepts defined above. The concepts of strong and weak local influence have very simple interpretations if we are modeling in discrete time and events occur every  $\Delta t$ . To see this, consider the expectation based weak conditionally local independence (not WCLi) in discrete time:

$$E[X_1(t + \Delta t) | X_1[t, -\infty], X_2[t, -\infty], X_3[t, -\infty]] = E[X_1(t + \Delta t) | X_1[t, -\infty], X_3[t, -\infty]]. \tag{14}$$

If this condition does not hold we have that  $X_2(t)$  weakly, conditionally and locally influences (WCLi)  $X_1(t)$  given  $X_3(t)$ . Strong local concepts are defined similarly by considering conditional independences. For the usual discrete time, real valued time series of Neuroimaging, all these concepts are equivalent as shown by Florens and Mouchart (1982) and Solo (2007). As an example, consider the multivariate autoregressive model of the previous section

$$X(t + \Delta t) = \sum_{k=1}^p A_k X(t - (k-1)\Delta t) + e(t + \Delta t) \tag{15}$$

with the innovation term  $e_{t+\Delta t}$  being GWN with covariance matrix  $\Sigma := \Sigma_{discrete}$ . For this familiar case  $E[X[t + \Delta t] | X[t, -\infty]] = \sum_k A_k X(t - (k-1)\Delta t)$ , and analyzing influence reduces to finding which coefficients of the autoregressive coefficients are zero. However, in continuous time there is a problem when  $\Delta t \rightarrow 0$ , since the stochastic processes we are dealing with are at least almost surely continuous and  $\lim_{\Delta t \rightarrow 0} E[X_1(t + \Delta t) | X_1[t, -\infty], X_2[t, -\infty], X_3[t, -\infty]] = \lim_{\Delta t \rightarrow 0} E[X_1(t + \Delta t)]$  is trivially satisfied (limits are now taken in the sense of a quadratic mean) because the  $X_1(t)$  process is path continuous—it will only depend on itself. To accommodate this situation instead we shall use the following definition for not WCLi (Commenges & Gégout-Petit, 2009; Comte & Renault, 1996; Florens & Fougere, 1996; Gégout-Petit & Commenges, 2010; Renault, Sekkat, & Szafarz, 1998):

$$\lim_{\Delta t \rightarrow 0} E \left[ \frac{X_1(t + \Delta t) - X_1(t)}{\Delta t} \middle| X_1(t, -\infty), X_2(t, -\infty), X_3(t, -\infty) \right] = \lim_{\Delta t \rightarrow 0} E \left[ \frac{X_1(t + \Delta t) - X_1(t + \Delta t)}{\Delta t} \middle| X_1(t, -\infty), X_3(t, -\infty) \right]. \tag{16}$$

As noted by Renault et al. (1998) (whom we follow closely here), for finite  $\Delta t$  this is equivalent to the usual definitions. Now how does this definition relate to the linear SDE in Eq. (3)?

For three time series:

$$\begin{pmatrix} dX_1(t) \\ dX_2(t) \\ dX_3(t) \end{pmatrix} = A \begin{bmatrix} X_1(t) \\ X_2(t) \\ X_3(t) \end{bmatrix} dt + dB(t). \tag{17}$$

Integrating from  $t$  to  $\Delta t$ , we have

$$X_1(t + \Delta t) - X_1(t) = \int_t^{t+\Delta t} [a(1,1)X_1(\tau) + a(1,2)X_2(\tau) + a(1,3)X_3(\tau) + \sigma_{bb}(B_1(t + \Delta t), -B_2(t))] d\tau$$

$$\Rightarrow \lim_{\Delta t \rightarrow 0} E \left[ \frac{X_1(t + \Delta t) - X_1(t)}{\Delta t} \middle| X_1(t), X_2(t), X_3(t) \right] = a(1,1)X_2(t) + a(1,2)X_2(t) + a(1,3)X_3(t).$$

This shows that, in effect, the detection of an influence will depend on whether the coefficients of the matrix  $A$  are zero or not. For nonlinear systems this holds with the local linear approximation. This treatment highlights the goal of WAGS, like structural causal modeling, is to detect conditional independencies; in this (AR) example, weak and local.

The issue of contemporaneous influence measures is quite problematic. In discrete time, it is clear that the covariance matrix of two or more time series may have cross-covariances that are due to an “environmental” or missing variable  $Z(t)$ . This was discussed by Akaike and a nice example of this effect is described in Wong and Ozaki (2007), which also explains the relation of the Akaike measures of influence to others used in the literature. For continuous time (Comte and Renault, 1996) define strong (second order) conditional contemporaneous independence (not SCCi) if:

$$\text{cov}[X_1(\infty, t], X_2(\infty, t) | X_1[t, -\infty], X_2[t, -\infty], X_3[t, -\infty]] = 0. \tag{18}$$

Note that this is the same definition for continuous time as for the discrete AR example (Eq. (15)) and is equivalent to requiring that the elements of the corresponding innovation covariance matrix  $\Sigma$  be zero. These authors then went on to define weak contemporaneous conditional independence (not WCCi) if:

$$\lim_{\Delta t \rightarrow 0} \{ \text{cov}[X_1(t + \Delta t), X_2(t + \Delta t) | X_1[t, -\infty], X_2[t, -\infty], X_3[t, -\infty]] \} = 0. \tag{19}$$

In the absence of these conditions we have strong (weak) contemporaneous conditional influences which are clearly non-

directional. In his initial paper (Granger, 1963) defined a contemporaneous version of his influence measure in discrete time. Much later, (Geweke, 1984) decomposed his own WAGS measure into a sum of parts, some depending on lag information and others reflecting contemporaneous (undirected) influences, see in these C&C (Bressler and Seth, 2010). However, Granger (in later discussions) felt that if the system included all relevant time series this concept would not be valid, unless these influences were assigned a directionality (see Granger, 1988, pp. 204–208). In this sense, he was proposing a Structural Equation Modeling approach to the covariance structure of the autoregressive model innovations. As will be mentioned below (WAGS influence section) this is something that has been explored in the econometrics literature by Demiralp and Hoover (2008), Moneta and Spirtes (2006), but not to our knowledge in Neuroimaging.

#### More general models

As we have seen, strong global measures of independence are equivalent to conditional independence and are therefore applicable to very general stochastic processes. For weak local conditional independence, the situation is a little more difficult and we have given examples, which involve a limit in the mean of a derivative-type operator expression. The more general theory, too technical to include here, entails successive generalizations by Mykland (1986), Aalen (1987), Commenges and Gégout-Petit (2009), and Gégout-Petit and Commenges (2010). The basic concept can be stated briefly as follows (we drop conditioning on a third time series for convenience). Suppose we have stochastic processes that are semi-martingales of the form,  $X(t) = P^X(t) + M^X(t)$ . Here  $P^X(t)$  is a predictable stochastic process<sup>12</sup> of bounded variation, which is known as the “compensator” of the semi-martingale, and  $M_t^X$  is a martingale.<sup>13</sup> Predictability is the key property that generalizes Wiener’s intuition. The martingale component is the unpredictable part of the stochastic process we are interested in.<sup>14</sup> Now suppose we have two stochastic processes  $X(t)$  and  $W(t)$ . If:

1. The martingales  $M^{X_1}$  and  $M^{X_2}$  are orthogonal (no contemporaneous interactions).
2.  $P^{X_1}(t)$  is measurable<sup>15</sup> with respect to  $X_1[t, -\infty]$  only (without considering  $X_2(t)$ ).

then  $X_1(t)$  is said to be weakly locally independent of  $X_2(t)$ . In Gégout-Petit and Commenges (2010) the concept of  $\sim$ WLCI is generalized to a general class of random phenomena that include random measures, marked point processes, diffusions, and diffusions with jumps, covering many of the models in Table 2. In fact, this theory may allow unification of the analysis of random behavioral events, LFP, spike recordings, and EEG, just to give a few examples.

<sup>12</sup> Roughly speaking, if  $P^X(t)$  is a predictable process, then it is “known” just ahead of time  $t$ . For a rigorous definition and some discussion see <http://myyn.org/m/article/predictable-process/>.

<sup>13</sup> For a martingale  $M(t)$ ,  $E(M(t+s)|X[t, -\infty]) = M(t)$  for all  $t$  and  $s$ . This states that the expected value of  $M(t+s)$  is that at time  $t$ , there is no “knowledge” (in the sense of expected value) for the future from the past, hence this type of process is taken as a representation of unpredictability.

<sup>14</sup> This is a form of the famous Doob–Meyer decomposition of a stochastic process (Medvegyev, 2007).

<sup>15</sup> Roughly speaking  $P^X(t)$  is measurable with respect to the process  $X_1[t, -\infty]$  and not  $X_2[t, -\infty]$  if all expected values of  $P^X(t)$  can be obtained by integrating  $X_1[t, -\infty]$  without reference to  $X_2[t, -\infty]$ . The technical definition can be found in Medvegyev (2007). Basically this definition is based on the concept of a “measurable function” extended to the sets of random variables that comprise the stochastic processes.

#### Direct influence

Weak local independence might be considered an unnecessarily technical condition for declaring the absence of an influence; in that strong (local or global) influence measures should be sufficient. An early counterexample of this was provided by Renault et al. (1998), where they considered a model where  $X(t)$  is  $\sim$ WLCI of  $W(t)$ , given  $Z(t)$ . See Fig. 7 for an illustration of this divergence between local and global influences. This has led Commenges and Gégout-Petit (2009) to define WLCI as the central concept for “direct influence” whereas SCGI is an influence that can be mediated directly or indirectly through other time series.

An important point here is the degree to which the definition of WAGS influence depends on the martingale concept or, indeed, on that of a stochastic process. As discussed in The observation equation section, there are a number of instances in which Markovian models developed for financial time series may not apply for Neuroimaging data. However, the concepts are probably generally valid, as we shall illustrate with some examples:

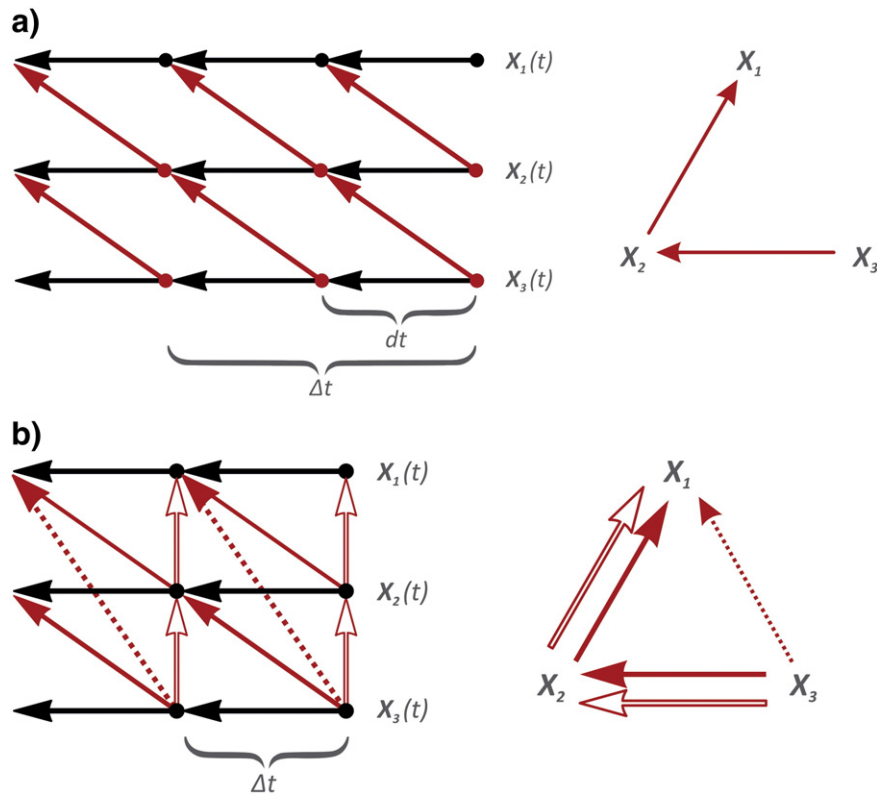
- The analytical random processes used in generalized coordinates are quite different from those usually studied in classical SDE theory but have been known for a long time (Belyaev, 1959). In fact, there has been quite a lot of work on their predictability (Lyman et al., 2000) and indeed there is even work on VARMA modeling of this type of process (Pollock, 2010).
- We have already seen that the definitions of influence do not depend on Markovian assumptions as noted by Aalen (1987).
- The use of deterministic bilinear systems in DCM (Penny et al., 2005) suggests that (non-stochastic) ODEs may be incorporated into the WAGS framework. This sort of assimilation has in fact been proposed by Commenges and Gégout-Petit (2009) as a limiting case of the definition based on semi-martingales above. Extensions of the definition might be required when dealing with chaotic dynamics but, even here, measure theoretic definitions are probably valid.<sup>16</sup> An interesting discussion of determinism versus stochastics can be found in Ozaki (1990).

The use or development of WAGS theory for systems that were not initially considered by the aforementioned papers may well be a fruitful area of mathematical research. In particular, WAGS may be especially powerful when applied to processes defined on continuous spatial manifolds (Valdes-Sosa, 2004; Valdés-Sosa et al., 2006). To our knowledge, WAGS has yet to be developed for the case of continuous time and space models; for example, those expressed as stochastic or random Partial Differential Equations.

#### Testing and measuring WAGS influence

Above, we have covered different types of WAGS influence. With these definitions in place we now distinguish between testing for the presence of an influence (inference on models) and estimating the strength of the influence (inference on parameters). There is an extensive literature on this, which we shall not go into here. Examples of testing versus measuring for discrete time VAR models include the Dickey–Fuller test and the Geweke measure of influence. In the electrophysiological literature, there are a number of measures proposed. A review and a toolbox for these measures can be found in Seth (2009). From the point of view of effective connectivity, many of these measures have an uncertain status. This is because effective connectivity is only defined in relation to a generative model. In turn, this means there are only two quantities of interest (that permit

<sup>16</sup> In particular the Sinai–Ruelle–Bowen measure for hyperbolic dynamical systems (Chueshov, 2002).



**Fig. 7.** The missing time problem. This figure provides a schematic representation of spurious causality produced by sub-sampling. a) Three time series  $X_1(t)$ ,  $X_2(t)$ , and  $X_3(t)$  are shown changing at an “infinitesimal” time scale with steps  $dt$ , as well as at a coarser sampled time scale with set  $\Delta t$ . Each time series, influences itself at later moments. In the example  $X_3(t)$  directly influences  $X_2(t)$ , with no direct influence on  $X_1(t)$ . In turn  $X_2(t)$  directly influences  $X_1(t)$ , with no direct influence on  $X_3(t)$ . Finally  $X_1(t)$  does not influence either  $X_3(t)$  nor  $X_2(t)$ . There are no contemporaneous influences. b) When only observing at the coarser time scale  $\Delta t$ , spurious contemporaneous influences (mediated by intermediate nodes) appear between  $X_2(t)$  and  $X_1(t)$  and between  $X_3(t)$  and  $X_2(t)$ . In addition a spurious direct influence appears between  $X_3(t)$  and  $X_1(t)$ . The graphical representations of the true and spurious causal relations are to the right of each figure where an arrow represents direct influence and a double arrow represents contemporaneous influence. Estimating these spurious influences can only be avoided by explicitly modeling their effect from continuous models or using models such as VARMA models which are resistant to this phenomena.

inference on models and parameters respectively): the relative evidence for a model with and without a connection and the estimate (conditional density over) the connection parameter. For DCM the first quantity is the Bayes factor and for GCM it is the equivalent likelihood ratio (Granger causal F-statistics). In DCM, the conditional expectation of the parameter (effective connectivity) measures the strength, while for GCM this is the conditional estimate of the corresponding autoregression coefficient. Other measures (e.g., partial directed coherence) are simply different ways of reporting these conditional estimates. The next section explores the use of WAGS measures of direct and indirect effects within the Structural Causal modeling framework, thus bringing together the two major strands of statistical causal modeling.

### Dynamic structural causal modeling

There have been recent theoretical efforts to embed WAGS into Structural Causal Modeling, which one could conceive of (in the language of Granger) as providing a means to find out which “prima facie causes” are actual “causes”. One of the first people to use the methods from Structural Causal Modeling was Granger himself: Swanson and Granger (1997) used Bayes-Net methods described in Spirtes et al. (2000) in combination with autoregressive modeling. Similar approaches have been adopted by Demiralp and Hoover (2008) and Moneta and Spirtes (2006), which address the search for directed contemporaneous influences mentioned above.

However, we should mention three current attempts to combine Structural Causal Modeling with WAGS influence analysis. We shall

follow White in calling models that can be described by both theoretical frameworks Dynamic Structural Systems:

1. Eichler has been developing graphical time series models that are based on discrete time WAGS. Recently, in work with Didelez the formalization of interventions has been introduced and equivalents for the backdoor and front-door criteria of Structural Causality have been defined. Thus, for discrete systems, this work could result in practical criteria for defining when it is possible to infer causal structure from WAGS in discrete time.
2. White has created a general formalism for Dynamical Structural Systems (White and Lu, 2010) based on the concept of settable systems (White and Chalak, 2009), which supports model optimization, equilibrium and learning. The effects of intervention are also dealt with explicitly.
3. Commenges and Gégout-Petit (2009) have also proposed a general framework for causal inference that combines elements of Bayes-Nets and WAGS influence and has been applied to epidemiology. Specifically, as mentioned above, they introduce a very general definition of WAGS that is valid for continuous/discrete time processes. This definition can be applied to a mixture of SDEs and point processes and distinguishes between direct influences and indirect influences. They then relate the definition to graphical models, with nodes connected by direct influences only and place their work in the context of General Systems Theory. Interestingly, they stress the need for an observation equation to assure causal explanatory power.

The common theme of all these efforts is to supplement predictability with additional criteria to extend WAGS influence to inference on

causal mechanisms. In the words of **Gégout-Petit and Commenges (2010)**: “A causal interpretation needs an epistemological act to link the mathematical model to a physical reality.” We will illustrate these ideas with a particular type of SSM, known as a (stochastic) dynamic causal model (DCM):

$$\begin{cases} \dot{x} = f(x, \theta, u) + \omega \\ y = g(x, \theta) + \varepsilon \end{cases} \quad (20)$$

where  $x$  are (hidden) states of the system,  $\theta$  are evolution parameters,  $u$  are the experimental control variables,  $\omega$  are random fluctuations and  $\varepsilon$  is observation noise. Inverting this model involves estimating the evolution parameters  $\theta$ , which is equivalent to characterizing the structural transition density  $p(\dot{x}|do(x))$ , having accounted for observational processes.<sup>17</sup> Here, time matters because it prevents instantaneous cyclic causation, but still allows for dynamics. This is because identifying the structural transition density  $p(\dot{x}|do(x))$  effectively decouples the children of  $X(t)$  (in the future) from its parents (in the past). Let us now examine a bilinear form of this model

$$f(x) = Ax + \sum_i u_i B^{(i)} x + Cu + \sum_j x_j D^{(j)} x. \quad (21)$$

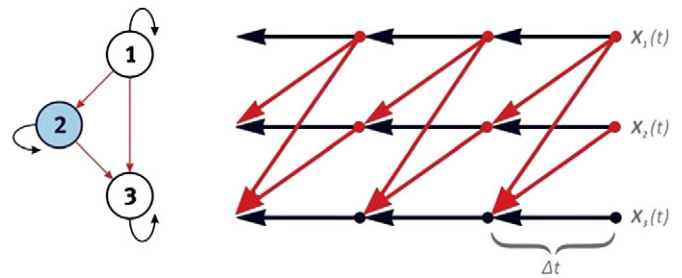
Then we have:

$$\begin{aligned} A &= \lim_{x,u \rightarrow 0} \frac{\partial}{\partial x} E[\dot{x}|do(x)] \\ B^{(i)} &= \frac{\partial^2}{\partial x \partial u_i} E[\dot{x}|do(x)] \\ C &= \lim_{x \rightarrow 0} \frac{\partial}{\partial u_i} E[\dot{x}|do(x)] \\ D^{(j)} &= \frac{\partial^2}{\partial x \partial x_j} E[\dot{x}|do(x)]. \end{aligned} \quad (22)$$

The meaning of  $A$ ; i.e. the effective connectivity is the rate of change (relative to  $x$ ) of the expected motion  $E[\dot{X}]$  where  $X$  is held at  $x \approx 0$ .<sup>18</sup> It measures the *direct effect* of connections. Importantly, *indirect effects* can be derived from the effective connectivity. To make things simple, consider the following 3-region DCM depicted in **Fig. 8**:

$$\begin{aligned} \dot{x}_1 &= A_{11}x_1 + \omega_1 \\ \dot{x}_2 &= A_{21}x_1 + A_{22}x_2 + \omega_2 \\ \dot{x}_3 &= A_{31}x_1 + A_{32}x_2 + A_{33}x_3 + \omega_3. \end{aligned} \quad (23)$$

The effect of node 1 on node 3 is derived from the calculus of the intervention  $do(X_1 = x_1)$ , where  $X_1$  is held constant at  $x_1$  but  $X_2$  is permitted to run its natural course. This intervention confirms that node 1 has both a direct and an indirect effect on node 3 (through node 2).<sup>19</sup> Interestingly, indirect effects can also be derived by



**Fig. 8.** Direct and indirect effects. Causal relationships implied by the DCM given in Eq. (23). On the left the apparent graph, that includes feedback which precludes causal analysis. Note that the causal links are actually expressed through implicit delays, which makes this graph a DAG, which is seen more clearly on the right where each node is expanded at several time instants.

projecting Eq. (20) onto generalized coordinates; i.e. by deriving the evolution function of the augmented state space  $\tilde{x} = (x, \dot{x}, \ddot{x}, \dots)^T$  (see Friston et al., 2008a,b for a variational treatment of stochastic dynamical systems in generalized coordinates). For example, deriving the left and the right hand side of the last equation in Eq. (23) with respect to time yields:

$$\begin{aligned} \ddot{x}_3 &= \tilde{A}_{31}x_1 + \tilde{A}_{32}x_2 + \tilde{A}_{33}x_3 + \tilde{\omega}_3 \\ \tilde{A}_{31} &= \underbrace{A_{31}(A_{11} + A_{33})}_{\text{direct effect}} + \underbrace{A_{32}A_{21}}_{\text{indirect effect}} \\ \tilde{A}_{32} &= A_{32}(A_{22} + A_{33}) \\ \tilde{A}_{33} &= A_{33}A_{33} \end{aligned} \quad (24)$$

where  $\tilde{\omega}_3$  lumps all stochastic inputs (and their time derivatives) together. The total effect of node 1 onto node 3 is thus simply decomposed through the above second order ODE (Eq. (24)), as the sum of direct and indirect effects. One can see that the indirect causal effect of node 1 on node 3 is proportional to the product  $A_{32}A_{21}$  of the path coefficients of the links  $[1 \rightarrow 2]$  and  $[2 \rightarrow 3]$ . This speaks to a partial equivalence of the *do* calculus and the use of generalized coordinates, when modeling both direct and mediated (indirect) effects. This is because embedding the evolution equation into a generalized coordinates of motion naturally accommodates dynamics and the respective contributions of direct/indirect connections (and correlations induced by non-Markovian state noise  $\omega$ ). However, the embedding (truncation) order has to be at least as great as the number of intermediary links to capture indirect effects.

This type of reasoning is very similar to the treatment of direct and indirect influences under WAGS influence and exemplifies a convergence of Structural Causal (Bayes-Net) Modeling and WAGS influence. One could summarize this ambition by noting the “arrow of time” converts realistic (cyclic) graphical models – that include feedback and cyclic connections – into a DAG formalism, to allow full causal inference. So what are the limits of this approach in Neuroimaging?

### Challenges for causal modeling in Neuroimaging

The papers in this C&C highlight challenges that face methods for detecting effective connectivity. These challenges arise mainly in the analysis of BOLD signals. To date, the only experimental examination of these issues is reported in the paper that originated this series (**David et al., 2008**). The main message from the ensuing exchanges is the need to account for the effect of the HRF; that is, to include an appropriate observation model in the analysis, along with careful evaluation of form, priors and Identifiability.

Another approach to testing the validity and limits of the methods discussed above has been through computer simulations. The results of these simulations have been mixed. A number of papers have supported the use of GCM in fMRI (**Deshpande et al., 2009; Stevenson**

<sup>17</sup> Note that the interventional interpretation of DCM is motivated by the (temporal) asymmetry between the left- and the right-hand terms in Eq. (24). Its right-hand term gives us the expected rate of change  $E[\dot{X}(t)]$  of  $X(t)$  if we fix  $X(t)$  to be  $x$  (i.e. if we perform the action  $do(x)$ ), but does not provide any information about what  $X(t)$  is likely to be if we fix its rate of change  $\dot{X}(t)$ . This is best seen by noting that the system’s motion  $\dot{X}(t)$  is a proxy for the system’s *future* state  $X(t + \Delta t)$ , which cannot influence its own past  $X(t)$ . Interestingly, this shows how interventional and prediction-over-time oriented (i.e. WAGS) interpretations of DCM are related.

<sup>18</sup> The original motivation for the neural evolution equation of DCM for fMRI data considered the system’s states  $x$  as being perturbations around the steady-state activity  $x_0$ . Thus,  $x = 0$  actually corresponds to steady (background) activity within the network ( $x_0$ ).

<sup>19</sup> Interventional probabilities in a dynamical setting have recently been derived in, e.g., **Eichler and Didelez (2010)**.



and Körding, 2010; Witt and Meyerand, 2009). Others have shown advantages for Bayes-Net methods in short time series and for GCM for longer time series (Zou et al., 2009).

An extensive set of simulations (NETSIM) has been carried out by Smith et al. (2010b) using non-stationary (Poisson-type) neural innovations in several configurations of nodes and simulating hemodynamics using the fMRI version of DCM. Many different methods were compared (apart from DCM), distinguishing between those that estimate undirected association (functional connectivity) from those that estimate “lagged” dependence (essentially a form of effective connectivity). The main conclusion was that a few undirected association methods that only used the information in the zero lag covariance matrixes perform well in identifying functional connectivity from fMRI. However, lag-based methods “perform worse”. We speculate that lag information is lost by filtering with a (regionally variable) HRF and sub-sampling. Thus one could expect that (stochastic) DCM might perform better, as supported by a comparison of SEM and DCM (Penny et al., 2004).

Interesting as these results are, several points remain unresolved. In the first place, more biophysically realistic simulations are called for, especially in the simulation of neurodynamics. The neurodynamics model in DCM for fMRI is intentionally generic, to ensure identifiability when deconvolving fMRI time-series. There is work suggesting that discrete time Vector Autoregressive Moving Average models are immune to sub-sampling and noise relative to VAR models (Amendola et al., 2010; Solo, 1986; 2007). Considering that WAGS influence modeling with VARMA models is in the standard time series textbooks (Lutkepohl, 2005), it is surprising that this model has not been used in Neuroimaging, with the notable exception of (Victor Solo, 2008).

NETSIM has not yet been tested using continuous time models. The problem, as pointed out by the creators of NETSIM and (Roebroeck et al., 2005), is not only sub-sampling but the combined effect of sub-sampling and the low pass filtering of the HRF. However, these problems only pertain to AR models. Continuous time DCMs have an explicit forward model of (fast) hidden states and are not confounded by sub-sampling or the HRF, provided both are modeled properly in the DCM. The key issue is whether DCM can infer hidden states in the absence of priors (i.e., stimulus functions) that are unavailable for design-free (resting state) fMRI studies of the sort generated by NETSIM. This is an unsettled issue that will surely be followed up in the near future, with the use of biophysically more informed models and new DCM developments; e.g., DCM in generalized coordinates, stochastic DCMs and the DCM–GCM combinations that are being tested at the moment.

It should further be noted that the effect of sub-sampling (and hemodynamic convolution) are only a problem at certain spatial and temporal scales. Undoubtedly it must be a concern, when inferring the dynamics of fast neural phenomena. However, it is clear that brain activity spans many different spatial (Michael Breakspear and Stam, 2005) and temporal (Vanhatalo et al., 2005) scales. Multi-scale time series methods (including WAGS influence measures) have already been used in econometrics (Gencay et al., 2002) and could be applied in neuroscience.

One example of events that occur at a time scale that is probably sufficiently slow to allow simple (AR) WAGS influence analysis are resting state fluctuations observed in concurrent EEG/fMRI recordings. The analysis of causal relations between EEG and BOLD have been studied by several authors (Eichler, 2005; Jiao et al., 2010; Valdés-Sosa et al., 2006) and is illustrated in Fig. 4: The autoregressive coefficients of this first order sparse VAR model suggest that:

1. There are hardly any lag 0 (or contemporaneous) interactions between ROIs.
2. The only coefficients that survive the FDR threshold in the fMRI are those that link each ROI to its own past.

3. There is no influence of the fMRI on the EEG.
4. There are many, interesting interactions, among the EEG sources.
5. There are a number of influences of the EEG sources on the fMRI.

This is a consistent causal model of EEG induced fMRI modulation—valid only for the slow phenomena that survive convolution with the HRF and for the alpha band EEG activity that was investigated here. Of course there are neural phenomena that might show up at as contemporaneous at this sampling rate—but we have filtered them out. An interesting analysis of information recoverable at each scale can be found in Deneux and Faugeras (2010).

#### Conclusion and suggestions for further work

1. We believe that the simulation efforts that are being carried currently out are very useful and should be extended to cover a greater realism in the neurodynamics, as well as to systematically test new proposals.
2. It will be also be important to have standardized experimental data from animals as a resource for model testing. Ideally this data set should provide intracranial recordings of possible neural drivers, BOLD-fMRI, surface EEG, diffusion MRI based structural connectivity and histological based connectivity matrices.<sup>20</sup>
3. There is a clear need for tools that can assess model evidence (and establish their Identifiability) when dealing with large model spaces of biophysically informed SSMs. These should be brought to bear on the issue of bounds on model complexity, imposed by the HRF convolution and sub-sampling in fMRI.
4. We foresee the following theoretical developments in Causal modeling for effective connectivity:
  - a. The fusion of Bayes-Net and WAGS methods.
  - b. The WAGS tools developed for combined point and continuous time stochastic processes may play an important role in the connectivity analysis of EEG/fMRI, LFP and spike train data.
  - c. WAGS methods must be extended to non-standard models, among others: non-Markovian, RDE, and delay differential equations.
5. The development of exploratory (nonparametric), large scale state-space methods that are biophysically constrained and contain modality specific observation equations. This objective will depend critically on the exploration of large model spaces and is in consistent with the recent surge of methods analyzing “Ultra-High” dimensional data.
6. The explicit decomposition of multiple spatial and frequency scales.
7. Effective connectivity in the setting of Neural Field Modeling

We hope to have focused attention on these issues, within a unifying framework that integrates apparently disparate and important approaches. We are not saying that DCM and GCM are equivalent, but rather that an integration is possible within a Bayesian SSM framework and the use of model comparison methods. Our review of the field has been based on the use of state space models (SSM). While we are aware that SSMs are not the only possible framework for analyzing effective connectivity, this formulation allowed us to present a particular view that we feel will stimulate further work.

Besides reviewing current work we have discussed a number of new mathematical tools: Random Differential Equations, non-Markovian models, infinitely differentiable sample path processes, as well as the use of graphical causality models. We also considered the use of

<sup>20</sup> Such a data set in an animal model including EEG, EcoG, DWI tractography and fMRI is being gathered by Jorge Riera (Tohoku University), within a collaboration including F. H. Lopes da Silva, Thomas Knoesche, Olivier David, and the authors of this paper. This data set will be made publicly available in the near future.

continuous-time AR and ARMA models. It may well be that some of these techniques will not live up to expectations, but we feel our field will benefit from these and other new tools that confront some of the particular challenges addressed in this discussion series.

## Acknowledgments

We dedicate this paper to Rolf Kötter for the many insights and the promotion of the Brain Connectivity Workshops that influenced this work. We also wish to thank Steve Smith for stimulating discussions, as well as to Daniel Commenges, Rolando Biscay, Juan Carlos Jimenez, Guido Nolte, Tohru Ozaki, Victor Solo, Nelson Trujillo, and Kamil Uludag for helpful input to the contents of this paper. An important part of the work described here was conceived and executed during the “The Keith Worsley workshop on Computational Modeling of Brain Dynamics: from stochastic models to Neuroimages (09w5092)” organized by the Banff International Research Station.

## Appendix A. Wiener's original definition of causality

This approach was first formalized by Wiener (1956) as follows.<sup>21</sup> Consider a strictly stationary (possibly complex) stochastic processes<sup>22</sup>  $X_1(t, \omega)$  defined as a collection of random variable for all integer time instants  $t$  and realizations  $\omega$ . Wiener showed how to construct its “innovation”—the unit variance white noise time series  $E_1(t, \omega)$  which is uncorrelated with the past of  $X_1(t, \omega)$ . The innovation  $E_2(t, \omega)$  can also be constructed for a second time series  $X_2(t, \omega)$ . Now consider the random variable  $K_1(\omega)$ , that part of  $E_1(t, \omega)$  uncorrelated with its own past and that of  $E_2(t, \omega)$ . The variance of this random variable lies between 0 and 1 and is the degree to which the time series  $X_1(t, \omega)$  does not depend on the past of  $X_2(t, \omega)$ . One minus this variance is the Wiener measure C of the causal effect of  $X_2(t, \omega)$  on  $X_1(t, \omega)$ . This measure of influence was in fact expressed by Wiener as an infinite sum:

$$I_{2 \rightarrow 1}^W = \sum_{m=1}^{\infty} |\rho(t, t-m)|^2 + \sum_{m=1}^{\infty} \left| \sum_{n=1}^{\infty} \rho(t, t-m) \overline{\rho(t-n, t-m)} \right|^2 + \dots$$

$$\rho(t, s) = E[X_1(t) \overline{X_2(s)}] \quad (25)$$

where  $\overline{X(s)}$  indicates the complex conjugate of a time series.

As pointed out in Bressler and Seth (2010) this definition is not practical. We elaborate on why: First, it is limited to strictly stationary processes and involves an infinite series of moments without specification of how to perform the requisite calculations. More seriously, it only involves a finite number of series and ignores the potential confounding effect of unobserved (or latent) causes. More importantly, it adopts the “functional formulation” of von Mises that lost out to the currently predominant “stochastic formulation” of Kolmogorov and Doob (Von Mises and Doob, 1941). Nevertheless Wiener's definition has several points that deserve to be highlighted:

1. It was not limited to autoregressive models but was based on the more general Moving Average Representation (MAR).
2. Although defined explicitly for discrete time stochastic processes, the extension to continuous time was mentioned explicitly.
3. Applications in neuroscience were anticipated. In fact, Wiener elaborated on its possible use: “Or again, in the study of brain waves we may be able to obtain electroencephalograms more or

less corresponding to electrical activity in different parts of the brain. Here the study of the coefficients of causality running both ways and of their analogs for sets of more than two functions  $f$  may be useful in determining what part of the brain is driving what other part of the brain *in its normal activity*”.

4. It is instructive to compare this initial definition with modern accounts of direct influence.

## References

- Aalen, O.O., 1987. Dynamic modeling and causality. *Scand. Actuarial J.* 13, 177–190.
- Aalen, O.O., Frigessi, A., 2007. What can statistics contribute to a causal understanding. *Scand. J. Stat.* 34 (1), 155–168. doi:10.1111/j.1467-9469.2006.00549.x
- Akaike, H., 1968. On the use of a linear model for the identification of feedback systems. *Annals of the Institute of Statistical Mathematics*, 20(1). Springer, pp. 425–439. Retrieved from <http://www.springerlink.com/index/MP5748216213R74Q.pdf>
- Amendola, A., Niglio, M., Vitale, C., 2010. Temporal aggregation and closure of VARMA models: some new results. In: Palumbo, F., Lauro, C.N., Greenacre, M.J. (Eds.), *Data Analysis and Classification*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 435–443. doi:10.1007/978-3-642-03739-9
- Angelova, M., Wennberg, B., 2010. On analytic and algebraic observability of nonlinear delay systems. *Automatica*, 46(4). Elsevier Ltd., pp. 682–686. doi:10.1016/j.automatica.2010.01.031
- Astrom, K.J., 1969. On the choice of sampling rates in parametric identification of time series. *Inf. Sci.* 1, 273–278.
- August, E., Papachristodoulou, A., 2009. A new computational tool for establishing model parameter identifiability. *J. Comput. Biol.* 16 (6), 875–885. doi:10.1089/cmb.2008.0211
- Belyaev, Y.K., 1959. Analytic random processes. *Theory Probab. Appl.* 4 (4), 402. doi:10.1137/1104040
- Bergstrom, A.R., 1966. Nonrecursive models as discrete approximations to systems of stochastic differential equations. *Econometrica* 34 (1), 173–182.
- Bergstrom, A.R., 1984. Continuous time stochastic models and issues of aggregation. In: Griliches, Z., Lnrilligato, M.D. (Eds.), *Handbook of Econometrics*, Volume II. Elsevier Science Publishers B.V.
- Bergstrom, A.R., 1988. Continuous-time models, realized volatilities, and testable distributional implications for daily stock returns. *Econometric Theory* 4 (3), 365–383. doi:10.1002/jae.1105
- Bojak, Ingo, Liley, D.T.J., 2010. Axonal velocity distributions in neural field equations. *PLoS Comput. Biol.* 6 (1), 1–25. doi:10.1371/journal.pcbi.1000653
- Bosch-Bayard, J., Valdés-Sosa, P., Virues-Alba, T., Aubert-Vázquez, E., John, E.R., Harmony, T., et al., 2001. 3D statistical parametric mapping of EEG source spectra by means of variable resolution electromagnetic tomography (VARETA). *Clinical EEG (Electroencephalography)*, 32(2). ECNS, pp. 47–61. Retrieved September 13, 2010, from <http://www.ncbi.nlm.nih.gov/pubmed/11360721>
- Brandt, S.F., Pelster, A., Wessel, R., 2007. Synchronization in a neuronal feedback loop through asymmetric temporal delays. *Europhys. Lett.* 79 (3), 38001. doi:10.1209/0295-5075/79/38001
- Breakspear, Michael, Stam, C.J., 2005. Dynamics of a neural system with a multiscale architecture. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360 (1457), 1051–1074. doi:10.1098/rstb.2005.1643
- Breakspear, M., Roberts, J.A., Terry, J.R., Rodrigues, S., Mahant, N., Robinson, P.A., 2006. A unifying explanation of primary generalized seizures through nonlinear brain modeling and bifurcation analysis. *Cereb. Cortex* 16 (9), 1296–1313. doi:10.1093/cercor/bhj072
- Bressler, S.L., Seth, A.K., 2011. Wiener–Granger causality: a well established methodology. *Neuroimage* 58, 323–329 (this issue).
- Bunge, M., 2009. *Causality and Modern Science*. Book, Fourth. Dover Publications, New York.
- Calbo, G., Cortés, J.-C., Jódar, L., 2010. Mean square power series solution of random linear differential equations. *Computers & Mathematics with Applications*, 59(1). Elsevier Ltd., pp. 559–572. doi:10.1016/j.camwa.2009.06.007
- Candy, J.V., 2006. *Model Based Signal Processing*. Book. IEEE Press, 701 pp.
- Carbonell, F., Biscay, R.J., Jimenez, J.C., de la Cruz, H., 2007. Numerical simulation of nonlinear dynamical systems driven by commutative noise. *J. Comput. Phys.* 226 (2), 1219–1233. doi:10.1016/j.jcp.2007.05.024
- Cartwright, N., 2007. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Politics. Cambridge University Press, Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, Sao Paulo.
- Chambers, Marcus J., Thornton, M.A., 2009. Discrete time representation of continuous time ARMA processes. *Por Clasificar*, pp. 1–19.
- Chen, C.C., Kiebel, S.J., Friston, K.J., 2008. Dynamic causal modelling of induced responses. *Neuroimage* 41 (4), 1293–1312. doi:10.1016/j.neuroimage.2008.03.026
- Chen, B., Hu, J., Zhu, Y., Sun, Z., 2009. Parameter identifiability with Kullback–Leibler information divergence criterion. *Int. J. Adapt. Control Signal Process.* 23 (10), 940–960. doi:10.1002/acs.1078
- Chueshov, I.D., 2002. Introduction to the Theory of Infinite-Dimensional Dissipative Systems. Book. ACTA Scientific Publishing House, Kharkov, pp. 1–419. Retrieved from <http://www.emis.de/monographs/Chueshov/>
- Commenges, D., Gégout-Petit, A., 2009. A general dynamical statistical model with causal interpretation. *J. R. Stat. Soc. B Stat. Methodol.* 71 (3), 719–736. doi:10.1111/j.1467-9868.2009.00703.x

<sup>21</sup> With some loss of rigor we have simplified the definitions, making our notation consistent with current time series analysis. For greater detail please consult the original references.

<sup>22</sup> That is  $\Pr(X_1(t_1, \omega), \dots, X_1(t_n, \omega)) = \Pr(X_1(t_1 + \tau, \omega), \dots, X_1(t_n + \tau, \omega))$  for all for all  $n$  and  $\tau$ .

- Comte, F., Renault, E., 1996. Noncausality in continuous time models. *Econometric Theory* 12, 215–256.
- Coombes, S., 2010. Large-scale neural dynamics: simple and complex. *NeuroImage* 52 (3), 731–739. doi:10.1016/j.neuroimage.2010.01.045
- Coombes, S., Venkov, N., Shiao, L., Bojak, I., Liley, D., Laing, C., 2007. Modeling electrocortical activity through improved local approximations of integral neural field equations. *Phys. Rev. E* 76 (5), 1–8. doi:10.1103/PhysRevE.76.051901
- Cox, D.R., Wermuth, N., 2004. Causality: a statistical view. *Int. Stat. Rev.* 72, 285–305.
- Daunizeau, J., David, O., Stephan, K.E., 2011a. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *NeuroImage* 58, 312–322 (this issue).
- Daunizeau, J., Friston, K.J., Kiebel, S.J., 2009b. Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D* 238 (21), 2089–2118. doi:10.1016/j.physd.2009.08.002
- Daunizeau, Jean, Kiebel, Stefan J., Friston, Karl J., 2009c. Dynamic causal modelling of distributed electromagnetic responses. *NeuroImage*, 47(2). Elsevier Inc., pp. 590–601. doi:10.1016/j.neuroimage.2009.04.062
- David, O., Kilner, J.M., Friston, K.J., 2006. Mechanisms of evoked and induced responses in MEG/EEG. *NeuroImage* 31 (4), 1580–1591. doi:10.1016/j.neuroimage.2006.02.034
- David, Olivier, 2011. fMRI connectivity, meaning and empiricism Comments on: Roebroeck et al. The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution. *NeuroImage* 58, 306–309 (this issue).
- David, Olivier, Guillemain, I., Saitlet, S., Rey, S., Deransart, C., Segebarth, C., et al., 2008. Identifying neural drivers with functional MRI: an electrophysiological validation. *PLoS Biol.* 6 (12), 2683–2697. doi:10.1371/journal.pbio.0060315
- Deco, G., Jirsa, V.K., Robinson, Peter A., Breakspear, Michael, Friston, Karl, 2008. The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Comput. Biol.* 4 (8), e1000092. doi:10.1371/journal.pcbi.1000092
- Demiralp, S., Hoover, K., 2008. A bootstrap method for identifying and evaluating a structural vector autoregression\*. *Oxford Bulletin of Economics and Statistics*. Retrieved August 27, 2010, from <http://www3.interscience.wiley.com/journal/120174048/abstract>
- Deneux, T., Fauergas, O., 2010. EEG-fMRI fusion of paradigm-free activity using Kalman filtering. *Neural Comput.* 22 (4), 906–948. doi:10.1162/neco.2009.05-08-793.
- Deshpande, G., Sathian, K., Hu, X., 2009. Effect of hemodynamic variability on Granger causality analysis of fMRI. *NeuroImage*, 52(3). Elsevier Inc., pp. 884–896. doi:10.1016/j.neuroimage.2009.11.060
- Dhamala, M., Rangarajan, G., Ding, M., 2008. Analyzing information flow in brain networks with nonparametric Granger causality. *NeuroImage* 41 (2), 354–362. doi:10.1016/j.neuroimage.2008.02.020
- Eichler, M., 2005. A graphical approach for evaluating effective connectivity in neural systems. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360 (1457), 953–967. doi:10.1098/rstb.2005.1641
- Eichler, M., Didelez, V., 2010. On Granger causality and the effect of interventions in time series. *Lifetime data analysis* 16 (1), 3–32. doi:10.1007/s10985-009-9143-3
- Faugeras, O., Touboul, J., Cessac, B., 2009. A constructive mean-field analysis of multi-population neural networks with random synaptic weights and stochastic inputs. *Front. Comput. Neurosci.* 3, 1. doi:10.3389/neuro.10.001.2009 (February).
- Florens, J.-P., 2003. Some technical issues in defining causality. *J. Econometrics* 112, 127–128.
- Florens, J.-P., Fougere, D., 1996. Noncausality in continuous time. *Econometrica* 64 (5), 1195–1212 Retrieved from <http://www.jstor.org/stable/2171962>
- Florens, J.-P., Mouchart, M., 1982. A note on noncausality. *Econometrica* 50 (3), 583–591 Retrieved from <http://www.jstor.org/stable/1912602>
- Florens, A.J., Mouchart, M., 1985. A linear theory for noncausality. *Econometrica* 53 (1), 157–176.
- Freiwald, W., Valdes-Sosa, P.A., Bosch-Bayard, Jorge, Biscay-Lirio, R., Jimenez, Juan Carlos, Rodríguez, L.M., et al., 1999. Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. *J. Neurosci. Methods* 94 (1), 105–119. doi:10.1016/S0165-0270(99)00129-6
- Friston, K.J., 2008a. Variational filtering. *NeuroImage* 41, 747–766.
- Friston, Karl J., 2008b. Hierarchical models in the brain. *PLoS Comput. Biol.* 4 (11), e1000211. doi:10.1371/journal.pcbi.1000211
- Friston, Karl, 2009a. Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS Biol.* 7 (2), e33. doi:10.1371/journal.pbio.1000033
- Friston, Karl, 2011. Dynamic causal modeling and Granger causality comments on: the identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *NeuroImage* 58, 303–305 (this issue).
- Friston, K.J., Daunizeau, J., 2008. DEM: a variational treatment of dynamic systems. *NeuroImage* 41, 849–885. doi:10.1016/j.neuroimage.2008.02.054
- Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. *NeuroImage* 19 (4), 1273–1302.
- Friston, K.J., Mechelli, A., Turner, R., Price, C.J., 2000. Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage* 12 (4), 466–477. doi:10.1006/nimg.2000.0630
- Frosini, B.V., 2006. Causality and causal models: a conceptual. *Int. Stat. Rev.* 305–334 (June 2004).
- Galka, A., Yamashita, O., Ozaki, Tohru, Biscay, Rolando, Valdés-Sosa, Pedro, 2004. A solution to the dynamical inverse problem of EEG generation using spatiotemporal Kalman filtering. *NeuroImage* 23 (2), 435–453. doi:10.1016/j.neuroimage.2004.02.022
- Galka, A., Ozaki, Tohru, Muhle, H., Stephani, U., Siniatchkin, M., 2008. A data-driven model of the generation of human EEG based on a spatially distributed stochastic wave equation. *Cogn. Neurodynamics* 2 (2), 101–113. doi:10.1007/s11571-008-9049-x
- Garnier, H., Wang, L., 2008. Identification of continuous time models from sampled data. In: Garnier, H., Wang, L. (Eds.), *Engineering*. Springer Verlag, London Limited.
- Ge, T., Kendrick, K.M., Feng, J., 2009. A novel extended Granger Causal Model approach demonstrates brain hemispheric differences during face recognition learning. *PLoS Comput. Biol.* 5 (11), e1000570. doi:10.1371/journal.pcbi.1000570
- Gégout-Petit, A., Commenges, D., 2010. A general definition of influence between stochastic processes. *Lifetime Data Anal.* 16 (1), 33–44. doi:10.1007/s10985-009-9131-7
- Gencay, R., Selcuk, F., Whitcher, B., 2002. An introduction to wavelets and other filtering methods in finance and economics. : Statistics, Vol. 12. Academic Press.
- Geweke, J., 1984. Measures of conditional linear dependence and feedback between time series. *J. Am. Stat. Assoc.* 79 (388), 907–915.
- Gill, J.B., Petrović, L., 1987. Causality and stochastic dynamical systems. *SIAM J. Appl. Math.* 47 (6), 1361–1366.
- Glover, G.H., 1999. Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage* 9 (4), 416–429 Retrieved August 21, 2010, from <http://www.ncbi.nlm.nih.gov/pubmed/10191170>
- Glymour, C., 2009. What Is Right with 'Bayes Net Methods' and What Is Wrong with 'Hunting Causes and Using Them'? *The British Journal for the Philosophy of Science* 61 (1), 161–211. doi:10.1093/bjps/axp039
- Gourieroux, C., Monfort, A., Renault, Eric, 1987. Kullback Causality Measures. *Granger, C.W.J.*, 1963. Economic processes involving feedback. *Inf. Control* 48, 28–48.
- Granger, C.W.J., 1988. Some recent developments in a concept of causality. *J. Econometrics* 39, 199–211.
- Hansen, L.P., Sargent, T.J., 1983. The dimensionality of the aliasing problem in models with rational spectral densities. *Econometrica* 51 (2), 377–387.
- Havlicek, M., Jan, J., Calhoun, V.M., 2009. Extended time-frequency Granger causality for evaluation of functional network connectivity in event-related fMRI data. Conference of the IEEE, pp. 4440–4443. Retrieved August 27, 2010, from <http://www.ncbi.nlm.nih.gov/pubmed/19963833>
- Havlicek, Martin, Jan, Jiri, Brazdil, M., Calhoun, V.D., 2010. Dynamic Granger causality based on Kalman filter for evaluation of functional network connectivity in fMRI data. *NeuroImage*, 53(1). Elsevier Inc., pp. 65–77. doi:10.1016/j.neuroimage.2010.05.063
- Havlicek, M., Friston, K.J., Jan, J., Brazdil, M., Calhoun, V.D., 2011. Dynamic modeling of neuronal responses in fMRI using cubature Kalman filtering. *NeuroImage*. Elsevier Inc. doi:10.1016/j.neuroimage.2011.03.005
- Holden, H., Oksendal, B., Ub, J., Zhang, T., 1996. Holden, Oksendal et al 1996—Stochastic Partial Differential Equations. Birkhauser.
- Jansen, B.H., Rit, V.G., 1995. Biological Cybernetics in a mathematical model of coupled cortical columns. *Biol. Cybern.* 366, 357–366.
- Jentzen, A., Kloeden, P.E., 2009. Pathwise Taylor schemes for random ordinary differential equations. *BIT Numer. Math.* 49 (1), 113–140. doi:10.1007/s10543-009-0211-6
- Jiao, Q., Lu, G., Zhang, Z., Zhong, Y., Wang, Z., Guo, Y., et al., 2010. Granger causal influence predicts BOLD activity levels in the default mode network. *Hum. Brain Mapp.* 1–8. doi:10.1002/hbm.21065
- Jirsa, V.K., et al., 2002. Spatiotemporal Forward Solution of the EEG and MEG Using Network Modeling. *IEEE Transactions on Medical Imaging* 21 (5), 493–504.
- Kailath, T., 1980. *Linear Systems*. Book. Prentice Hall, New Jersey.
- Kalitzin, S.N., Parra, J., Velis, D.N., Lopes da Silva, F.H., 2007. Quantification of unidirectional nonlinear associations between multidimensional signals. *IEEE Trans. Biomed. Eng.* 54 (3), 454–461. doi:10.1109/TBME.2006.888828
- Larsson, E.K., Mossberg, M., Soderstrom, T., 2006. An overview of important practical aspects of continuous-time ARMA system identification. *Circuits Syst. Signal Process.* 25 (1), 17–46. doi:10.1007/s00034-004-0423-6
- Lauritzen, S., 1996. Graphical Models. Oxford University Press.
- Ljung, L., Glad, T., 1994. On global identifiability for arbitrary model parametrizations. *Automatica* 30 (2), 265–276.
- Łuczka, J., 2005. Non-Markovian stochastic processes: colored noise. *Chaos (Woodbury, N.Y.)* 15 (2), 26107. doi:10.1063/1.1860471
- Lutkepohl, H., 2005. *New Introduction to Multiple Time Series Analysis*. Book. Springer, pp. 1–764.
- Lyman, R.J., Edmonson, W.W., McCullough, S., Rao, M., 2000. The predictability of continuous-time, bandlimited processes. *IEEE Trans. Signal Process.* 48 (2), 311–316.
- Machamer, P., Darden, L., Craver, C.F., Machamer, P., 2000. Thinking about mechanisms. *Philos. Sci.* 67 (1), 1–25.
- Maiwald, Thomas, Timmer, Jens, 2008. Dynamical modeling and multi-experiment fitting with PottersWheel. *Bioinformatics (Oxford, England)* 24 (18), 2037–2043. doi:10.1093/bioinformatics/btn350
- Marinazzo, D., Liao, W., Chen, H., Stramaglia, S., 2011. Nonlinear connectivity by Granger causality. *NeuroImage* 58, 330–338 (this issue).
- Marreiros, A.C., Kiebel, Stefan J., Daunizeau, Jean, Harrison, L.M., Friston, Karl J., 2009. Population dynamics under the Laplace assumption. *NeuroImage*, 44(3). Elsevier Inc., pp. 701–714. doi:10.1016/j.neuroimage.2008.10.008
- Marrelec, G., Benali, H., Ciuciu, P., Pélégriani-Issac, M., Poline, J.-B., 2003. Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. *Hum. Brain Mapp.* 19 (1), 1–17. doi:10.1002/hbm.10100
- Martínez-Montes, E., Valdés-Sosa, P.A., Miwakeichi, F., Goldman, R.I., Cohen, M.S., 2004. Concurrent EEG/fMRI analysis by multiway Partial Least Squares. *NeuroImage* 22 (3), 1023–1034. doi:10.1016/j.neuroimage.2004.03.038
- Marzetti, L., Del Gratta, C., Nolte, G., 2008. Understanding brain connectivity from EEG data by identifying systems composed of interacting sources. *NeuroImage* 42 (1), 87–98. doi:10.1016/j.neuroimage.2008.04.250
- Mccrorie, J. Roderick, 2003. The problem of aliasing in identifying finite parameter continuous time stochastic models. *Acta Applicandae Mathematicae* 79, 9–16.
- Mccrorie, J.R., Chambers, M.J., 2006. Granger causality and the sampling of economic. *J. Econometrics* 132, 311–326.
- Medvedev, P., 2007. *Stochastic Integration Theory*. Oxford University Press, USA. Retrieved July 8, 2010, from <http://books.google.com/books?hl=en&lr=&id=>

- pZGKC\_PVvBsC&oi=fnd&pg=PR13&dq=Stochastic+Integration+Theory&ots=IogSuzUxwK&sig=j29s0LLApDxAueHhDp5qNylcOQ
- Minchev, B., Wright, W., 2005. A review of exponential integrators for first order semi-linear problems. Preprint Numerics. Trondheim, Norway. Retrieved August 29, 2010, from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+review+of+exponential+integrators+for+first+order+semi-linear+problems#0>
- Moneta, A., Spirtes, Peter, 2006. Graphical models for the identification of causal structures in multivariate time series models. Proceedings of the 9th Joint Conference on Information Sciences (JCIS), 1. Atlantis Press, Paris, France, pp. 1–4. doi:10.2991/jcis.2006.171
- Moran, R.J., Stephan, K.E., Kiebel, S.J., Rombach, N., O'Connor, W.T., Murphy, K.J., et al., 2008. Bayesian estimation of synaptic physiology from the spectral responses of neural masses. *NeuroImage* 42 (1), 272–284. doi:10.1016/j.neuroimage.2008.01.025
- Mykland, P., 1986. Statistical Causality.
- Nalatore, H., Ding, M., Rangarajan, G., 2007. Mitigating the effects of measurement noise on Granger causality. *Phys. Rev. E* 75 (3). doi:10.1103/PhysRevE.75.031123
- Nolte, G., Meinecke, F., Ziehe, A., Müller, K.-R., 2006. Identifying interactions in mixed and noisy complex systems. *Phys. Rev. E* 73 (5), 1–6. doi:10.1103/PhysRevE.73.051913
- Nolte, G., Ziehe, A., Nikulin, V., Schlögl, A., Krämer, N., Brismar, T., et al., 2008. Robustly estimating the flow direction of information in complex physical systems. *Phys. Rev. Lett.* 100 (23), 1–4. doi:10.1103/PhysRevLett.100.234101
- Nolte, G., Marzetti, L., Valdes Sosa, P., 2009. Minimum Overlap Component Analysis (MOCA) of EEG/MEG data for more than two sources. *J. Neurosci. Methods* 183 (1), 72–76. doi:10.1016/j.jneumeth.2009.07.006
- Ozaki, T., 1990. Contribution to the discussion of M.S. Bartlett's paper, 'Chance and chaos'. *J. R. Stat. Soc. Ser. A* 153, 330–346.
- Ozaki, Tohru, 1992. A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach. *Stat. Sin.* 2, 113–135.
- Ozaki, Tohru, 2011. Statistical Time Series Modelling Approach to Signal Decomposition, Inverse Problems and Causality Analysis for Neuroscience Data. CRC Press.
- Pearl, J., 2000. *Causality: Models, Reasoning and Inference*. Cambridge University Press.
- Pearl, J., 2003. Statistics and causal inference: a review. *Test* 12 (2), 281–345.
- Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Modelling functional integration: a comparison of structural equation and dynamic causal models. *NeuroImage* 23 (Suppl 1), S264–S274. doi:10.1016/j.neuroimage.2004.07.041
- Penny, W., Ghahramani, Z., Friston, K., 2005. Bilinear dynamical systems. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360 (1457), 983–993. doi:10.1098/rstb.2005.1642
- Petrović, L., Stanojević, D., 2010. Statistical causality, extremal measures and weak solutions of stochastic differential equations with driving semimartingales. *J. Math. Modell. Algorithms* 9 (1), 113–128. doi:10.1007/s10852-009-9121-5
- Phillips, P.C.B., 1973. The problem of identification in finite parameter continuous time models. *J. Econometrics* 1, 351–362.
- Phillips, P.C.B., 1974. The estimation of some continuous time models. *Econometrica* 42 (5), 803–823.
- Pollock, D., 2010. Oversampling of stochastic processes. Working Papers, 2. Retrieved August 27, 2010, from <http://ideas.repec.org/p/wse/wpaper/44.html>
- Ramsey, J.D., Hanson, S.J., Hanson, C., Halchenko, Y.O., Poldrack, R.A., Glymour, C., 2010. Six problems for causal inference from fMRI. *NeuroImage*, 49(2). Elsevier Inc., pp. 1545–1558. doi:10.1016/j.neuroimage.2009.08.065
- Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., et al., 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics Oxford England* 25 (15), 1923–1929. doi:10.1093/bioinformatics/btp358
- Renault, Eric, Sekkat, K., Szafarz, A., 1998. Testing for spurious causality in exchange rates. *J. Empir. Finance* 5, 47–66.
- Riera, J.J., Jimenez, J.C., Wan, X., Kawashima, R., Ozaki, T., 2007a. Nonlinear local electrovascular coupling. II: from data to neuronal masses. *Hum. Brain Mapp.* 354 (August 2006), 335–354. doi:10.1002/hbm.20278
- Riera, J.J., Jimenez, J.C., Wan, X., Kawashima, R., Ozaki, T., 2007b. Nonlinear local electrovascular coupling. II: from data to neuronal masses. *Hum. Brain Mapp.* 354 (August 2006), 335–354. doi:10.1002/hbm.20278
- Riera, Jorge J., Wan, Xiaohong, Jimenez, Juan Carlos, Kawashima, Ryuta, 2006. Nonlinear local electrovascular coupling. I: a theoretical model. *Hum. Brain Mapp.* 27 (11), 896–914. doi:10.1002/hbm.20230
- Robinson, P.M., 1991. Automatic frequency domain inference on semiparametric and nonparametric models. *Econometrica* 59 (5), 1329. doi:10.2307/2938370
- Robinson, P.A., Chen, P.-chia, Yang, L., 2008. Physiologically based calculation of steady-state evoked potentials and cortical wave velocities. *Biol. Cybern.* 98 (1), 1–10. doi:10.1007/s00422-007-0191-z
- Roebroek, A., Formisano, E., Goebel, R., 2005. Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage* 25 (1), 230–242 Retrieved from <Go to ISI>: <http://WOS:000227369600021>
- Roebroek, Alard, Formisano, Elia, Goebel, Rainer, 2011a. Reply to Friston and David fMRI: model selection, causality and deconvolution. *NeuroImage* 58, 310–311 (this issue).
- Roebroek, Alard, Formisano, Elia, Goebel, Rainer, 2011b. The identification of interacting networks in the brain using fMRI: model selection, causality and deconvolution. *NeuroImage* 58, 296–302 (this issue).
- Saccomani, M.P., Audoly, S., Bellu, G., D'Angiò, L., 2010. Examples of testing global identifiability of biological and biomedical models with the DAISY software. *Comput. Biol. Med.* 40 (4), 402–407. doi:10.1016/j.compbiomed.2010.02.004
- Sanchez-Bornot, J., Martinez-Montes, E., Lage-Castellanos, Agustín, Vega-Hernández, M., Valdes-Sosa, P.A., 2008. Uncovering sparse brain effective connectivity: a voxel-based approach using penalized regression. *Stat. Sin.* 18, 1501–1518 Retrieved from <http://www3.stat.sinica.edu.tw/statistica/password.asp?vol=18&num=4&art=14>
- Sargan, J.D., 1974. Some discrete approximations to continuous time stochastic models. *J. R. Stat. Soc. B* 36 (1), 74–90.
- Schwartz, E.L., 1977. Spatial mapping in the primate sensory projection: analytic structure and relevance to perception. *Biol. Cybern.* 25, 181–194.
- Schweder, T., 1970. Composable Markov processes. *J. Appl. Probab* 7 (2), 400. doi:10.2307/3211973
- Seth, A.K., 2009. Granger Causal Connectivity Analysis: A MATLAB Toolbox.
- Shampine, L.F., Gahinet, P., 2006. Delay-differential-algebraic equations in control theory. *Appl. Numer. Math.* 56, 574–588. doi:10.1016/j.apnum.2005.04.025
- Shardlow, T., 2003. Numerical simulation of stochastic PDEs for excitable media. Analysis University of Manchester. Numerical Analysis Report 437.
- Smith, J.F., Pillai, A., Chen, K., Horwitz, B., 2010a. Identification and validation of effective connectivity networks in functional magnetic resonance imaging using switching linear dynamic systems. Manuscript. *NeuroImage* 52 (3), 1027–1040. doi:10.1016/j.neuroimage.2009.11.081
- Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., et al., 2010b. Network modelling methods for fMRI. *NeuroImage*. doi:10.1016/j.neuroimage.2010.08.063
- Solo, V., 1986. Topics in advanced time series analysis. In: Pino, G., Rebollo, R. (Eds.), *Lectures in Probability and Statistics*, Vol. 1215. Springer, Berlin Heidelberg. doi:10.1007/BFb0075871
- Solo, V., 2007. On causality I: sampling and noise. 46th IEEE Conference on Decision and Control. IEEE, pp. 3634–3639. Retrieved June 29, 2010, from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:On+Causality+I:+Sampling+and+Noise#0>
- Solo, V., 2008. On causality and mutual information. Proceedings of the 47th IEEE Conference on Decision and Control, pp. 4939–4944.
- Spirtes, P., Glymour, C., Scheines, R., 2000. *Causation, Prediction and Search*. Cambridge University Press.
- Stephan, Klaas Enno, Kasper, L., Harrison, L.M., Daunizeau, Jean, den Ouden, H.E.M., Breakspear, Michael, et al., 2008. Nonlinear dynamic causal models for fMRI. *NeuroImage* 42 (2), 649–662. doi:10.1016/j.neuroimage.2008.04.262
- Stevenson, I.H., Kording, K.P., 2010. On the similarity of functional connectivity between neurons estimated across timescales. *PLoS One* 5 (2), e9206. doi:10.1371/journal.pone.0009206
- Supp, G.G., Schlögl, A., Trujillo-Barreto, Nelson, Müller, M.M., Gruber, T., 2007. Directed cortical information flow during human object recognition: analyzing induced EEG gamma-band responses in brain's source space. *PLoS One* 2 (1), e684. doi:10.1371/journal.pone.0000684
- Sussmann, H., 1977. An interpretation of stochastic differential equations as ordinary differential equations which depend on the sample point. *Am. Math. Soc.* 83 (2), 296–298 Retrieved August 21, 2010, from <http://www.ams.org/journals/bull/1977-83-02/S0002-9904-1977-14312-7/S0002-9904-1977-14312-7.pdf>
- Swanson, N.R., Granger, C.W.J., 1997. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *J. Am. Stat. Assoc.* 92 (437), 357. doi:10.2307/2291481
- Triacca, U., 2007. Granger causality and contiguity between stochastic processes. *Phys. Lett. A* 362, 252–255. doi:10.1016/j.physleta.2006.10.024
- Valdes-Sosa, P.A., Jimenez, J.C., Riera, J., Biscay, R., Ozaki, T., 1999. Nonlinear EEG analysis based on a neural mass model. *Biol. Cybern.* 81 (5–6), 415–424 Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10592017>
- Valdés-Sosa, P.A., Bosch, J., Jiménez, J., Trujillo, N., Biscay, R., Morales, F., et al., 1999. The statistical identification of nonlinear brain dynamics: a progress report. *Non linear Dynamic and Brain Functioning*, pp. 1–22. Retrieved August 28, 2010, from [http://www.ism.ac.jp/~ozaki/publications/paper/1999\\_Valdes\\_Bosch\\_nov\\_sci.pdf](http://www.ism.ac.jp/~ozaki/publications/paper/1999_Valdes_Bosch_nov_sci.pdf)
- Valdes-Sosa, P.A., 2004. Spatio-temporal autoregressive models defined over brain manifolds. *Neuroinformatics* 2 (2), 239–250. doi:10.1385/Nl:2:2:239
- Valdes-Sosa, P.A., Riera, J., Casanova, R., 1996. Spatio temporal distributed inverse solutions. In: Aine, C.J., Okada, Y., Stroink, G., Switthenby, S.J., Wood, C.C. (Eds.), *Biomag 96: Proceedings of the Tenth International Conference on Biomagnetism*, Volume 1. Springer, pp. 377–380.
- Valdés-Sosa, P.A., Hernández, J., Vila, P., 1996. EEG spike and wave modelled by a stochastic limit cycle. *Neuroreport* Retrieved August 21, 2010, from [http://journals.lww.com/neuroreport/Abstract/1996/09020/EEG\\_spike\\_and\\_wave\\_modelled\\_by\\_a\\_stochastic\\_limit.37.aspx](http://journals.lww.com/neuroreport/Abstract/1996/09020/EEG_spike_and_wave_modelled_by_a_stochastic_limit.37.aspx)
- Valdés-Sosa, P.A., Sánchez-Bornot, J.M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., et al., 2005. Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 360 (1457), 969–981. doi:10.1098/rstb.2005.1654
- Valdés-Sosa, P.A., Sánchez-Bornot, J., Vega-Hernández, M., Melie-García, L., Lage-Castellanos, A., Canales-Rodríguez, E., 2006. Granger causality on spatial manifolds: applications to neuroimaging. *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*, pp. 1–53.
- Valdes-Sosa, P.A., Sanchez-Bornot, J.M., Sotero, R.C., Iturria-Medina, Y., Aleman-Gomez, Y., Bosch-Bayard, Jorge, et al., 2009a. Model driven EEG/fMRI fusion of brain oscillations. *Hum. Brain Mapp.* 30 (9), 2701–2721. doi:10.1002/hbm.20704
- Valdés-Sosa, P.A., Vega-Hernández, Mayrim, Sánchez-Bornot, J.M., Martínez-Montes, E., Bobes, M.A., 2009b. EEG source imaging with spatio-temporal tomographic nonnegative independent component analysis. *Hum. Brain Mapp.* 30 (6), 1898–1910. doi:10.1002/hbm.20784
- Vanhatalo, S., Voipio, J., Kaila, K., 2005. Full-band EEG (FbEEG): an emerging standard in electroencephalography. *Clinical Neurophysiology*, 116(1). Elsevier, pp. 1–8.

- Retrieved August 29, 2010, from <http://linkinghub.elsevier.com/retrieve/pii/S1388245704003748>
- Victor Solo, 2008. Spurious causality and noise with fMRI and MEG. Organization for Human Brain Mapping annual Meeting.
- Von Mises, R., Doob, J.L., 1941. Discussion of papers on probability theory. *Ann. Math. Stat.* 12 (2), 215–217.
- White, H., Chalak, K., 2009. Settable systems: an extension of Pearl's causal model with optimization, equilibrium, and learning. *J. Mach. Learn. Res.* 10, 1–49.
- White, H., Lu, X., 2010. Granger causality and dynamic structural systems. *J. Financ. Econometrics* 8 (2), 193–243. doi:10.1093/jfifnec/nbq006
- Wiener, N., 1956. The theory of prediction. In: BeckenBach, E. (Ed.), *Modern Mathematics for Engineers*. McGraw-Hill, New York.
- Witt, S.T., Meyerand, M.E., 2009. The effects of computational method, data modeling, and TR on effective connectivity results. *Brain Imaging Behav.* 3 (2), 220–231. doi:10.1007/s11682-009-9064-5
- Wong, K.F.K., Ozaki, Tohru, 2007. Akaike causality in state space. Instantaneous causality between visual cortex in fMRI time series. *Biol. Cybern.* 97 (2), 151–157. doi:10.1007/s00422-007-0165-1
- Woodward, J., 2003. *Making Things Happen: A Theory of Causal Explanations*. Oxford University Press, Oxford, New York, Bangkok, Buenos Aires, Cape Town.
- Wright, S., 1921. Correlation and causation. *J. Agric. Res.* 20, 557–585.
- Zou, C., Denby, K.J., Feng, J., 2009. Granger causality vs. dynamic Bayesian network inference: a comparative study. *BMC Bioinformatics* 10, 122. doi:10.1186/1471-2105-10-122

## **Conclusiones**

1. Se demuestra que el uso de métodos de regresión penalizados permite extender el concepto de causalidad de Granger para el análisis de sistemas definidos sobre variedades espacialmente extendidos, como es el caso de las estructuras cerebrales.
2. Se generaliza así el mapeo paramétrico estadístico (SPM) de Neuroimágenes dinámicas para conectividades y no solo sobre activaciones (como era hasta ahora).
3. Se demuestra la utilidad de estos métodos para el estudio del funcionamiento cerebral espontáneo y durante tareas cognitivas utilizando tanto la resonancia magnética funcional (fMRI) como el registro concurrente de fMRI y electroencefalograma (EEG).

## **Recomendaciones**

1. Aplicar los métodos desarrollados para la evaluación de Neuroimágenes de pacientes neurológicos.
  
2. Continuar la comprobación de los métodos y determinar sus limitaciones con dos tipos de datos:
  - a. Obtenidas de simulaciones a gran escala de sistemas neurales realistas in silico.
  
  - b. Con datos experimentales estandarizados de animales como del siguiente tipo: registros intracraneales de los posibles generadores neurales, BOLD fMRI-, la superficie EEG, conectividad estructural basada en MRI de difusión y matrices de conectividad basadas en la histología.
  
3. Avanzar en los siguientes desarrollos teóricos en modelos causales para la conectividad efectiva:
  - a. La fusión de métodos de redes Bayesianas y los métodos basados en Causalidad de Granger (WAGS).
  
  - b. Las herramientas de WAGS desarrolladas de los procesos estocásticos de tiempo pueden jugar un papel importante en el análisis de la conectividad de EEG / fMRI, potenciales de campos locales y los datos de trenes de espigas.

c. Los métodos WAGS deben extenderse a modelos no estándar, entre ellos: los no markovianos, RDE, y las ecuaciones diferenciales con retardo así como para modelos de campos neurales