NeuroImage 54 (2011) 2695–2705

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Effective connectivity of fMRI data using ancestral graph theory: Dealing with missing regions

Lourens Waldorp [a,*], Ingrid Christoffels [b], Vincent van de Ven [c]

[a] Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands
[b] Department of Cognitive Psychology, Leiden Institute of Brain and Cognition, Leiden University, Leiden, The Netherlands
[c] Department of Cognitive Neuroscience, Maastricht University, Maastricht, The Netherlands

ARTICLE INFO

Article history:
Received 16 December 2009
Revised 21 August 2010
Accepted 18 October 2010
Available online 1 November 2010

Keywords:
Graphical model
Network analysis
Model selection
Random effects
Sandwich

ABSTRACT

Most of the current methods to assess effective connectivity from functional magnetic resonance imaging (fMRI) rely on the assumption that all relevant brain regions are entered into the analysis. If this assumption is untenable, which we believe is most often the case, then spurious connections between brain regions can appear. In this paper we propose to use an ancestral graph to model connectivity, which provides a way to avoid spurious connections. The ancestral graph is determined from trial-by-trial variation and not from the time series. A random effects model is defined for ancestral graphs which allows for individual differences in terms of graph parameters (e.g., connection strength). Procedures for model selection, model fit, and hypothesis testing of ancestral graphs are proposed. The hypothesis test can be used to find differences in connection strength between, for example, conditions. Monte Carlo simulations show that the ancestral graph is appropriate to model connectivity from fMRI condition specific trial data. To assess the accuracy further, the proposed method is applied to real fMRI data to determine how brain regions interact during speech monitoring.

Introduction

Effective connectivity between neuronal systems A and B has been defined as the direct influence of A on B (Büchel and Friston, 2000; Friston, 2007). In terms of deterministic systems this means that the inputs of B come directly from A, and do not go through any other variable. Several methods to estimate effective connectivity have been suggested: structural equation modeling (Mclntosh and Gonzalez-Lima, 1997; Buchel and Friston, 1997), dynamic causal modeling (DCM, Friston et al., 2003; Penny and Holmes, 2004) or its nonlinear extension (Stephan et al., 2008), dynamic Bayesian models (Rajapakse and Zhou, 2007), and Granger causality analysis with bivariate (Roebroeck et al., 2005) or multivariate time series (Eichler, 2005). There are many differences between these methods. For instance, DCM only takes instantaneous relations into account (because it is based purely on differential equations), whereas Granger causal models take both instantaneous and lagged relations into account. But a commonality of these methods is that they all use the time series to estimate effective connectivity. In functional magnetic resonance imaging (fMRI), however, spurious connections could result from analyzing the time series because the temporal resolution is too low (Eichler, 2005). Furthermore, except for the method of Eichler (2005), which was applied to simultaneous EEG and fMRI, each of these methods assumes that all relevant neuronal systems which could explain the connection between A and B are in the analysis. It is however very likely that there are neuronal systems not entered into the analysis, and that therefore some of the connections are in fact spurious (Eichler, 2005; Roebroeck et al., 2009). A neuronal system could be missing because, for example, it did not survive some statistical threshold, or it could be considered a priori irrelevant to the network.

In this paper we argue that a particular graphical model, an ancestral graph, can be used to account for neuronal systems not entered into the analysis when determining effective connectivity. In fMRI a graphical model is a representation of a joint distribution of several neuronal systems. We propose to estimate this joint distribution from the replications of condition specific trials and not from the time series. By using the replications instead of time series both instantaneous and lagged connections end up in the model. A major benefit of graphical models is that by considering a picture of a complicated multivariate situation, the underlying probability distribution is simultaneously considered (Whittaker, 1990; Cox and Wermuth, 1996; Lauritzen, 1996; Pearl, 2000; Edwards, 2003). An ancestral graph has three types of connections (Richardson and Spirtes, 2002) which we will give the interpretations of effective connectivity (influence), functional connectivity (correlation), and connectivity due to unobserved neuronal systems (unobserved common cause). Unobserved neuronal systems are therefore explicitly taken into account in an ancestral graph.

[*] Corresponding author. Department of Psychological Methods, University of Amsterdam, Roetersstraat 15, 1018WB, Amsterdam, The Netherlands. Fax: +31 20 639 0026.
E-mail address: waldorp@uva.nl (L. Waldorp).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.10.054

To test differences in strength of connectivity within a network or between conditions a random effects model is used like that in Beckmann et al. (2003), but transformed to variance parameters for the ancestral graph. In the random effects model, subjects are assumed to have different connection strengths originating from one population distribution.

First, ancestral graphs and the interpretation of its parameters in terms of connectivity are described. Second, estimation of the random effects parameters is discussed. Third, an existing algorithm (Zhang, 2008) to ascertain connections in an ancestral graph is presented, which together with a method to determine the best fitting model, makes it possible to ascertain which model represents the group network best. Fourth, a robust testing procedure is developed for testing between ancestral graph parameters. And finally, the ancestral graph model is applied to a data set on speech monitoring.

Ancestral graphs

We aim to model the connectivity of fMRI data by an ancestral graph. This is achieved by considering whether conditional independencies in the data are similar to those implied by the ancestral graph. The key, therefore, to comparing data and model is to determine the conditional independencies implied by the ancestral graph. First, the definition of an ancestral graph is given and it is described what the relation between a graph and a probability distribution is. Then the interpretation of the ancestral graph parameters is given when the assumed distribution is Gaussian.

Definition

A graph G consists of nodes in V and edges in E. Two examples are given in Fig. 1, in which each node could represent a brain region. The most popular graph is the directed acyclic graph (DAG), shown in Fig. 1(a). In this type of graph all edges are directional ($\rightarrow$) and there can be no cycle (Pearl, 2000). An example of a cycle is $1\rightarrow 2\rightarrow 3\rightarrow 1$. In Fig. 1(b) there are directed, undirected (-), and bidirected ($\leftrightarrow$) edges.

The graph G cannot contain an edge from a node to itself or have more than one edge between any two nodes. A node i is a parent of j whenever $i\rightarrow j$, i is a spouse of j whenever $i\leftrightarrow j$, and i is a neighbor of j whenever $i - j$. The set of nodes that are parents of i is denoted by $\mathrm{pa}(i)$, the set of nodes that are spouses of i is denoted by $\mathrm{sp}(i)$, and the set of nodes that are neighbors of i is denoted by $\mathrm{ne}(i)$. A sequence of edges with no repetitions of nodes is called a path. If there is a path like $i\rightarrow\cdots\rightarrow j$ with only directed edges pointing to j, or $i=j$, then i is an ancestor of j, denoted by $i\in\mathrm{an}(j)$. With these definitions, the definition of an ancestral graph can be given as follows (Drton and Richardson, 2004).

Definition ancestral graph

A graph $G=(V,E)$ with undirected, directed, and bidirected edges is an ancestral graph if for all $i\in V$

(i) $i\notin\mathrm{an}(\mathrm{pa}(i)\cup\mathrm{sp}(i))$;
(ii) if $\mathrm{ne}(i)\neq\emptyset$, then $\mathrm{pa}(i)\cup\mathrm{sp}(i)=\emptyset$

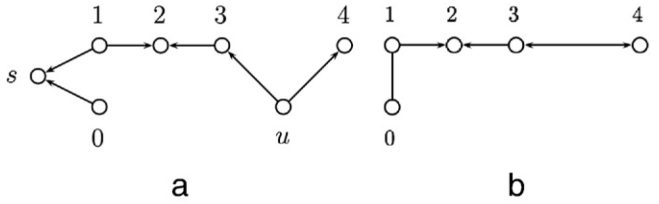

Fig. 1. In (a) a directed acyclic graph is shown (only directed edges and no cycles). In (b) an ancestral graph is shown based on the DAG in (a). There are three types of edges: directed, undirected and bidirected. The node u in (a) is unobserved resulting in a bidirected edge between nodes 3 and 4 in (b), and s in (a) is conditioned on resulting in a correlation between nodes 0 and 1 in (b).

The first condition says that there can be no cycles consisting of directed or bidirected edges. The second condition says that there can be no directional or bidirectional edges pointing to undirected edges. An example of an ancestral graph is given in Fig. 1(b).

One of the key elements of a graph is separation. When a probability distribution is compatible with a graph (see next subsection), then it possible to determine conditional independencies by considering separation in the graph. A collider at node i on a path involving i has two arrowheads at i, that is, $\rightarrow i\leftarrow$, $\rightarrow i\leftrightarrow$, $\leftrightarrow i\leftarrow$, or $\leftrightarrow i\leftrightarrow$. Two nodes a and b are m-separated by c if (i) on the path between a and b there is a collider node which is not c, or (ii) on the path between a and b there is no collider and c intercepts the path between a and b (Richardson and Spirtes, 2002). In Fig. 1(b) 0 is separated from 2 by 1, which illustrates condition (ii); and 1 and 3 are separated (without 2), which illustrates condition (i).

Probabilities and ancestral graphs

A probability distribution P can be associated with a graph G. By doing so, conditional independencies can be determined from separation in the graph. Let the nodes in V be identified with random variables Y_i for all $i\in V$ and let P be the joint distribution of all Y_i. Furthermore, let Y_A denote the set of Y_i such that $i\in A$, and $Y_A\perp\!\!\!\perp Y_B|Y_C$ denotes that Y_A is independent of Y_B when conditioned on Y_C. The connection between m-separation in graphs and conditional independence is given by the global Markov property (Richardson and Spirtes, 2002): Y_A is independent of Y_B given Y_C whenever A and B are m-separated by Y_C. Whenever this holds P is said to be globally Markov with respect to G. In Fig. 1(b) a probability distribution P is globally Markov with respect to G if the following conditional independencies hold: $Y_0\perp\!\!\!\perp Y_2|Y_1$, $(Y_0,Y_1)\perp\!\!\!\perp(Y_3,Y_4)$, and $Y_2\perp\!\!\!\perp Y_4|Y_3$. These conditional independencies are, of course, precisely the m-separations of the graph G.

The global Markov property indicates that a probability distribution P is compatible to G. This does not exclude, however, the possibility that another graph G_* is also compatible with P. In other words, there need not be a one-to-one correspondence between a probability distribution and a graph. Consider two configurations and their joint probability distributions

$$\begin{array}{ll} a\rightarrow c\rightarrow b & P(a,b,c)=P(a)P(c|a)P(b|c) \\ a\rightarrow c\leftarrow b & P(a,b,c)=P(a)P(b)P(c|a,b) \end{array} \tag{1}$$

Because their factorizations in terms of components are different due to the conditional independencies, it is possible to distinguish these two configurations only with information of the probability distribution (e.g., Pearl, 2000; Lauritzen, 1996). The best fit of the data to a specific probability distribution is therefore essential (see Section 3). However, there are configurations which are different but have identical components in the factorization (Verma and Pearl, 1991). For example, $a\rightarrow c\rightarrow b$ is equivalent in terms of probability distributions to $a\leftarrow c\rightarrow b$ with joint probability distribution $P(a,b,c)=P(c)P(a|c)P(b|c)$. They are equivalent because of the symmetry in conditional probability: $P(c)P(a|c)=P(a,c)=P(a)P(c|a)$. So, a collider gives rise to different conditional independencies, whereas other configurations do not. The key to differentiating between models is, therefore, determined by the set of colliders (Andersson et al., 1997).

In an ancestral graph the undirected and bidirected connections have an interpretation that relates DAGs to ancestral graphs (Richardson and Spirtes, 2002). Suppose that the node u in Fig. 1(a) is unobserved. Then the conditional independencies that result in the marginalized distribution are those that hold in Fig. 1(b) of the ancestral graph. Therefore, a bidirected connection can be interpreted as a connection resulting from an unobserved variable in a DAG. Suppose now that s in Fig. 1(a) is conditioned on. Then, again, the conditional independencies of the resulting distribution are those entailed by the ancestral graph in Fig. 1(b). And so, the undirected connection between 0 and 1 can be interpreted as the consequence of a variable (possibly unobserved) being conditioned on. In fact, any conditioning or marginalizing will always result in an ancestral graph (Richardson and Spirtes, 2002).

The property of ancestral graphs that marginalization and conditioning result again in an ancestral graph can be used in neuroimaging. Suppose that the node u in Fig. 1(a) represents a neuronal system that is not observed because, for example, it is below the multiple comparison threshold. The result of the unobserved u is that it is marginalized over. In the ancestral graph in Fig. 1(b) this results in the bidirected connection between 3 and 4. Similarly, if s in Fig. 1(a) is unobserved but constant then the undirected connection between 0 and 1 results. In both cases the resulting graph is still an ancestral graph.

Ancestral graphs and connectivity

The interpretation of the three different types of edges between nodes in an ancestral graph has a simple statistical interpretation when it is assumed that the joint distribution of all variables is Gaussian (Richardson and Spirtes, 2002). Functional connectivity will be associated with an undirected edge i-j, effective connectivity will be associated with a directed edge $j \rightarrow i$, and a connection due to an unobserved neuronal system will be associated with a bidirected edge $i \leftrightarrow j$.

The parameterization associates each edge and node of the graph with a parameter of the Gaussian distribution. An undirected connection i-j is obtained whenever the parameter from the (positive definite) matrix $\Lambda = (\lambda_{ij})$ is $\lambda_{ij} \neq 0$. A directed connection $j \rightarrow i$ is obtained whenever the parameter from the lower diagonal matrix $B = (\beta_{ij})$ is $\beta_{ij} \neq 0$. And finally, a bidirected connection $i \leftrightarrow j$ is obtained whenever the parameter from the (positive definite) matrix $\Omega = (\omega_{ij})$ is $\omega_{ij} \neq 0$. The variance matrix Σ of the p variables with joint Gaussian distribution modeled by an ancestral graph by is (Richardson and Spirtes, 2002)

$$\Sigma = B^{-1}\begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & \Omega \end{pmatrix}\left(B^{-1}\right)' \tag{2}$$

The interpretation of the three types of parameters each associated with its own connection in the graph is now straightforward. First, the matrix Λ is the inverse of the covariance matrix (precision matrix) for the undirected part of G. That means that λ_{ij} associated with an undirected edge i-j, can be interpreted as a partial covariance between i and j, that is, the covariance between i and j with all other variables in the undirected part of G conditioned on. Second, the parameter β_{ij} associated with a directed edge $j \rightarrow i$, is the regression coefficient for variable j in the regression of i. This is similar to the interpretation of the parameters of a DAG with a Gaussian distribution. And third, the parameter ω_{ij} associated with a bidirected edge $i \leftrightarrow j$, is the covariance between residuals e_i and e_j where

$$e_i = Y_i - \sum_{j \in \mathrm{pa}(i)} \beta_j Y_j. \tag{3}$$

The covariance between residuals has been associated with unobserved common causes before (see e.g., Andersson et al., 1997). In Fig. 2 the statistical interpretation of the edges of an ancestral graph is given. The number of parameters q is the sum of the number of nodes p in V and the number of connections in E.

To obtain estimates of these parameters, replications of observed responses to each of the nodes is required (see Section 4 on how these can be obtained). From these replications a joint probability distribution is determined combined with assumption of normality for the ancestral graph. In our application in Section 4, general linear model parameters associated with a region of interest (ROI) are estimated for each replication and are then used to obtain a probability distribution for those ROIs. For purposes of testing differences between connections the variance of the parameters is also required. Even if the model is incorrect, statistical inference should be possible. These issues are addressed in the next section.

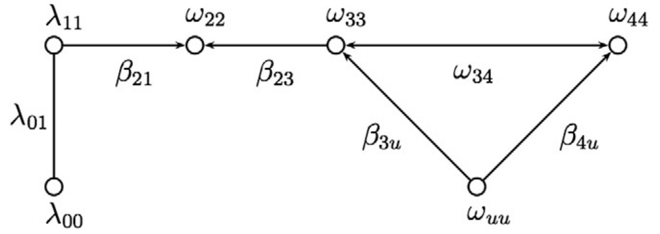

Fig. 2. Example of an ancestral graph showing the interpretation of the edges in terms of the statistical parameters of an ancestral graph.

Random effects model

We follow Henderson (see e.g., Robinson, 1991) for the derivation of the random effects model. A joint density for the data and random effects is defined. The random effects are then estimated by maximizing this density with respect to the effects and population parameters, that is, estimates are obtained by the method of maximum likelihood. The random effects can be interpreted as if the parameters of each subject are randomly drawn from a normal distribution with unknown mean and variance (Verbeke and Molenberghs, 2000; Beckmann et al., 2003).

Estimation

Let $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in})$ be the $p \times n$ matrix of observed responses on p regions and n independent replications (trials) of subject $i = 1, \ldots, N$. It is assumed that for each subject i and each replication j the $p \times 1$ vector y_{ij} is normally distributed with mean zero (possibly after subtraction of the mean) and $p \times p$ variance matrix Σ_i. The matrix Σ_i is modeled for each subject i by an ancestral graph with parameters in B_i, Λ_i, and Ω_i. The unique parameters for subject i are collected in the $q \times 1$ vector θ_i. We also assume that the parameters in θ_i are normally distributed with unknown mean μ and $q \times q$ variance matrix Ψ. We assume additionally that each subject has its own true value θ_{0i}, since we have n replications of each subject (see the Appendix for more details).

The parameters of interest are the population mean μ and variance matrix Ψ. The individual effects θ_i for all N subjects are sometimes of interest. The parameters and effects are estimated by maximum likelihood (ML). ML-estimates are obtained from the joint density of y and θ_i, denoted by $f(y, \theta_i; \mu, \Psi)$. Let $S_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij} y_{ij}'$ be the $p \times p$ sample variance matrix for subject i with $n > p$ and let $\Sigma_i = \Sigma(\theta_i)$. The logarithm of the joint density of the data and the random effects, denoted by $L(\mu, \Psi; y, \theta_i) = \log f(y, \theta_i; \mu, \Psi)$, is

$$L(\mu, \Psi; y, \theta_i) = -\frac{N(n+q)}{2}\log(2\pi) - \frac{n}{2}\sum_{i=1}^{N} \log|\Sigma_i| - \frac{n}{2}\sum_{i=1}^{N} \mathrm{tr}\left[S_i \Sigma_i^{-1}\right] - \frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_{i=1}^{N} (\theta_i - \mu)' \Psi^{-1} (\theta_i - \mu). \tag{4}$$

To obtain ML-estimates the maximum of $L(\mu, \Psi; y, \theta_i)$ is required (see e.g., Demidenko, 2004; Robinson, 1991). The parameter estimates of μ and Ψ are easily obtained as

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \theta_i \quad \text{and} \quad \hat{\Psi} = \frac{1}{N}\sum_{i=1}^{N} (\theta_i - \mu)(\theta_i - \mu)' \tag{5}$$

Of course, the individual effects in θ_i are required for these estimates. In the Appendix it is shown that the ML-estimate of θ_i obtained from the conditional log-likelihood $L_c(\theta_i;\mu,\Psi,y)$ is consistent, which results in consistent estimates of μ and Ψ. This means that only the first part of the log-likelihood is required for the estimation of the subject specific effects in θ_i. These subject specific effects are obtained from the iterative conditional fitting (ICF) algorithm by Drton and Richardson (2004) implemented in R. The ICF algorithm makes use of the independence of the undirected and the directed part of the ancestral graph, such that parameters in Λ_i are obtained directly, and parameters in B_i and Ω_i are obtained by fitting a conditional distribution with one node removed, iterated over all nodes in the directed part until convergence. We refer to Drton and Richardson (2004) for a full explanation of ICF.

The estimate of Ψ, the population variance matrix, could be improved to also be able to estimate reliably the individual variance matrix of the parameters. This can be achieved by using the asymptotic representation of the estimate $\hat{\theta}_{ni}$ of the individual parameters.

$$\hat{\theta}_{ni} = \theta_{0i} + \frac{1}{n}\sum_{j=1}^{n} H(\theta_{0i})^{-1} J(y_{ij};\theta_{0i}) + o_p(n^{-1/2}), \tag{6}$$

where $J(y_{ij};\theta_{0i})$ is the $q\times 1$ vector of first-order partial derivatives of the log-likelihood for observation y_{ij} evaluated at θ_{0i}, the true value of subject i, and $H(\theta_{0i})$ is the $q\times q$ matrix with second-order partial derivatives of the log-likelihood. This representation can be used in the estimate $\hat{\Psi}$ in Eq. (5) to obtain

$$\tilde{\Psi} = \frac{1}{Nn}\sum_{i=1}^{N} H(\theta_{0i})^{-1}\left(\frac{1}{n}\sum_{j=1}^{n} J(y_{ij};\theta_{0i})J(y_{ij};\theta_{0i})'\right)H(\theta_{0i})^{-1}. \tag{7}$$

This estimate is similar to the sandwich (e.g., White, 1982; Waldorp, 2009) in that it has the second-order derivatives outside of the product of first-order derivatives. If it is assumed that the model is correct then the product of first-order derivatives is equal to the Hessian, and only the inverse of the Hessian remains (Van der Vaart, 1998; Waldorp et al., 2005a). The estimate $\tilde{\Psi}$ also shows that the sandwich can be defined for each subject separately based on n replications, which is a single element of the sum over N subjects. Since these results are derived for large samples, it is of great interest to see what happens when the sample sizes are small to moderate.

Monte Carlo simulations

To show that the estimates of the variance in $\tilde{\Psi}$ are accurate, we perform Monte Carlo simulations. We show the performance of the estimator obtained in (7), referred to as the sandwich, and compare it to a more traditional estimate where only $H(\theta_{0i})$ is used (see e.g., Van der Vaart, 1998). We compute both types of variance parameters when there are one or more nodes missing.

To generate data we used the graph of Fig. 3 with all nodes observed and using linear regressions. In equations, this model is

$$\begin{aligned} Y_0 &= e_0 & Y_3 &= \beta_{3u}Y_u + e_3 \\ Y_1 &= \beta_{10}Y_0 + e_1 & Y_4 &= \beta_{4u}Y_u + e_4 \\ Y_2 &= \beta_{21}Y_1 + \beta_{23}Y_3 + e_2 & Y_u &= e_u \end{aligned} \tag{8}$$

where $\beta_{ij}=0.5$ for all i,j, and the e_i are independent Gaussian white noise (see Eichler, 2005, for similar settings). We used two different models to estimate parameters. Model 1 is shown in Fig. 3(a), where only node u is unobserved. Model 2 has in addition to unobserved node u also unobserved nodes 0 and 1, shown in Fig. 3(b). In Fig. 4 the ratio of estimated variance to true variance (obtained from 500 simulations) for sample sizes ranging from 10 to 100, is displayed for the directional edge 2←3, that is the variance of parameter β_{23}. In model 1 the Hessian has a ratio closer to one at low sample sizes than the sandwich, indicating that at lower sample sizes the Hessian estimates the variance better. From a sample size of 30 both the Hessian and sandwich are accurate. In model 2, however, the Hessian can be seen to overestimate the variance for any number of replications. In contrast, the sandwich seems to settle on the correct variance from about 20 replications.

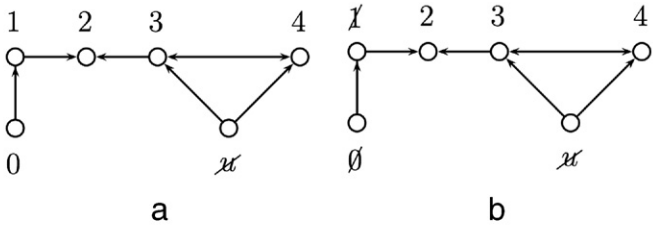

Fig. 3. Model 1 is shown in (a) where u is unobserved (signified by the strike through), resulting in the bidirected connection between nodes 3 and 4. In (b) model 2 is shown with nodes 0, 1, and u unobserved. The data generating model is with all nodes observed.

Model selection and model fit

There are two issues to consider in determining connectivity: which model best represents the group of subjects and does the model fit each subject. Determining the model that best represents the group is in line with a random effects model. Individual fits can be

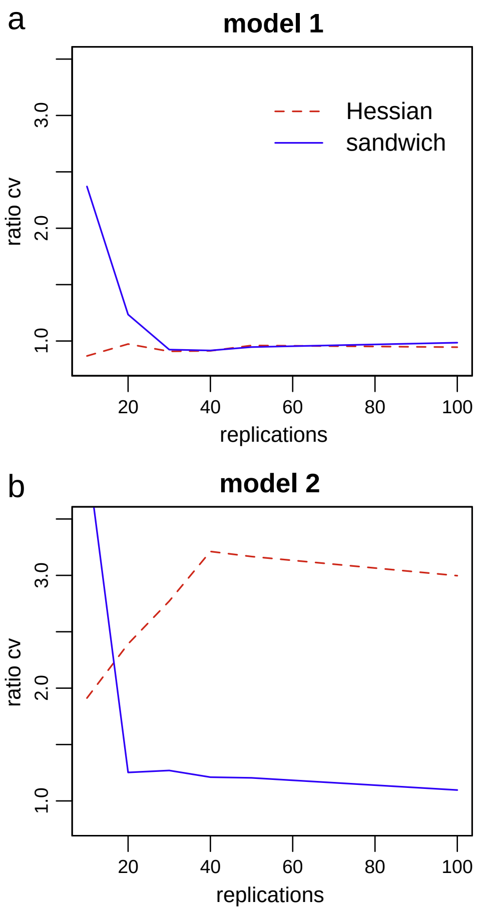

Fig. 4. Ratio of estimated to true variance of the parameter β_{23} for the Hessian (red and dashed) and sandwich (blue and continuous) methods. When the ratio is 1 the variance estimate is good. In (a) for model 1 and in (b) for model 2.

important to determine what differences there are between subjects in terms of networks. Both issues are addressed.

In modeling networks the number of models increases exponentially (Whittaker, 1990; Marrelec et al., 2006). In the case of ancestral graphs with p regions there are at least $3^{p(p-1)/2}$ possible models but always less than $4^{p(p-1)/2}$, because there are three types of edges or no edge. So with $p=4$ regions, the number of models is between 729 and approximately 4000. The problem is that this is a huge number of models to select from. There are two ways to approach this issue: (i) use predefined regions, hypotheses, to model connectivity, as in DCM, or (ii) perform an automated search. We use a semi-automated search combined with a local search to obtain the best maximal ancestral graphs. The automated search is obtained from an algorithm described in an excellent paper by Zhang (2008) and also in Spirtes et al. (1993). The algorithm searches for the most informative ancestral graph that can be obtained given the probability distribution, referred to as the fast causal inference (FCI) algorithm (we used the original FCI algorithm). The result is not necessarily a maximal ancestral graph. Therefore, a second stage is required to obtain a maximal ancestral graph. We refer to Zhang (2008) for a complete description of the FCI algorithm.

The FCI algorithm is implemented in R using the SIN method by Drton and Perlman (2004) to take into account the multiple comparison problem when determining conditional independencies. In this procedure the edge set E of G contains only undirected connections, and hence the covariance matrix is modeled by Λ, the inverse of the covariance matrix containing partial covariances. In SIN the null hypothesis $H_{0,ij}:\lambda_{ij}=0$ is tested for each of the $p(p-1)/2$ possible connections for p nodes while controlling for multiple comparisons (Drton and Perlman, 2004). The tests on conditional independencies are used in the FCI algorithm to decide whether edges should be in the set of edges in the graph for that subject. In the second stage several ancestral graphs that are minor variations of the FCI output are compared. These ancestral graphs are compared using a score function.

The score function we use is the Akaike information criterion (Akaike, 1973). We use the Akaike information criterion (AIC) because it fits well with the possibility of missing regions (and hence connections). The AIC is predicated on the notion that the true model is not in the set of possible models, but is in theory attainable asymptotically (Bozdogan, 2000). But due to our lack of information (limited number of observations), the true model is out of reach. The AIC is defined for ancestral graph model G_q with q parameters as (Akaike, 1973)

$$\mathrm{AIC}(G_q) = -2L(\theta_i, \mu, \Psi; y) + 2q \tag{9}$$

The main aim of the AIC is to keep in balance the bias and variance of the estimated parameters (Casteren and Gooijer, 1997). Of course, changing direction in an ancestral graph will not change the number of parameters and so the penalty will remain the same. But if the log-likelihood decreases because of the conditional probabilities in correspondence with the change of direction, then the AIC will indicate this. The AIC can be used to determine the best model at the group and at the individual level.

That the model is best according to the AIC does not entail that it fits, as compared to the saturated model. Together with the AIC we therefore require a model fit procedure to obtain the best fitting model for all subjects. Many fit procedures exist (see e.g., Claeskens and Hjort, 2008). However, our objective is to fit a model that is approximately correct (se e.g., Waldorp et al., 2005b). To that end, a modification of the likelihood ratio test is used. The likelihood ratio (LR) test has asymptotically a chi-square distribution when the model is correct (Young and Smith, 2005). With approximate models, however, the LR rejects the model more often than expected.

Let the null hypothesis be $H_0: \theta \in A \subset \mathbb{R}^q$ and the alternative $H_A: \theta \notin A$, which corresponds to using the unrestricted model with the variance matrix S. The LR test is defined for the ancestral graph as

$$\lambda_A = log \frac{max_{\theta \in A} L}{max_{\theta \notin A} L}. \tag{10}$$

As is seen below, when an ancestral graph is fitted to data with an unobserved node u (as in Fig. 3(a)), then the LR test rejects H_0 more often than a level of 5%, say. However, an ancestral graph with a bidirected edge represents a good approximation. Therefore, we would like to accept the model in H_0 when the ancestral graph represents missing nodes as bidirected edges. In order to account for such approximations, we use a modified version of the LR test. This modification was proposed as an improvement to a Bartlett-type correction for small samples (Yuan and Bentler, 1997). But the modification is of the order $O_p(n^{-1})$, and so can be considered as a small correction towards the null hypothesis depending on the size of LR. The modification was proposed for least squares type estimators, but Browne (1974) has shown that the difference between the LR and least square type estimator is asymptotically zero. The modified version is defined as

$$T_A = \frac{\lambda_A}{1 + \lambda_A / n}. \tag{11}$$

The test T_A has asymptotically a chi-square distribution with $p(p+1)/2-q$ degrees of freedom, where p is the number of nodes and q is the number of parameters of the ancestral graph. Because this is still an asymptotic result, Monte Carlo simulations are required to determine the behavior in small samples.

Monte Carlo simulations for model selection and model fit

To show the performance of the model selection with the AIC and model fit with the LR and T_A procedures, we consider the two models of the previous section shown in Fig. 3. The performance measure for model selection is percentage of correct decisions using the AIC, and for model fit the measures are false positive rate (FPR) and power of T_A. The FPR refers to the probability of rejecting a model when in fact it is a good approximation. And power is the probability that the model is rejected given that the graph is a poor approximation.

Data are generated by the ancestral graph of Fig. 3(a). To evaluate performance of the AIC, the percentage of correct decisions by the AIC was computed for two competing ancestral graphs for each of the two models. The competing graphs were the correct one with the edge from $2 \leftarrow 3$ and the incorrect one with the edge reversed to $2 \rightarrow 3$. Fig. 5 shows that the AIC is quite accurate for model 1 with only one unobserved node but less accurate when there are many unobserved nodes, as in model 2.

To evaluate model fit we used a common significance level of 5%. A correct model fit procedure should reject the correct model no more than 5%, that is, the false positive rate (FPR) should not exceed 0.05. In model 1 there are $5(5+1)/2-9=6$ degrees of freedom, which results in a threshold value of $\chi^2_6(0.05)=12.592$. For model 2 there is $3(3+1)/2-5=1$ degree of freedom, which leads to a threshold of $\chi^2_1(0.05)=3.841$. To compute the power the same modification was applied to both models 1 and 2, in which the edge $2 \leftarrow 3$ was changed to $2 \rightarrow 3$.

For model 1 in Fig. 6(a) it is clearly seen that the FPR of the LR is too high at lower number of replications. The modified version T_A remains below the 5% level for all n. Both the LR and T_A converge to the 5% level. The power for model 1 in Fig. 6(c) is higher for the LR than for T_A. This was expected because T_A is a smaller version of LR, making it more difficult to reject a graph with T_A than with LR. For model 2 the FPR of the LR is only slightly too high (Fig. 6(b)), and for T_A remains

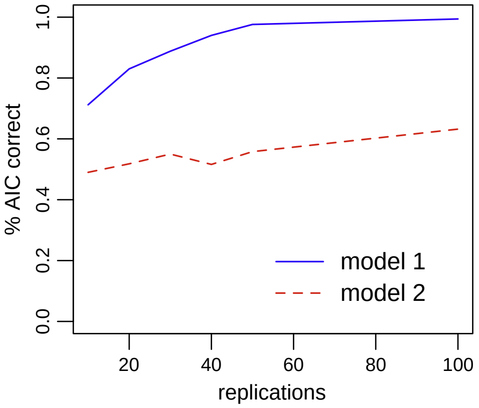

Fig. 5. Percentage of correct decisions by the AIC for the two competing graphs. The percentages are given for model 1 (continuous and blue) and 2 (dashed and red) from Fig. 3.

around 0.05 for all number of replications. The power of the LR and T_A in Fig. 6(d) is about the same for model 2 and does not exceed 0.20.

In conclusion, we can say that when the model is reasonably close to the data generating ancestral graph, then it is likely that the best fitting model can be found. If, however, the model has many unobserved nodes, then it is much more difficult to find the best fitting model.

Testing parameters

When a model has been selected, parameters of the model can be tested. We follow general methods for random effects analysis (e.g., Beckmann et al., 2003; Verbeke and Molenberghs, 2000; Penny and Holmes, 2004)). The parameters of the ancestral graph model for each subject are collected in θ_i for $i=1,\ldots,N$. All subjects can be concatenated into a single $Nq\times 1$ vector θ. Then a contrast C can be constructed to test the null hypothesis $H_0: C\theta = u$, where u is a constant, often zero. The variance of the estimate $C\hat{\theta}_n$ is then $C\tilde{\Psi}_b C'$, where $\tilde{\Psi}_b$ is the $Nq\times Nq$ block diagonal matrix $\text{diag}\left(\tilde{\Psi}_1,\ldots,\tilde{\Psi}_N\right)$. We can then construct a Wald test (Young and Smith, 2005; Waldorp, 2009)

$$W = \frac{n-k}{nk}(C\hat{\theta}_n - u)'(C\tilde{\Psi}_b C')^{-1}(C\hat{\theta}_n - u), \tag{12}$$

where n is the number of replications and k is the number of independent contrasts in C. This test is approximately F-distributed with degrees of freedom k and n-k under H_0. This result is based on two approximations: (i) the estimate $\hat{\theta}_n$ is locally asymptotically normal, and (ii) in the first-order partial derivatives of the likelihood, the difference between the sample and modeled variance matrix is asymptotically distributed as normal (see Appendix for details). A simple example of the Wald test is the test on the average across subjects (Beckmann et al., 2003). We average the estimate of both the parameters and the variance across subjects to obtain estimates of the

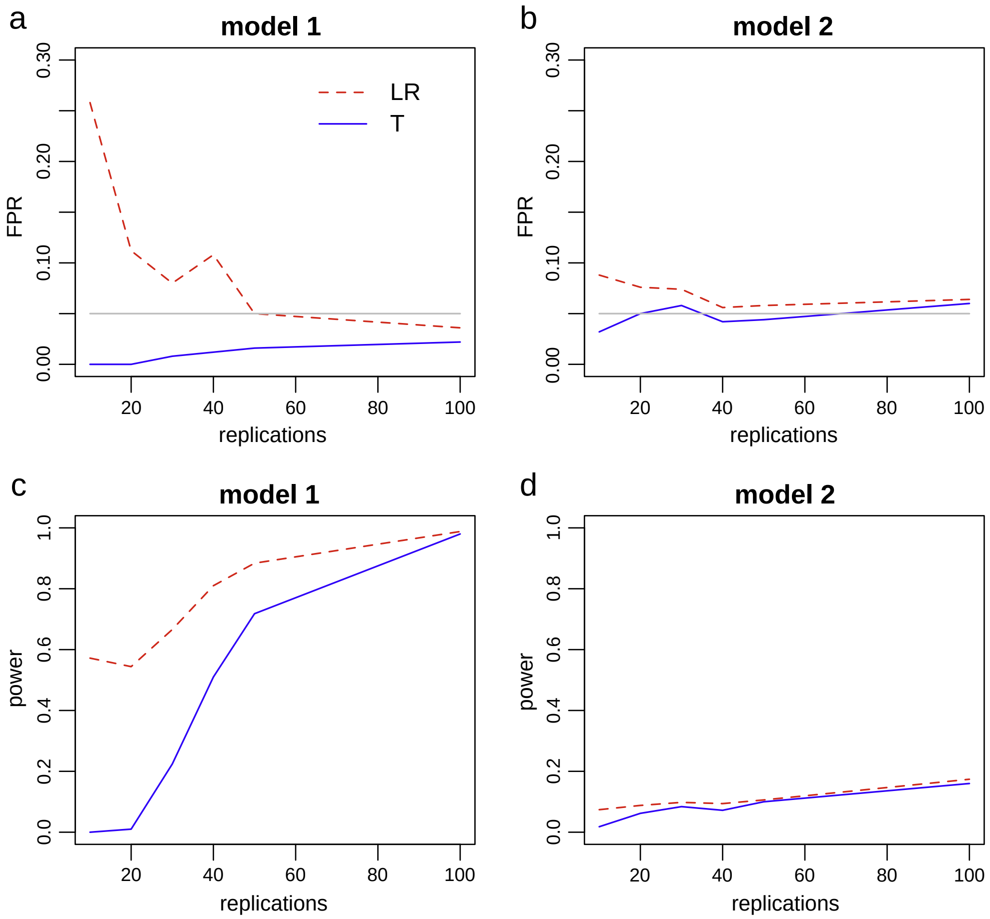

Fig. 6. False positive rate (FPR) and power of the fit procedures. The FPR (a and b) and the power of the fit (c and d) of model 1 (left panels) and model 2 (right panels). Model 1 is shown in Fig. 3(a), and model 2 is shown in Fig. 3(b).

population mean and variance matrix as in Eq. (5). Then we can construct a simple contrast c' to test the difference between two average parameters, and obtain $c' = (1, -1, 0, \ldots, 0)$, for example. We then have

$$W_c = \frac{Nn-1}{Nn} \frac{(c'\hat{\mu})^2}{c'\bar{\Psi}c}. \tag{14}$$

This test is applied to real data in the next section.

Monte Carlo simulations for the Wald test

To determine whether the asymptotic approximations of the Wald test are adequate, Monte Carlo simulations were performed. Model 1 (Fig. 3(a)) from the previous sections was used to evaluate the false positive rate (FPR) and power of the Wald test on averages in Eq. (13). The null hypothesis $H_0: \beta_{01} = \beta_{21}$ was tested. For the FPR the null hypothesis was true; to compute the power the values of the parameters were set to $\beta_{01} = 0.2$ and $\beta_{21} = 0.7$. In Fig. 7 it can be seen that the FPR is approximately at the nominal level of 5% and that the power increases quickly with the number of replications, and is high with about 50 replications. In conclusion, the Wald test appears to be valid and has reasonable power at moderate sample sizes.

Application to real data

An ancestral graph combined with a random effects model appears to be appropriate theoretically. But, of course, it is essential to determine whether the results of the method actually make sense using real data. In this section we describe how an ancestral graph can be obtained and how its parameters can be tested using data from an fMRI experiment on speech monitoring.

One objective of the study was to determine which brain regions are involved in speech monitoring and how these regions interact (Christoffels et al., submitted). The task was to name pictures out loud while noise masked the self-produced speech (similar to Christoffels et al., 2007). The noise masking was varied with four levels, from no noise to loud noise. A blocked design was used to implement the different conditions. It was predicted that as the level of the noise mask increased, the attenuation in the superior temporal gyrus decreased due to a cancelation process.

A contrast of the main effect of speaking with and without noise was tested with a random effects analysis of the general linear model (GLM) implemented in BrainVoyager QX (Goebel et al., 2006), and was thresholded at a false discovery rate of 0.05. The analysis included functional imaging data of $N = 11$ subjects. We selected five regions from a random effects result: supplementary motor area (SMA), insula (INS), cerebellum (CB), superior temporal gyrus (STG), and anterior cingulate cortex (ACC). For these regions we obtained the standardized amplitude parameters of the GLM for each of 12 replications in all four noise masking conditions (for more details, see Christoffels et al., submitted). Since it seems reasonable to assume that the network consists of the same set of regions in each of the conditions, we obtained $n = 48$ replications for each subject.

The assumption of normality for the ancestral graph can be investigated for each region by comparing the quantiles of the sample and the quantiles of the normal distribution. The comparisons for INS and SMA are shown in Fig. 8 for one subject. The figure shows that the observations from the sample with $n = 48$ are all reasonably close to the line of expected quantiles of the normal distribution. This indicates that the GLM parameters are approximately normally distributed.

The two stage procedure to find the ancestral graph starts with the FCI algorithm (Zhang, 2008) using the SIN approach (Drton and Perlman, 2004) to determine connections. Following Drton and Perlman (2004) the algorithm was used with a nominal significance level of 0.3, which made it easier for connections to enter the model (see Drton and Perlman, 2004, for a discussion on this). In Fig. 9 the plots for all 11 subjects are shown with the connections derived from FCI for each subject separately (the connections from FCI corresponded almost exactly to the SIN procedure for undirected graphs with all regions entered). From the FCI output a connection was thought to be relevant at the group level if at least three (an arbitrary

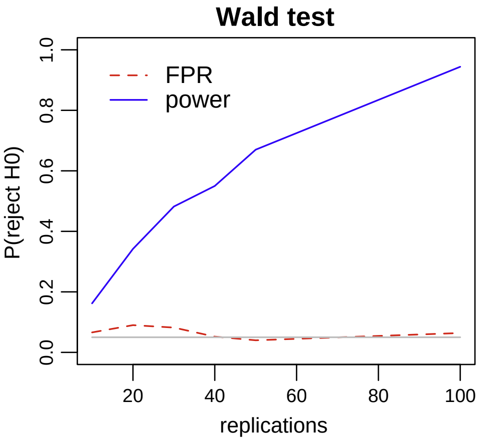

Fig. 7. Probability to reject the null hypothesis conditional on either the null being true (FPR, red, dashed line) or conditional on the null being false (power, blue, continuous line). The gray line at 0.05 is the nominal significance level.

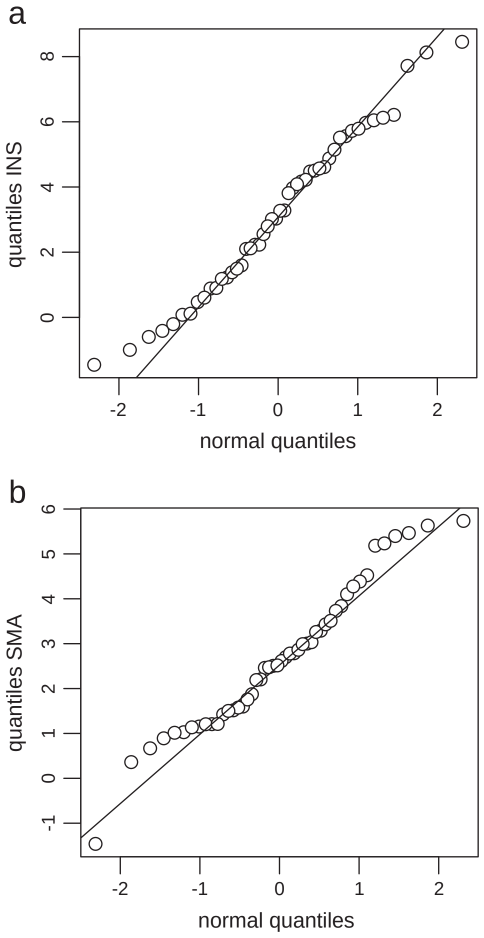

Fig. 8. Comparison of sample and theoretical quantiles (qq-plot) based on $n = 48$. In (a) the comparison for INS and in (b) the comparison for SMA for one subject. The line indicates where the observations (open circles) are expected to be according to the theoretical normal distribution.

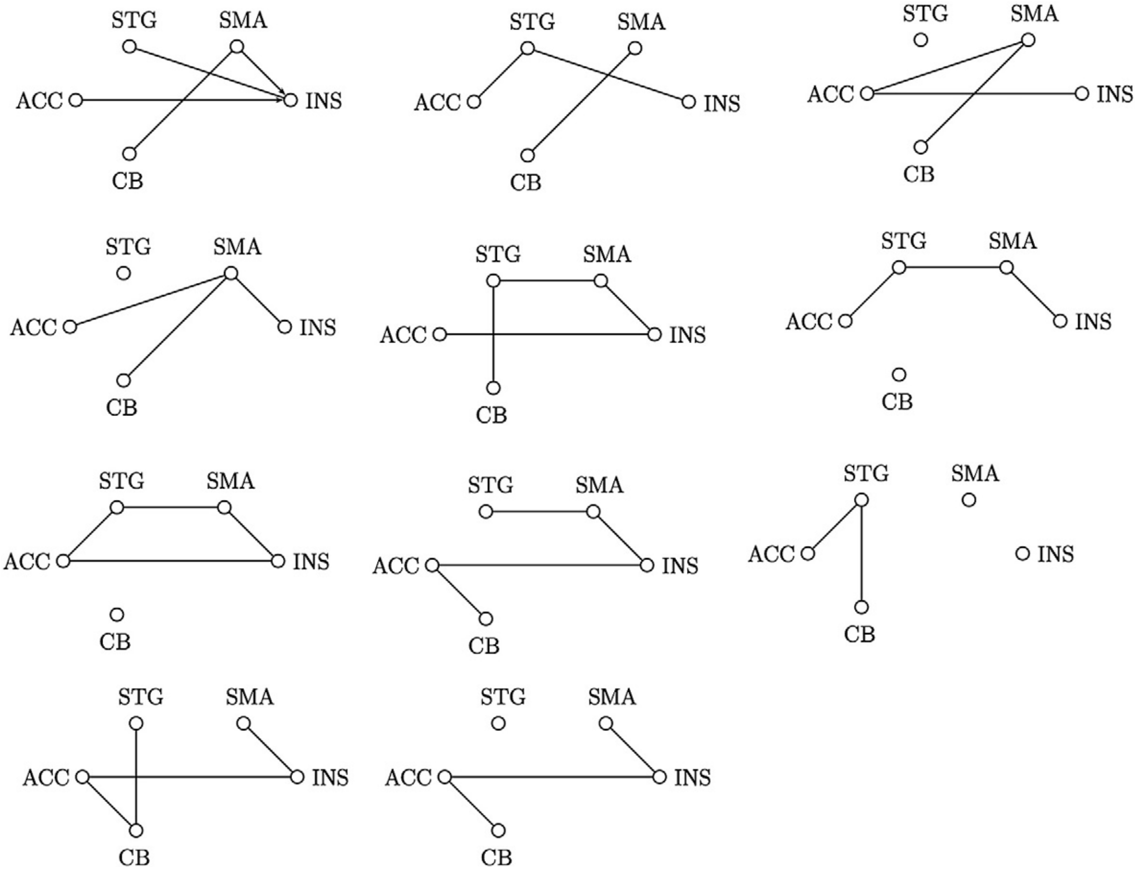

Fig. 9. Output of the FCI algorithm in combination with SIN for each of the 11 subjects. Note that only subject 1 has directed connections to the INS. Subjects are ordered from left to right and top to bottom.

choice, which becomes irrelevant because of subsequent group modelling) subjects had the connection. This resulted in the connections INS-SMA, STG-ACC, STG-CB, STG-SMA, and ACC-INS. From these five connections three models containing undirected, directed and bidirected connections were generated. Model A is shown in Fig. 10. Model B has one additional connection CB→STG and an undirected connection replaced by SMA→STG. Model C has two additional connections compared with A, STG→INS and STG→ACC. For these models AIC values were computed and compared at the group level. From Table 1 it can be seen that model A had the lowest AIC value and is to be preferred at the group level. To investigate individual differences, AIC values were also compared at the individual level, also shown in Table 1. It can be seen that 9 out of the 11 subjects had the lowest AIC value for Model A.

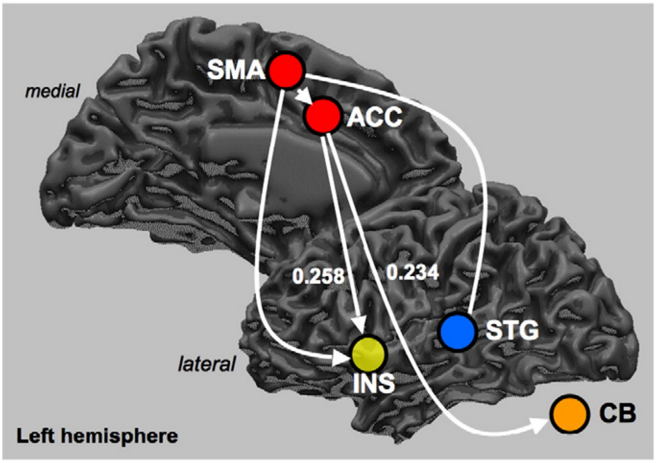

Fig. 10. Ancestral graph model A represented on the brain. The model nodes (regions) and paths are superimposed on a three-dimensional inflated representation of the cortical sheet of the left hemisphere. The figure shows the lateral and medial views of the same hemisphere. Note that the connection between SMA and STG is undirected. The transparent yellow of the INS refers to the fact that the INS is behind the opercula. For explanations of abbreviations, see text.

To make sure that there were no missing regions which could have caused some of the connections, bidirectional edges were used to see if the fit improved. For none of the edges did a bidirectional edge improve the fit. This means that model A is the best candidate for nine subjects and that it is unlikely that there are spurious connections.

Finally, T_A was used to determine that for the same nine subjects the model fits. In Table 1 it can be seen that model A fits the same nine subjects as the AIC indicated was the best model. Interestingly, for subject 2 model B was best according the AIC but no model fits according to T_A. The combination of the AIC and T_A provides evidence that the network is similar for nine subjects. These nine subjects can therefore be used in the random effects analysis to test for differences in connection strengths.

An interesting comparison to make is between the connections from the ACC to either the INS or the CB. A stronger connection from the ACC to either to INS or CB means that one of the connections is more dominant with respect to the other. To test for a difference in connection strength, the Wald test was used, introduced in Section 3.3. This is the test across averaged parameters and variances given in the example. The null hypothesis states that in the population there is no difference between $\beta_{ACC\rightarrow INS}$ and $\beta_{ACC\rightarrow CB}$. The connection strengths of the average across the 9 subjects are $\overline{\beta}_{ACC\rightarrow INS}=0.258$ and $\overline{\beta}_{ACC\rightarrow CB}=0.234$ (multiplied by -1 for the usual regression interpretation). Then the Wald test is $W=0.0056$ with degrees of freedom 1 and $Nn-1=9(48)-1=431$

Table 1
Model comparisons for models A, B, and C using the AIC at the group level in the first row and at the individual level in the second row. In the third to fifth rows are the fits of the three models where dot indicates whether model A, B, or C fits according to T_A at level 0.05.

Model		
A	B	C
−7819.548	−7625.408	−7592.915

	Subjects										
	1	2	3	4	5	6	7	8	9	10	11
AIC min{A, B, C}	A	C	A	A	A	A	A	A	B	A	A
Fit T_A(A	•		•	•	•	•	•	•	•	•	•
Fit T_A(B)	•		•		•			•	•	•	•
Fit T_A(C)	•		•		•			•	•	•	•

and p-value $p = 0.940$. This result indicates that there is no evidence against the hypotheses that on average there is no difference in strength between the two edges.

Conclusion

Assessing effective connectivity relies on methods that take into account that all brain regions which influence the network are in the analysis. If this assumption is untenable, which we believe is most often the case, then spurious connections between brain regions can result. The framework we proposed using ancestral graphs provides a way to avoid spurious connections. An ancestral graph can distinguish between effective and functional connectivity on the one hand and unobserved causes on the other hand.

The simulations showed that an ancestral graph provides a good approximation when some of the connections or some of the brain regions are not in the analysis. The estimation of parameters and model selection work because omitting regions still yields an ancestral graph, a property specific to ancestral graphs. The analysis of the speech monitoring data showed that it is possible to select a model and test specific hypotheses about the network using the random effects model.

Any method relies on the tenability of its assumptions. The assumptions of the framework presented here are (i) the ancestral graph is a reasonable approximation to the network, and (ii) the data obtained from the brain regions should be approximately normal. As to the first assumption, the simulations showed that when the ancestral graph had most of the nodes and connections of the underlying network, then it is useful to assess connectivity. When many of the nodes and connections are missing, assessing effective connectivity becomes more difficult. The second assumption about normality was investigated using the data on speech monitoring. It appeared that the coefficients from the GLM were indeed approximately normally distributed. It is at present unclear how severe the consequences are when this assumption is violated. However, when many voxels are used to obtain a single random variable to represent a brain region (e.g., an average over voxels of GLM β-coefficients), then according to the central limit theorem, it is reasonable in general to assume that this parameter is approximately normally distributed. Furthermore, it has been found that a contrast of two fMRI images is almost indistinguishable from a normal distribution (Wink and Roerdink, 2006).

The selection of the network depends to a great extent on the criterion used to determine the best network for the data. Here we used the AIC because it is in line with the idea of ancestral graphs that some nodes may be missing and because the AIC is relatively straightforward to implement. However, the AIC is a very general procedure to select a model and is known to be inconsistent (Burnham and Anderson, 2004; Grunwald, 2000). Currently, we are improving the selection procedure by considering a problem specific method to determine network connectivity (Claeskens and Hjort, 2008). In connection with this model selection, we are also working on improving the current version of the FCI algorithm to include all rules to make the procedure consistent (Zhang, 2008). Furthermore, we are working on large scale networks (e.g., (Valdés-Sosa et al., 2005)) in combination with ancestral graphs. When large scale networks can be analyzed using ancestral graphs, it may be better possible to combine structural and functional information in a single graph.

In the application to the speech monitoring data we used the GLM to obtain ROIs. Then we obtained an average of the GLM coefficients from these ROIs to obtain the variance matrix to compare with the variance matrix of the ancestral graph. It can be argued that both the way in which the ROIs were determined and using the average to represent the entire ROI are suboptimal. A better way of obtaining a ROI is to use a spatial model for BOLD activity like that in Weeda et al. (2009). Subsequently, the amplitude coefficient of such a ROI would be a better representative of the ROI than the average of GLM coefficients. We plan to combine both methods in the future.

Appendix

Model assumptions

To include the true scores for each subject we are in fact assuming a third level in the random effects model. To obtain the two-level model again requires additional assumptions. We already had the assumption that the data Y are $N(0, \Sigma)$. We need additionally that (i) the variance of the random effects is $N(\theta_{0i}, \sigma^2\Psi)$, (ii) that σ is small, and (iii) that θ_{0i} are from the population $N(\mu, \Psi)$. The first and second assumptions say that the variation of the individual effects is proportional to the population variance but much smaller. This is not an unreasonable assumption since the subjects are derived from that population and it is likely that within subject variation is smaller than between subject variation. Intuitively you would expect from these assumptions that with small σ the third level has a minor impact because the probability is high that we obtain a value close to the true value for that subject. From the assumptions above the log-likelihood is (omitting constants)

$$\begin{aligned} L(\theta_i, \mu, \Psi; y) = & -\frac{n}{2}\sum_{i=1}^{N} log|\Sigma_i| - \frac{n}{2}\sum_{i=1}^{N} \mathrm{tr}\left[S_i\Sigma_i^{-1}\right] \\ & -\frac{N}{2} log|\sigma^2\Psi| - \frac{1}{2}\sum_{i=1}^{N} (\theta_i - \theta_{0i})'\sigma^{-2}\Psi^{-1}(\theta_i - \theta_{0i}) \\ & -\frac{N}{2} log|\Psi| - \frac{1}{2}\sum_{i=1}^{N} (\theta_{0i} - \mu)'\Psi^{-1}(\theta_{0i} - \mu). \end{aligned}$$

The estimate of θ_{0i} is then $\hat{\theta}_{0i} = \left(1 - \sigma^{-2}\right)^{-1}\left(\mu - \sigma^{-2}\theta_i\right)$. It is easy to see that when $\sigma \downarrow 0$ then $\hat{\theta}_{0i} = \theta_i$. Hence, for values of σ close to zero, the three-level model can be approximated by the two-level model where the true value of each subject is approximately obtained.

Consistency

We will show that the parameters obtained by estimating the random effects for each subject separately and then using them to estimate the population parameters leads to consistent estimates. Furthermore, we show that the order of taking limits (subjects and trials) has no consequence on the estimate. That is, the maximizer of the (nonstochastic) log-likelihood is the same.

We use a classical approach to show consistency (e.g., Amemiya, 1985, Theorem 4.1.1). If the assumptions of this theorem hold, then the estimates are consistent. This theorem makes it easy to consider more than one limit process at a time (see e.g., Amemiya, 1971). Let $\gamma = (\mu, \Psi, \theta) \in \mathbb{R}^s$ be the collection of all parameters, and let $S_{ni} =$

$n^{-1}\sum_{j=1}^{n} y_{ij}y'_{ij}$, where the subscript n is used to emphasize its dependence on the number of trials n. The three assumptions of that theorem are (i) the parameter space Γ is a compact subset of $\mathbb{R}^s$, (ii) $L_{nN}(\gamma)$ is continuous in γ for all (measurable) y, and (iii) the normalized log-likelihood $(nN)^{-1}L_{nN}(\gamma)$ converges in probability to a nonstochastic function $L(\gamma)$ uniformly in γ as $n,N\rightarrow\infty$, and attains a unique global maximum at γ_0. Assumptions (i) and (ii) are easy to satisfy. Assumption (iii) needs to be verified. We are dealing with the limits of trials $n\rightarrow\infty$ and subjects $N\rightarrow\infty$. The normalized log-likelihood is (omitting the irrelevant constant)

$$\frac{1}{nN}L_{nN}(\gamma) = -\frac{1}{2N}\sum_{i=1}^{N} log|\Sigma_i| - \frac{1}{2N}\sum_{i=1}^{N} \text{tr}\left[S_{ni}\Sigma_i^{-1}\right]$$
$$-\frac{1}{2n}log|\Psi| - \frac{1}{2n}\text{tr}\left[\frac{1}{N}\sum_{i=1}^{N}(\theta_i-\mu)'\Psi^{-1}(\theta_i-\mu)\right].$$

The first term involves only the nonstochastic components Σ_i and so is irrelevant. The last two terms will vanish because they contain n^{-1} and no sum over trials. The second term $-(2N)^{-1}\sum_{i=1}^{N}\text{tr}[S_{ni}\Sigma_i^{-1}]$ is of interest. (Note that the same result is obtained for the three-level model, and hence the following is also true for the three-level model.) We intend to show that assumption (iii) holds. Assume in addition to the assumptions in the text and (i) and (ii) above that (iv) $E\{S_{ni}\}=\Sigma_{0i}$ for each i, where $\Sigma_{0i}=\Sigma(\theta_{0i})$ is the true variance matrix of subject i, (v) that $max_i sup_\theta E\{|\text{tr}[S_{ni}\Sigma_i]|\}<\infty$, (vi) that $max_i E\{||J(y_{ij};\theta_{0i})J(y_{ij};\theta_{0i})'||\}<\infty$, (vii) that the trials are identically and independently distributed, and (viii) that subjects are uncorrelated. If convergence is first with respect to the trials and then to the subjects $(n,N)\rightarrow\infty$, then by (iv) and (vii) we can apply the weak law of large numbers (WLLN) and we have that $S_{ni}-\Sigma_{0i}=o_p(1)$ for all i, and so by the continuity theorem (see e.g., Van der Vaart, 1998)

$$\frac{1}{N}\sum_{i=1}^{N}\text{tr}\left[S_{ni}\Sigma_i^{-1}\right] - \frac{1}{N}\sum_{i=1}^{N}\text{tr}\left[\Sigma_{0i}\Sigma_i^{-1}\right] = o_p(1) \qquad (n,N)\rightarrow\infty.$$

For the reverse order of limits we assume additionally (viii) and use Chebyshev's WLLN (see e.g., Serfling, 1980) to establish the same convergence but now for the reverse order $(N,n)\rightarrow\infty$. Note that in neither case of the order of taking limits need the term converge to a constant, the nonstochastic part is required to follow the stochastic part closely. Then by (v) the log-likelihood converges in probability to a nonstochastic function, which has unique maximizer $\theta_i=\theta_{0i}$ for all i. It is unique because the mapping $\theta\mapsto\Sigma_\theta$ is one-to-one (Richardson and Spirtes, 2002, Cor. 8.8). The consistency of $\hat{\mu}$ and $\hat{\Psi}$ follow by applying the WLLN. Using the asymptotic approximation of $\hat{\theta}_{ni}$ in Eq. (6) we have that $E\{\hat{\theta}_i\} = \theta_{0i}$ for all i. It follows from the finite variance assumption in (vi) that $N^{-1}\sum_{i=1}^{N}\hat{\theta}_{ni}-N^{-1}\sum_{i=1}^{N}\theta_0 = o_p(1)$ as $N\rightarrow\infty$. If the average is μ then we are done, since then also $\hat{\Psi}(\hat{\theta})-\hat{\Psi}(\theta_0) = o_p(1)$ as $N\rightarrow\infty$.

The Wald statistic

To determine the degrees of freedom for the Wald test, the distributions of the two components $C\hat{\theta}_n$ and $C\tilde{\Psi}_bC$ need to be established (Bilodeau and Brenner, 1999). The estimate $\hat{\theta}_n$ is locally asymptotically normal, when we are not too far from the truth ($O(n^{-1/2})$). Hence, taking the inner product in the Wald test gives a sum of squared normal variables consisting of k components, which provides the first set of degrees of freedom. As for the second set of degrees of freedom, the distribution of $\tilde{\Psi}_i$ needs to be established. The stochasticity in this case comes from the differences of the squares of the data and the model: $\frac{1}{n}y_{ij}y'_{ij}-\Sigma_i$. This converges for large n to a normal variable with zero mean and variance proportional to the sandwich (Van der Vaart, 1998). And since the sum contains n elements constrained by k components in C, then by the same reasoning as before, we obtain n-k degrees of freedom.

References

Akaike, H., 1973. Information and an extension of the maximum likelihood principle. In: Petrov, B., Csáki, F. (Eds.), Proceedings of the second international symposium on information theory. Supplement to problems of control and information theory. Akademiai Kiado, Budapest, pp. 267–281.

Amemiya, T., 1971. The estimation of the variances in a variance components model. Int. Econ. Rev. 12 (1), 1–13.

Amemiya, T., 1985. Advanced econometrics. Basil Blackwell, Oxford.

Andersson, S., Madigan, D., Perlman, M., 1997. A characterization of markov equivalence classes for acyclic digraphs. Ann. Stat. 25, 505–541.

Beckmann, C., Jenkinson, M., Smith, S., 2003. General multilevel linear modeling for group analysis in FMRI. Neuroimage 20, 1052–1063.

Bilodeau, M., Brenner, D., 1999. Theory of multivariate statistics. Springer-Verlag, New York.

Bozdogan, H., 2000. Akaike's information criterion and recent developments in information complexity. J. Math. Psychol. 44, 62–91.

Browne, M., 1974. Generalized least squares estimators in the analysis of covariance structures. S. Afr. Stat. J. 8, 1–24.

Buchel, C., Friston, K., 1997. Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fmri. Cereb. Cortex 7, 768–778.

Büchel, C., Friston, K., 2000. Assessing interactions among neuronal systems using functional neuroimaging. Neural Netw. 13, 871–882.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference: understanding aic and bic in model selection. Sociol. Meth. Res. 33, 261–304.

Casteren, P., Gooijer, J., 1997. Model selection by maximum entropy. In: Fomby, T., Hill, R. (Eds.), Advances in econometrics: applying maximum entropy to econometric problems, Vol.12. JAI Press, Greenwich, Connecticut, pp. 135–161.

Christoffels, I., Formisano, E., Schiller, N., 2007. Neural correlates of verbal feedback processing: an fmri study employing overt speech. Hum. Brain Mapp. 28, 868–879.

Claeskens, G., Hjort, N., 2008. Model selection and model averaging. Cambridge University Press, Cambridge.

Cox, D., Wermuth, N., 1996. Multivariate dependencies: models, analysis and interpretation. Monographs on statistics and applied probability. Chapman & Hall.

Demidenko, E., 2004. Mixed models: theory and applications. John Wiley and Sons.

Drton, M., Perlman, M., 2004. A sinful approach to gaussian graphical model selection. Tech. Rep. 457 August, University of Washington, Department of Statistics.

Drton, M., Richardson, T., 2004. Iterative conditional fitting for gaussian ancestral graph models. Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, pp. 130–137.

Edwards, 2003. Introduction to graphical modelling. Springer-Verlag, New York.

Eichler, M., 2005. A graphical approach for evaluating effective connectivity in neural systems. Philos. Trans. R. Soc. Lond. B 360, 953–967.

Friston, K., 2007. Functional integration. In: Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W. (Eds.), Statistical parametric mapping, Ch.36. Academic Press, London, pp. 471–491.

Friston, K., Harrison, L., Penny, W., 2003. Dynamic causal modelling. Neuroimage 19, 1273–1302.

Goebel, R., Esposito, F., Formisano, E., 2006. Analysis of fiac data with brainvoyager qx: from single-subject to cortically aligned group glm analysis and self-organizing group ica. Hum. Brain Mapp. 27 (5), 392–401.

Grunwald, P., 2000. Model selection based on minimum description length. J. Math. Psychol. 44, 133–152.

Lauritzen, S., 1996. Graphical Models. Oxford University Press.

Marrelec, G., Krainik, A., Duffau, H., Pélégrini-Issac, M., Lehéricy, S., Doyon, J., Benali, H., 2006. Partial correlation for functional brain interactivity investigation in functional mri. Neuroimage 32 (1), 228–237 Aug.

Mclntosh, A., Gonzalez-Lima, F., 1997. Structural equation modeling and its application to network analysis in functional brain imaging. Hum. Brain Mapp. 2, 2–22.

Pearl, J., 2000. Causality: models and prediction. Cambridge University Press.

Penny, W., Holmes, A., 2004. Random effects analysis. In: Frackowiak, R.S., Ashburner, J.T., Penny, W.D., Zeki, S., Friston, K.J., Frith, C.D., Dolan, R.J., Price, C.J. (Eds.), Human Brain Function. Academic Press.

Rajapakse, J., Zhou, J., 2007. Learning effective brain connectivity with dynamic bayesian networks. Neuroimage 37 (3), 749–760.

Richardson, T., Spirtes, P., 2002. Ancestral graph markov models. Ann. Stat. 30 (4), 962–1030.

Robinson, G.K., 1991. That blup is a good thing: the estimation of random effects. Stat. Sci. 6 (1), 15–32.

Roebroeck, A., Formisano, E., Goebel, R., 2005. Mapping directed influence over the brain using granger causality and fmri. Neuroimage 25, 230–242.

Roebroeck A., Formisano E., Goebel R., 2009. The identification of interacting networks in the brain using fmri: model selection, causality and deconvolution. NeuroImage. doi:10.1016/j.neuroimage.2009.09.036.

Serfling, R., 1980. Approximation theorems for mathematical statistics. John Wiley and Sons.

Spirtes, P., Glymour, C., Sceines, R., 1993. Causation, prediction, and search. Springer-Verlag.

Stephan, K.E., Kasper, L., Harrison, L.M., Daunizeau, J., den Ouden, H.E.M., Breakspear, M., Friston, K.J., 2008. Nonlinear dynamic causal models for fmri. Neuroimage 42 (2), 649–662.

Valdés-Sosa, P.A., Sánchez-Bornot, J.M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L., Canales-Rodríguez, E., 2005. Estimating brain functional connectivity with sparse multivariate autoregression. Philos. Trans. R. Soc. Lond. B 360, 969–981.

Van der Vaart, A., 1998. Asymptotic statistics. Cambridge University Press, New York.

Verbeke, G., Molenberghs, G., 2000. Linear mixed models. Springer.

Verma, T., Pearl, J., 1991. Equivalence and synthesis of causal models. Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI-91). Elsevier Science, New York, NY.

Waldorp, L., 2009. Robust and unbiased variance of glm coefficients for misspecified autocorrelation and hemodynamic response models in fmri. Int. J. Biomed. Imaging 2009, 723912.

Waldorp, L., Huizenga, H., Grasman, R., 2005a. The wald test and Cramer-Rao bound for misspecified models in electromagnetic source analysis. IEEE Trans. Signal Process. 53 (9), 3427–3435.

Waldorp, L.J., Grasman, R.P.P.P., Huizenga, H.M., 2005b. Goodness-of-fit and confidence intervals of approximate models. J. Math. Psychol. 50 (2), 203–213.

Weeda, W., Waldorp, L., Christoffels, I., Huizenga, H., 2009. Activated region fitting: a robust high power method for fmri analysis using parameterized regions of activation. Hum. Brain Mapp. 30, 2595–2605.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50 (1), 1–25.

Whittaker, J., 1990. Grapphical models in applied multivariate statistics. John Wiley and Sons.

Wink, A., Roerdink, J., 2006. BOLD noise assumptions in fMRI. Int. J. Biomed. Imaging 1–11.

Young, G., Smith, R., 2005. Essentials of statistical inference. Cambridge University Press.

Yuan, K.-H., Bentler, P., 1997. Mean and covariance structure analysis: theoretical and practical improvements. J. Am. Stat. Assoc. 92, 767–774.

Zhang, J., 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artif. Intell. 172, 1873–1896.