Stat

The ISI's Journal for the Rapid Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.66

# Efficient sampling of Gaussian graphical models using conditional Bayes factors

**Max Hinne[a,b,*], Alex Lenkoski[c], Tom Heskes[b] and Marcel van Gerven[a]**

**Received 15 September 2014; Accepted 23 October 2014; Published 17 November 2014**

**Bayesian estimation of Gaussian graphical models has proven to be challenging because the conjugate prior distribution on the Gaussian precision matrix, the *G*-Wishart distribution, has a doubly intractable partition function. Recent developments provide a direct way to sample from the *G*-Wishart distribution, which allows for more efficient algorithms for model selection than previously possible. Still, estimating Gaussian graphical models with more than a handful of variables remains a nearly infeasible task. Here, we propose two novel algorithms that use the direct sampler to more efficiently approximate the posterior distribution of the Gaussian graphical model. The first algorithm uses conditional Bayes factors to compare models in a Metropolis–Hastings framework. The second algorithm is based on a continuous time Markov process. We show that both algorithms are substantially faster than state-of-the-art alternatives. Finally, we show how the algorithms may be used to simultaneously estimate both structural and functional connectivity between subcortical brain regions using resting-state functional magnetic resonance imaging. Copyright © 2014 John Wiley & Sons, Ltd.**

**Keywords: conditional Bayes factors; functional connectivity; Gaussian graphical models**

## 1 Introduction

A key objective in many areas of science is to uncover the interactions amongst a large number of variables based on a limited amount of data. Examples include gene regulatory networks, where one wants to identify the interactions amongst DNA segments; market basket analysis, where the relations are studied between customers based on their purchase behaviour; or neuroscience, where the connections between segregated neuronal populations are linked to cognitive ability and impairment. One way to estimate these relations is to employ Gaussian graphical models, where the non-zero entries in the off-diagonal of a precision matrix correspond to the edges in a conditional independence graph (Dempster, 1972). However, fully Bayesian estimation of the posterior of a Gaussian graphical model has proven to be notoriously hard.

To allow Bayesian inference of the Gaussian graphical model, a conjugate prior (Diaconis & Ylvisaker, 1979) on a precision matrix restricted by the conditional independence graph $G$ was constructed for decomposable graphs (Dawid & Lauritzen, 1993) and later generalized to arbitrary graphs (Roverato, 2002). Subsequent work coined this distribution the $G$-Wishart distribution (Atay-Kayis & Massam, 2005). A number of Monte Carlo algorithms for model estimation

[a]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen 6525 EZ, The Netherlands
[b]Institute for Computing and Information Sciences, Radboud University Nijmegen, Nijmegen 6525 AJ, The Netherlands
[c]Norwegian Computing Center, Oslo NO-0373, Norway
*Email: mhinne@cs.ru.nl

(wileyonlinelibrary.com) DOI: 10.1002/sta4.66

using the *G*-Wishart distribution have been developed (Piccioni, 2000; Mitsakakis et al., 2011; Dobra et al., 2011; Wang & Li, 2012), but each of these algorithms required substantial computational resources due to difficulty with sampling from the *G*-Wishart distribution. To address this bottleneck, a recent study proposed an efficient way to directly sample from the *G*-Wishart distribution (Lenkoski, 2013) by scaling samples from a regular Wishart distribution to fit the required dependency structure (Hastie et al., 2009). Even with the direct sampler, approximating the Gaussian graphical model remained difficult because of the doubly intractable partition function of the *G*-Wishart distribution. However, by combining features of the exchange algorithm (Murray et al., 2006) with reversible jump sampling (Green, 1995), calculating the partition function may be circumvented (Lenkoski, 2013). The algorithm that implements this idea, named the double reversible jump (DRJ) algorithm, provides substantial computational gains compared with earlier approaches (Lenkoski, 2013).

Although the DRJ algorithm enables model selection in a more efficient manner than previously possible, computational costs remain a limiting factor in practical applications with a large number of variables. In this paper, we propose two novel, faster, algorithms for Bayesian estimation of the Gaussian graphical model. In the first algorithm, we combine the direct sampler (Lenkoski, 2013) with an efficient representation of the conditional Bayes factor (Cheng & Lenkoski, 2012), which results in an elegant Metropolis–Hastings algorithm to which we will refer as the double conditional Bayes factor sampler. In the second algorithm, we cast the double conditional Bayes factors algorithm in a birth–death Markov chain Monte Carlo (MCMC) setting (Mohammadi & Wit, 2014). Here, rather than accepting or rejecting a new state with an edge added or removed, we associate with these changes birth and death events, respectively. These events occur with such rates that their equilibrium coincides with the posterior of interest (Stephens, 2000). Both algorithms provide substantial speed improvement over the status quo, as we show in simulations.

We also provide an application of our algorithms by estimating structural and functional connectivity between subcortical structures using resting-state functional magnetic resonance imaging (fMRI). It is a major goal in cognitive neuroscience to understand how spatially segregated neural populations are coupled, using indirect measures of neural activity such as functional magnetic resonance imaging (Smith et al., 2013; Salinas & Sejnowski, 2001). In this context, the anatomical pathways between neural populations are referred to as structural connectivity, whereas correlated activity patterns between these populations are referred to as functional connectivity (Friston, 2011). Both forms of connectivity may be estimated simultaneously using Gaussian graphical models. Here, the precision matrix captures the functional interactions between variables, and the associated conditional independence graph represents the direct connections between variables. Bayesian estimation of Gaussian graphical models is particularly relevant because the posterior over precision matrices provides complete information about the strength of functional interactions, and the posterior over conditional independence graphs allows one to associate a probability with a putative direct connection between variables of interest.

# 2 Gaussian graphical models

## 2.1. Preliminaries

Let observed data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ consist of $n$ independent draws from a $p$-dimensional multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{K}^{-1})$, with zero mean and precision (inverse covariance) matrix $\mathbf{K}$. Here, $\mathbf{K} \in \mathbb{P}_p$, with $\mathbb{P}_p$ the space of positive definite $p \times p$ matrices. The likelihood of $\mathbf{K}$ is given by

$$P(\mathbf{X} \mid \mathbf{K}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i \mid \mathbf{0}, \mathbf{K}^{-1}) \propto |\mathbf{K}|^{n/2} \exp\left[-\frac{1}{2}\langle \mathbf{K}, \mathbf{S}\rangle\right], \quad (1)$$

 

where $\mathbf{S} = \mathbf{X}^T\mathbf{X}$ is the empirical covariance and $\langle\cdot,\cdot\rangle$ the trace inner product operator. The precision matrix has the important property that zero elements correspond to conditional independencies. In other words, (1) specifies a Gaussian Markov random field with respect to a graph $G = (V, E)$, with $V = \{1, \ldots, p\}$ and $E \subset V \times V$, in which the absence of a connection indicates independence, that is, $(i, j) \notin E \rightarrow k_{ij} = 0$. For convenience, throughout this paper, we slightly abuse notation and use $(i, j) \in G$ to indicate that the edge $(i, j)$ is present in $E$.

The dependency graph may be used to specify a prior distribution on the precision matrix, which is known as the $G$-Wishart distribution (Roverato, 2002)

$$P(\mathbf{K} \mid G, \delta, \mathbf{D}) = \mathcal{W}_G(\delta, \mathbf{D}) = \frac{|\mathbf{K}|^{(\delta-2)/2}}{Z_G(\delta, \mathbf{D})} \exp\left[-\frac{1}{2}\langle \mathbf{K}, \mathbf{D}\rangle\right] \mathbf{1}_{\mathbf{K}\in\mathbb{P}_G}, \tag{2}$$

in which $\mathbb{P}_G$ is the space of positive definite $p \times p$ matrices that have zero elements wherever $(i, j) \notin G$, $\delta$ is the prior degrees of freedom, $\mathbf{D}$ is the prior scaling matrix and $\mathbf{1}_x$ evaluates to 1 if and only if $x$ holds and to 0 otherwise. The $G$-Wishart distribution is conjugate to the multivariate Gaussian likelihood in (1), so that

$$P(\mathbf{K} \mid G, \delta, \mathbf{D}, \mathbf{X}) = \mathcal{W}_G(\delta + n, \mathbf{D} + \mathbf{S}) = \frac{|\mathbf{K}|^{(n+\delta-2)/2}}{Z_G(\delta + n, \mathbf{D} + \mathbf{S})} \exp\left[-\frac{1}{2}\langle \mathbf{K}, \mathbf{D} + \mathbf{S}\rangle\right]. \tag{3}$$

Note that the Wishart distribution is a special case of the $G$-Wishart distribution, with which it coincides if $G$ is a fully connected graph. Importantly, the partition function $Z_G(\delta, \mathbf{D})$ depends on $G$, which makes the $G$-Wishart a doubly intractable distribution. We return to the implications of this fact later on.

Central to this work is that we wish to perform model selection in Gaussian graphical models, which revolves around the joint posterior

$$P(G, \mathbf{K} \mid \mathbf{X}) \propto P(\mathbf{X} \mid \mathbf{K})P(\mathbf{K} \mid G)P(G). \tag{4}$$

In the remainder, we outline several algorithms to approximate this distribution.

## 2.2. Direct samples from the G-Wishart distribution

Because the prior $P(\mathbf{K} \mid G)$ is $\mathcal{W}_G(\delta, \mathbf{D})$, we need a way to draw samples from the $G$-Wishart distribution. Up until recently, this was achieved using a block Gibbs sampler that updates $\mathbf{K}$ according to either the edges of $G$ (Wang & Li, 2012) or its clique decomposition (Piccioni, 2000). Although this enables model inference of $P(G, \mathbf{K} \mid \mathbf{X})$, as desired, both approaches require substantial computational effort, making them prohibitive for use in contexts with a large number of variables. An alternative method was proposed that is more efficient (Lenkoski, 2013), which is an adaptation of an algorithm for estimating the mode $\hat{\mathbf{K}}$ of the $G$-Wishart distribution (Hastie et al., 2009; Moghaddam et al., 2009). The algorithm is as follows:

1. Sample $\mathbf{\Sigma} \sim \mathcal{W}(\delta, \mathbf{D})$.
2. Let $\mathbf{W} = \mathbf{\Sigma}$.
3. Repeat for $j = 1, 2, \ldots, p$ until convergence:
   (a) Let $N_j \subset V$ be the set of variables that are connected to $j$ in $G$. Form $\mathbf{W}_{N_j}$ and $\mathbf{\Sigma}_{N_j j}$ and solve
   $$\hat{\beta}_j^* = \mathbf{W}_{N_j}^{-1}\mathbf{\Sigma}_{N_j j}.$$
   (b) Form $\hat{\beta}_j \in \mathbb{R}^{p-1}$ by copying the elements of $\hat{\beta}_j^*$ to the appropriate locations and imputing zeros in those locations not connected to $j$ in $G$.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.66

(c) Replace $\mathbf{W}_{j,-j}$ and $\mathbf{W}_{-j,j}$ with $\mathbf{W}_{-j,-j}\hat{\beta}_j$.

4. Return $\mathbf{K} = \mathbf{W}^{-1}$.

Conceptually, the algorithm draws a sample from a Wishart distribution, which is then iteratively scaled according to the dependence structure in $G$. In practice, we observe that convergence (see step 3) is typically reached within a handful of iterations, even for moderate to large $p$.

# 3 Sampling algorithms

The direct sampler paves the way for novel inference algorithms. Here, we introduce two novel algorithms for approximation of the joint posterior in (4).

## 3.1. Double reversible jump sampler

As a baseline for comparison, we use the DRJ sampler (Lenkoski, 2013). This algorithm was shown to be more efficient as previously used approaches and may be considered state of the art. It builds upon the reversible jump sampler discussed in Dobra & Lenkoski (2011). The key idea offered by this approach is that it introduces an auxiliary variable $\mathbf{K}^0 \sim \mathcal{W}_G(\delta, \mathbf{D})$, as in the exchange algorithm (Murray et al., 2006), that is efficiently sampled using the direct sampler discussed earlier. Because of the way this auxiliary variable is constructed, the doubly intractable partition functions of the $G$-Wishart distribution are cancelled out in the calculation of the acceptance ratios of newly proposed graphs.

## 3.2. Direct double conditional Bayes factor (DCBF) sampler

The DRJ algorithm provides a substantial improvement over previous algorithms, as it avoids the need to approximate the ratio of partition functions or invoke the Gibbs sampling algorithm for drawing samples from the $G$-Wishart distribution. Nonetheless, the algorithm can be simplified. In Cheng & Lenkoski (2012), it is shown that if $G$ and $\tilde{G}$ differ only in the edge $e = (p-1, p)$ and $G \subset \tilde{G}$, the odds ratio of these two models may be expressed as

$$\frac{P\left(\mathbf{X} \mid \tilde{G}, \mathbf{K}, \mathbf{D}\right)}{P(\mathbf{X} \mid G, \mathbf{K}, \mathbf{D})} = N(\mathbf{K}, \mathbf{D} + \mathbf{S}) \frac{Z_G(\delta, \mathbf{D})}{Z_{\tilde{G}}(\delta, \mathbf{D})} \tag{5}$$

with

$$N(\mathbf{K}, \mathbf{U}) \equiv \phi_{p-1,p-1} \left(\frac{2\pi}{u_{pp}}\right)^{1/2} \exp\left[\frac{1}{2} u_{pp} \left(\frac{\phi_{p-1,p-1} u_{p-1,p}}{u_{pp}} - \frac{\sum_{l=1}^{p-2} \phi_{lp-1}\phi_{lp}}{\phi_{p-1,p-1}}\right)^2\right], \tag{6}$$

where $\mathbf{K} = \Phi^T\Phi$, with $\Phi$ the Cholesky decomposition of $\mathbf{K}$. The term in (5) can be considered the conditional Bayes factor of the comparison between $G$ and $\tilde{G}$. Similar to the DRJ approach, Cheng & Lenkoski (2012) propose to augment the sampling process with an auxiliary variable $\mathbf{K}^0 \sim \mathcal{W}_G(\delta, \mathbf{D})$. This results in a convenient acceptance ratio for the addition of an edge to $G$

$$\alpha = \frac{N(\mathbf{K}, \mathbf{D} + \mathbf{S})}{N\left(\tilde{\mathbf{K}}^0, \mathbf{D}\right)} \frac{P(\tilde{G})}{P(G)}, \tag{7}$$

 

(wileyonlinelibrary.com) DOI: 10.1002/sta4.66

The ISI's Journal for the Rapid Dissemination of Statistics Research

where the ratio is inverted if the edge is removed from $G$ instead. Note that the variables $G, \mathbf{K}, \mathbf{U}$ and $\mathbf{D}$ must be permuted for each edge flip to place the particular edge under consideration in the position $(p-1, p)$.

The algorithm described in Cheng & Lenkoski (2012) employs the block Gibbs sampler to sample from the $G$-Wishart distribution. Instead we propose to make use of the direct sampler explained in Section 2.2 to arrive at the following procedure for estimation of the Gaussian graphical model:

1. Let $G = G^{[s]}$ be the current graph and let $\mathbf{K} = \mathbf{K}^{[s]} \sim \mathcal{W}_G(\delta + n, \mathbf{D} + \mathbf{S})$.
2. For each edge $(i, j) \in G$, do the following:
   (a) Create a permutation of the variables so that $(i, j) \to (p-1, p)$. Permute $G, \mathbf{K}, \mathbf{D}$ and $\mathbf{S}$ accordingly.
   (b) Let $\tilde{G} = G \cup (p-1, p)$ if $(p-1, p) \notin G$ or $\tilde{G} = G \setminus (p-1, p)$ if $(p-1, p) \in G$.
   (c) Draw $\tilde{\mathbf{K}}^0 \sim \mathcal{W}_{\tilde{G}}(\delta, \mathbf{D})$.
   (d) Accept the move from $G$ to $\tilde{G}$ with probability $\alpha$ as in (7).
   (e) Restore the original ordering of $G, \mathbf{K}, \mathbf{D}$ and $\mathbf{S}$ and draw $\tilde{\mathbf{K}} \sim \mathcal{W}_{\tilde{G}}(\delta + n, \mathbf{D} + \mathbf{S})$
3. Set $G^{[s+1]} = \tilde{G}$ and $\mathbf{K}^{[s+1]} \sim \mathcal{W}_{\tilde{G}}(\delta + n, \mathbf{D} + \mathbf{S})$.

The usage of the direct sampler instead of the block Gibbs updates makes this direct DCBF algorithm computationally much more efficient (Liang, 2010).

## 3.3. Double continuous time (DCT) sampler

A downside of the usage of an auxiliary variable scheme is that it decreases the acceptance probability of proposals, as essentially two moves have to be accepted at once. This hampers mixing of the Markov chain, so that multimodal distributions are approximated poorly. To improve acceptance, Mohammadi & Wit (2014) introduce a birth–death continuous-time Markov process (Cappé et al., 2003) for Gaussian graphical models. Rather than accepting the addition or removal of an edge, Mohammadi & Wit (2014) associate birth and death events with these changes, respectively. Each edge dies independently of all others as a Poisson process with death rate $d_e(G, \mathbf{K})$. Because the edges are independent, the overall death rate at a particular pair of graph $G$ and precision $\mathbf{K}$ is $d(\mathbf{K}) = \sum_e d_e(G, \mathbf{K})$. Birth rates $b(\mathbf{K})$ are defined similarly but for non-edges instead.

Because the birth and death processes are independent Poisson processes, the expected time between two events is $1/(d(\mathbf{K}) + b(\mathbf{K}))$. This time can be considered the process spent at any particular instance of $(G, \mathbf{K})$. The probability of the death event of edge $e \in G$ is

$$P(\text{death of edge } e) = \frac{d(G, \mathbf{K})}{b(G, \mathbf{K}) + d(G, \mathbf{K})}, \tag{8}$$

with again an analogous definition for the birth event for a non-edge.

Mohammadi & Wit (2014) show that the birth–death process has the posterior $P(G, \mathbf{K} \mid \mathbf{X})$ as stationary distribution, if for all edges and non-edges $e$

$$d_e(\tilde{G}, \tilde{\mathbf{K}}) P(\tilde{G}, \tilde{\mathbf{K}} \mid \mathbf{X}) = b_e(G, \mathbf{K}) P(G, \mathbf{K} \mid \mathbf{X}), \tag{9}$$

for $\tilde{G} = G \cup e$. The birth and death rates may be chosen accordingly as

$$b_e(G, \mathbf{K}) = \frac{P(\tilde{G}, \tilde{\mathbf{K}} \mid \mathbf{X})}{P(G, \mathbf{K} \mid \mathbf{X})} \quad \text{for } e \notin G \quad \text{and} \quad d_e(G, \mathbf{K}) = \frac{P(G, \mathbf{K} \mid \mathbf{X})}{P(\tilde{G}, \tilde{\mathbf{K}} \mid \mathbf{X})} \quad \text{for } e \in G. \tag{10}$$

with again $\tilde{G} = G \cup e$.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.66

The key observation is that these birth–death rates can be computed using the double conditional Bayes factors as in (7). Here, again, we make use of the exchange framework by introducing the auxiliary variable $\mathbf{K}^0$, such that explicit evaluation of the partition functions is circumvented. This leads to a novel approach that we will refer to as the DCT sampler given by the following:

1. Let $G = G^{[s]}$ be the current graph and let $\mathbf{K} = \mathbf{K}^{[s]} \sim \mathcal{W}_G(\delta + n, \mathbf{D} + \mathbf{S})$.
2. For each non-edge $e \notin G$,
   (a) Create a random permutation of the variables so that $(i,j) \rightarrow (p-1,p)$. Permute $G, \mathbf{K}, \mathbf{D}$ and $\mathbf{S}$ accordingly.
   (b) Let $\tilde{G} = G \cup e$. Draw $\mathbf{K}^0 \sim \mathcal{W}_{\tilde{G}}(\delta, \mathbf{D})$
   (c) Compute the birth rate $b_e(G, \mathbf{K})$ using (10).
3. Compute the total birth rate of the current state $b(G, \mathbf{K})$.
4. For each edge $e \in G$,
   (a) Create a random permutation of the variables so that $(i,j) \rightarrow (p-1,p)$. Permute $G, \mathbf{K}, \mathbf{D}$ and $\mathbf{S}$ accordingly.
   (b) Let $\tilde{G} = G \setminus e$. Draw $\mathbf{K}^0 \sim \mathcal{W}_{\tilde{G}}(\delta, \mathbf{D})$
   (c) Compute the death rate $d_e(G, \mathbf{K})$ using (10).
5. Compute the total death rate of the current state $d(G, \mathbf{K})$ and the waiting time between events $w(G, \mathbf{K}) = 1/(d(\mathbf{K}) + b(\mathbf{K}))$.
6. Create a birth or death event according to the probabilities of death events (8) and birth events, and set $G^{[s+1]} = \tilde{G}$ and $\mathbf{K}^{[s+1]} \sim \mathcal{W}_{\tilde{G}}(\delta + n, \mathbf{D} + \mathbf{S})$.

# 4 Experiments

In this section, we first analyse the validity of the two proposed methods using an example with a known precision matrix. Subsequently, we apply the algorithms in an explorative study to identify structural and functional connectivity between subcortical brain structures.

## 4.1. Simulation

We compared the performance of the DRJ algorithm and the two novel algorithms using a simulation proposed by Wang & Li (2012). In this example, we have $p = 6$ and $n = 18$. Furthermore, the precision matrix $\mathbf{K}$ is given by $k_{ii} = 1$ for $i = 1, \ldots, p$, $k_{i,i+1} = k_{i+1,i} = 0.5$ for $i = 1, \ldots, p-1$ and finally, $k_{1p} = k_{p1} = 0.4$. The associated conditional independence graph $G$ follows as $(i,j) \in G \leftrightarrow k_{ij} \neq 0$. The scatter matrix is then constructed as $\mathbf{S} = \mathbf{X}\mathbf{X}^T = n\mathbf{K}^{-1}$, which corresponds to $n$ independent observations of $\mathcal{N}(\mathbf{0}, \mathbf{K}^{-1})$. Through exhaustive enumeration of all 32,768 possible graphs of size $p$, Wang & Li (2012) show that the posterior edge probabilities are

$$P((i,j) \in G \mid \mathbf{X}) = \begin{pmatrix} 1 & 0.969 & 0.106 & 0.085 & 0.113 & 0.850 \\ 0.969 & 1 & 0.980 & 0.098 & 0.081 & 0.115 \\ 0.106 & 0.980 & 1 & 0.982 & 0.098 & 0.086 \\ 0.085 & 0.098 & 0.982 & 1 & 0.980 & 0.106 \\ 0.113 & 0.081 & 0.98 & 0.980 & 1 & 0.970 \\ 0.850 & 0.115 & 0.086 & 0.106 & 0.970 & 1 \end{pmatrix}, \tag{11}$$

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.66

The ISI's Journal for the Rapid Dissemination of Statistics Research

and the expectation of $\mathbf{K}$ is

$$\mathbb{E}(\mathbf{K} \mid \mathbf{X}) = \begin{pmatrix} 1.139 & 0.569 & -0.011 & 0.006 & -0.013 & 0.403 \\ 0.569 & 1.175 & 0.574 & -0.008 & 0.005 & -0.014 \\ -0.011 & 0.574 & 1.176 & 0.574 & -0.008 & 0.006 \\ 0.006 & -0.008 & 0.574 & 1.175 & 0.573 & -0.011 \\ -0.013 & 0.005 & -0.008 & 0.573 & 1.175 & 0.569 \\ 0.403 & -0.014 & 0.006 & -0.011 & 0.569 & 1.138 \end{pmatrix}. \tag{12}$$

We approximate this ground truth using the three different algorithms, each implemented in MATLAB (The MathWorks, Inc., Natick, MA, USA, Release 2014). Throughout, we use vague priors in the form of $P(G) \propto 1$ for $G$ and $P(\mathbf{K} \mid G) = \mathcal{W}_G(3, \mathbf{I}_p)$. The algorithms are each executed for 100,000 iterations, of which the first 50,000 are discarded as burn-in. Conditional expectations for edges (i.e. edge probabilities) and precision matrices are then calculated as

$$\mathbb{E}((i,j) \in G \mid \mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{(i,j)\in G_t} \quad \text{and} \quad \mathbb{E}(\mathbf{K} \mid \mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{K}_t \tag{13}$$

for the DRJ and the double conditional Bayes factor algorithms, with $T$ as the number of samples. For the DCT algorithm, these expectations are calculated as

$$\mathbb{E}((i,j) \in G \mid \mathbf{X}) = \frac{1}{W}\sum_{t=1}^{T} w_t \mathbf{1}_{(i,j)\in G_t} \quad \text{and} \quad \mathbb{E}(\mathbf{K} \mid \mathbf{X}) = \frac{1}{W}\sum_{t=1}^{T} w_t \mathbf{K}_t, \tag{14}$$

with $W = \sum_{t=1}^{T} w_t$. It is easy to see that this idea generalizes the discrete time MCMC approach by assuming $w_t = 1$ for all $t$.

We quantify the approximation accuracy of the three algorithms in a number of ways. First, the accuracy of the edge probabilities is expressed using the mean squared error with respect to the true probabilities in (11). Second, we compute the Kullback–Leibler divergence (Kullback & Leibler, 1951) between the precision matrix obtained by Wang & Li (2012) as given in (12) and $\hat{\mathbf{K}} \equiv \mathbb{E}(\mathbf{K} \mid \mathbf{X})$ using either of the algorithms. We also count the number of unique models that each algorithm considers to express mixing behaviour. Next, we compute the marginal posterior probability of the true graph. Finally, we compute the relative computational speeds of the algorithms. The results of the comparison are shown in Table I. The algorithms have similar performance in approximating the desired posterior

**Table I.** Results for the comparison between the three described samplers on a simulated example, averaged over ten simulations. Standard errors are indicated in parentheses. Shown are the MSE of edge probabilities relative to (11), the KL divergence between the expected precision matrix and (12), the number of unique models visited, the marginal posterior probability of the true graph $P(G \mid \mathbf{S})$ and the relative speed of the algorithms compared with the DRJ baseline.

| Algorithm | MSE | KL | No. of models | $P(G \mid \mathbf{S})$ | Rel. speed |
|---|---|---|---|---|---|
| DRJ | 5e-04 (4e-05) | 1e-04 (2e-05) | 1299 (31) | 0.3674 (0.0008) | 1 (0) |
| DCBF | 5e-04 (2e-05) | 1e-04 (1e-05) | 1472 (23) | 0.3826 (0.0040) | 3.57 (1e-01) |
| DCT | 1e-03 (1e-05) | 7e-04 (3e-04) | 1187 (35) | 0.4277 (0.0008) | 3.80 (1e-02) |

MSE, mean squared error; KL, Kullback-Leibler divergence; DRJ, double reversible jump; DCBF, double conditional Bayes factor; DCT, double continuous time.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.66

distribution, and each obtains the true graph as the mode of the approximated distribution. Contrary to Mohammadi & Wit (2014), we do not find the continuous time algorithm to have the best mixing. In fact, of the three considered models, the continuous time MCMC approach finds the smallest number of unique models. Note that the continuous time approach may converge faster (Rao & Teh, 2012), but this is not apparent in this simulation. Finally, the efficiency of our way of computing the conditional Bayes factor (see (5)) is demonstrated by a substantial speed increase, as the DCBF algorithm is 3.57 times faster than the DRJ sampler, and the DCT algorithm is 3.80 times faster than the DRJ algorithm, whereas the algorithm in Mohammadi & Wit (2014) is 1.79 times slower than the DRJ sampler.

## 4.2. Subcortical brain connectivity

As an explorative example, we estimate structural and functional connectivity in a fully Bayesian setting. In previous work, functional connectivity has been estimated under the assumption that the underlying structural connectivity was known (Hinne et al., 2014). Here, we address the more challenging problem of simultaneously estimating the posterior distribution of both structural and functional connectivity.

4.2.1. Empirical data. The data consist of resting-state functional MRI data collected for one subject. We refer the reader to van Oort et al. (2014) for details of the acquisition protocol. Preprocessing was performed using FSL 5.0 (Jenkinson et al., 2012) and consisted of the following steps. T1 images were linearly registered to MNI-152 space. Multi-echo volumes at each TR were combined (Poser et al., 2006). Motion correction was performed using MCFLIRT, and estimated motion parameters were regressed out together with their temporal derivatives and mean time courses for both WM and CSF. Finally, data were high-pass filtered at 0.001 Hz. Subcortical structures were segmented using FSL FIRST (Patenaude et al., 2011), resulting in data for a total of 14 regions, consisting of bilateral accumbens, amygdala, caudate, hippocampus, pallidum, putamen and thalamus (see Figure 1A). For each of these regions, the signal was averaged over all voxels in that region and subsequently standardized to have zero mean and unit variance.

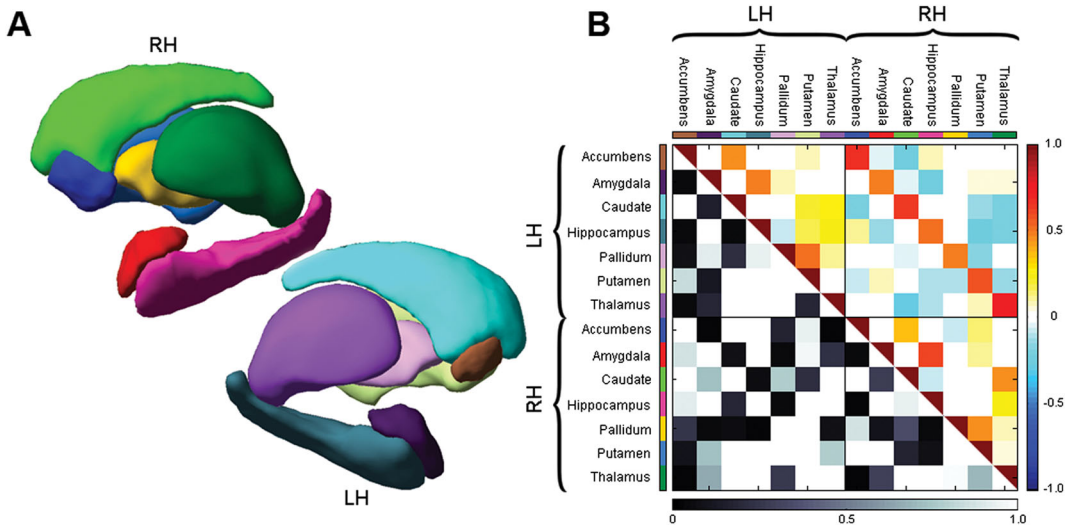


**Figure 1.** Subcortical connectivity. A. Subcortical structures, consisting of bilateral accumbens, amygdala, caudate, hippocampus, pallidum, putamen and thalamus. B. Posterior probabilities of structural connectivity (lower triangle) and expected partial correlations between these structures (upper triangle). LH and RH indicate left hemisphere and right hemisphere, respectively.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.66

The ISI's Journal for the Rapid Dissemination of Statistics Research

4.2.2. Bayesian structural and functional connectivity estimation. The human brain can be viewed as a complex dynamical system, where ongoing changes in neuronal dynamics produce adaptive behaviour (Bullmore & Sporns, 2009). These dynamics can be expressed in terms of interactions between brain regions, which are commonly referred to as functional connectivity. At the same time, direct functional interactions presuppose anatomical links between brain regions, known as structural connectivity. For this reason, structural and functional connectivity must be intimately related (Akil et al., 2011).

Functional connectivity is most easily expressed using a covariance matrix that, in the case of standardized data, provides the correlation structure between different brain regions. However, this approach suffers from the drawback that it cannot distinguish between direct and indirect connections. Alternatively, one may use partial correlations that capture only direct effects, in the absence of confounding factors. The matrix of partial correlations $\mathbf{R}$ may be obtained from a precision matrix using $r_{ij} = 1$ if $i = j$ and $r_{ij} = -k_{ij}/\sqrt{k_{ii}k_{jj}}$ otherwise. Because functional coupling must be accompanied by an anatomical connection, partial correlations between brain regions not only reveal the strength of these couplings but also indicate which regions are physically connected. In other words, the joint posterior in (4) becomes a distribution over functional connectivity $\mathbf{K}$ (or, equivalently, $\mathbf{R}$) and structural connectivity $G$.

We proceed by approximating the joint posterior using both the DCBF algorithm and the DCT sampler. Both algorithms were executed for 100,000 iterations, of which the first 50,000 were discarded as burn-in. Once again, we set $P(G) \propto 1$ and $P(\mathbf{K} \mid G) = \mathcal{W}_G(3, \mathbf{I}_p)$. The algorithms yield almost identical results, as shown by an MSE of edge probabilities of 0.0006 and a symmetrized Kullback–Leibler divergence of 0.0002.

Figure 1B shows the posterior edge probabilities and partial correlations produced by the DCBF algorithm. The structural connectivity estimate shows that the majority of edges is associated with either very high or very low edge probabilities. The functional connectivity estimate shows that functional homologues in the left and right hemispheres are associated with high partial correlations (expected partial correlations $\langle r \rangle$ in the range $[0.48, 0.73]$), indicating that these functional homologues have similar functional roles. Within a cortical hemisphere, the most salient functional interactions (highest expected partial correlations with $\langle r \rangle$ in the range $[0.23, 0.61]$) are given bilaterally by amygdala–hippocampus, pallidum–putamen, accumbens–caudate, caudate–thalamus and hippocampus–thalamus. These functional interactions can be explained by direct pathways and unobserved common inputs that induce a high partial correlation. Interestingly, edges with high posterior probability (edge probability higher than 0.999) can be associated with weak absolute partial correlations (with $\langle r \rangle$ as low as 0.1). This indicates that there exist weakly coupled regions (from the linear correlation point of view) that cannot be explained away by other functional interactions.

# 5 Discussion

We have proposed two novel algorithms for Bayesian model selection in a Gaussian graphical model. The first algorithm combines a direct manner to sample $G$-Wishart variates (Lenkoski, 2013) with an efficient way of computing conditional Bayes factors when comparing two different models (Cheng & Lenkoski, 2012), resulting in an improved Metropolis–Hastings approach. The second approach integrates the direct sampler within a birth–death continuous time Markov process (Mohammadi & Wit, 2014). Both algorithms provide accurate estimates of the posterior graphs and precision matrices and are substantially faster (up to a factor of 3.80) than previously available alternatives. We demonstrate the use of the algorithms by estimating, for the first time, both structural and functional connectivity simultaneously using fMRI data.

In future work, we aim to improve mixing of the samplers by introducing moves between graphs that differ by more than a single edge. Similarly, one may conceive events other than births and deaths of edges. In either case, the corresponding conditional Bayes factors must be derived, and these should be more efficient to compute than a

(wileyonlinelibrary.com) DOI: 10.1002/sta4.66

series of consecutive edge additions and removals. We expect that this will further contribute to efficient estimation of Gaussian graphical models.

## Acknowledgements

The authors gratefully acknowledge the support of the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science. Alex Lenkoski's work is supported by the Statistics for Innovation, $(sfi)^2$, in Oslo. The authors thank Erik van Oort and David Norris for the acquisition of the fMRI data.

## References

Akil, H, Martone, ME & van Essen, DC (2011), 'Challenges and opportunities in mining neuroscience data', *Science*, **331**(6018), 708–712.

Atay-Kayis, A & Massam, H (2005), 'A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models', *Biometrika*, **92**(2), 317–335.

Bullmore, E & Sporns, O (2009), 'Complex brain networks: graph theoretical analysis of structural and functional systems', *Nature Reviews Neuroscience*, **10**(3), 186–198.

Cappé, O, Robert, CP & Rydén, T (2003), 'Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(3), 679–700.

Cheng, Y & Lenkoski, A (2012), 'Hierarchical Gaussian graphical models: beyond reversible jump', *Electronic Journal of Satistics*, **6**, 2309–2331.

Dawid, AP & Lauritzen, SL (1993), 'Hyper Markov laws in the statistical analysis of decomposable graphical models', *Annals of Statistics*, **21**, 1272–1317.

Dempster, AP (1972), 'Covariance selection', *Biometrics*, **28**, 157–175.

Diaconis, P & Ylvisaker, D (1979), 'Conjugate priors for exponential families', *Annals of Statistics*, **7**(2), 269–281.

Dobra, A & Lenkoski, A (2011), 'Copula Gaussian graphical models and their application to modeling functional disability data', *Annals of Applied Statistics*, **5**(2A), 969–993.

Dobra, A, Lenkoski, A & Rodriguez, A (2011), 'Bayesian inference for general Gaussian graphical models with application to multivariate lattice data', *Journal of the American Statistical Association*, **106**, 1418–1433.

Friston, K (2011), 'Functional and effective connectivity: a review', *Brain Connectivity*, **1**(1), 13–35.

Green, PJ (1995), 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika*, **82**, 711–732.

Hastie, T, Tibshirani, R & Friedman, J (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edn., *Springer*, New York.

Hinne, M, Ambrogioni, L, Janssen, RJ, Heskes, T & van Gerven, M (2014), 'Structurally-informed Bayesian functional connectivity analysis', *NeuroImage*, **68**, 294–305.

 

Jenkinson, M, Beckmann, CF, Behrens, TE, Woolrich, MW & Smith, SM (2012), 'FSL', *NeuroImage*, **62**, 782–790.

Kullback, S & Leibler, RA (1951), 'On information and sufficiency', *Annals of Mathematical Statistics*, **22**(1), 79–86.

Lenkoski, A (2013), 'A direct sampler for *G*-Wishart variates', *Stat*, **2**(1), 119–128.

Liang, F (2010), 'A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants', *Journal of Statistical Computation and Simulation*, **80**, 1007–1022.

Mitsakakis, N, Massam, H & Escobar, MD (2011), 'A Metropolis–Hastings based method for sampling from the *G*-Wishart distribution in Gaussian graphical models', *Electronic Journal of Satistics*, **5**, 18–30.

Moghaddam, B, Marlin, B, Khan, E & Murphy, K (2009), 'Accelerating Bayesian structural inference for non-decomposable Gaussian graphical models', Adv Neural Inf Process in Bengio, Y, Schuurmans, D, Lafferty, J, Williams, CKI & Culotta, A (eds), Curran Associates Inc., Red Hook, NY, USA, 1285–1293.

Mohammadi, A & Wit, EC (2014), 'Bayesian structure learning in sparse Gaussian graphical models', *Bayesian Analysis*, In press.

Murray, I, Ghahramani, Z & MacKay, DJC (2006), 'MCMC for doubly-intractable distributions', Proc 22nd Ann Conf Uncertainty in Artificial Intelligence (UAI-06), AUAI Press, Cambridge, MA, USA, 359–366.

Patenaude, B, Smith, SM, Kennedy, D & Jenkinson, M (2011), 'A Bayesian model of shape and appearance for subcortical brain', *NeuroImage*, **56**(3), 907–922.

Piccioni, M (2000), 'Independence structure of natural conjugate densities to exponential families and the Gibbs sampler', *Scandinavian Journal of Statistics*, **27**, 111–127.

Poser, BA, Versluis, MJ, Hoogduin, JM & Norris, DG (2006), 'BOLD contrast sensitivity enhancement and artifact reduction with multiecho EPI: parallel-acquired inhomogeneity-desensitized fMRI', *Magnetic Resonance in Medicine*, **55**(6), 1227–1235.

Rao, V & Teh, YW (2012), 'MCMC for continuous-time discrete-state systems', Adv Neural Inf Process 25 in Pereira, F, Burges, CJC, Bottou, L & Weinberger, KQ (eds), Curran Associates, Inc., Red Hook, NY, USA, 701–709.

Roverato, A (2002), 'Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models', *Scandinavian Journal of Statistics*, **29**(3), 391–411.

Salinas, E & Sejnowski, TJ (2001), 'Correlated neuronal activity and the flow of neural information', *Nature Reviews Neuroscience*, **2**(1), 539–550.

Smith, SM, Vidaurre, D, Beckmann, CF, Glasser, MF, Jenkinson, M, Miller, KL, Nichols, TE, Robinson, EC, Salimi-Khorshidi, G, Woolrich, MW, Barch, DM, Uğurbil, K & Van Essen, DC (2013), 'Functional connectomics from resting-state fMRI', *Trends in Cognitive Sciences*, **17**(12), 666–682.

Stephens, M (2000), 'Bayesian analysis of mixture models with an unknown number of components — an alternative to reversible jump methods', *Annals of Statistics*, **28**(1), 40–74.

van Oort, ESB, van Cappellen van Walsum, AM & Norris, DG (2014), 'An investigation into the functional and structural connectivity of the default mode network', *NeuroImage*, **90**, 381–389.

Wang, H & Li, SZ (2012), 'Efficient Gaussian graphical model determination under *G*-Wishart distributions', *Electronic Journal of Statistics*, **6**, 168–198.