Simulation of Hyper-Inverse Wishart Distributions in Graphical Models

By CARLOS M. CARVALHO

Institute of Statistics and Decision Sciences, Duke University, Durham,

North Carolina 27708-0251, U.S.A.

carlos@stat.duke.edu

HÉLÉNE MASSAM

Department of Mathematics & Statistics, York University,

Toronto M3J1P3, Canada.

massamh@mathstat.yorku.ca

and MIKE WEST

Institute of Statistics and Decision Sciences, Duke University, Durham,

North Carolina 27708-0251, U.S.A.

mw@stat.duke.edu

Summary

We introduce and exemplify an efficient method for direct sampling from hyper-inverse Wishart distributions. The method relies very naturally on the use of standard junction-tree representation of graphs, and couples these with matrix results for inverse Wishart distributions. We describe the theory and resulting computational algorithms for both decomposable and non-decomposable graphical models. An example drawn from financial time series demonstrates application in a context where inferences on a structured covariance model are required. We discuss and investigate questions of scalability of the simulation methods to higher dimensional distributions. The paper concludes with general comments about the approach, including its use in connection with existing Markov chain Monte Carlo methods that deal with uncertainty about the graphical model structure.

Some key words: Gaussian graphical models; Hyper-inverse Wishart; Junction trees; Portfolio analysis; Posterior simulation.

1 Introduction

Recent developments in Markov chain Monte Carlo and stochastic search in graphical models have led to the methodology of Gaussian graphical models now being effectively routinely applicable in multivariate analysis in problems of increasing dimension. In both decomposable and non-decomposable models we now have access to increasingly efficient methods for model specification and graphical model structure search, such as described in Dobra et al. (2004), Giudici & Green (1999), Jones et al. (2005), Atay-Kayis & Massam (2005) and Wong et al. (2003). Jones et al. (2005) present a detailed overview and description of existing and novel methods of model determination, and compare their implementation in a number of examples and simulation studies. A central element of all these methods is the family of conjugate priors defined by Dawid & Lauritzen (1993), based on the class of the hyper-Markov laws known as the hyper-inverse Wishart distributions.

Our interest here is in the efficient simulation of hyper-inverse Wishart distributions. The recent literature has focussed on graphical model structure search, with little mention of the key and complementary problem of efficient inference on the parameters of a structured covariance matrix on a given graph. One likely reason for this is the difficulties faced in dealing

with non-decomposable graphical models. This paper addresses this issue directly, using recent theoretical innovations for non-decomposable graphical models developed for a different reason by Atay-Kayis & Massam (2005), and defines a comprehensive and effective method for direct simulation of both decomposable and non-decomposable hyper-inverse Wishart distributions. We explicitly do not address the complementary well researched questions of posterior inference about graphical model structure, but note that our methods are naturally and easily embeddable within any existing Markov chain Monte Carlo or stochastic search method.

As we mention above, direct sampling from this class of distributions has not yet been explicitly addressed at any level of generality. Giudici & Green (1999) use importance sampling, while Roverato (2000) suggests an alternative parameterisation, based on the Cholesky decomposition of the precision matrix, that could provide a way to sample from hyper-inverse Wishart models on decomposable graphs and might on first glance be viewed as attractive. However, in large-scale problems a method based on the Cholesky decomposition rapidly becomes unattractive. Furthermore, the challenge of sampling hyper-inverse Wishart models on non-decomposable graphs remains open. In application of Gaussian graphical models we often also require inference for

complicated functions of the parameters of a variance matrix and so an approach to direct simulation of posteriors under hyper-inverse Wishart models is highly desirable.

Our strategy naturally uses the junction tree of a graph to decompose the hyper-inverse Wishart distribution, and so allows us to work sequentially at the level of prime components. In decomposable models this decomposition provides access to standard distributional theory for the inverse Wishart distribution. In the non-decomposable case, standard distributional results no longer hold and properties of the inverse of hyper-inverse Wishart distributions are used; see Atay-Kayis & Massam (2005) for theoretical contributions relevant to our development later.

2 Background

2.1 *Basic graph theory*

Let $G = (V, E)$ be an undirected graph with vertex set V of p elements and edge-set E. Vertices a and b are said to be neighbours in G if there is an edge $(a, b) \in E$. A graph, or subgraph, is complete if all of its vertices are connected by edges in E. A clique is a complete subgraph that is not contained within another complete subgraph. Subgraphs (A, B, C) form a

decomposition of G if $V = A \cup B$, $C = A \cap B$ is complete and C separates A from B, i.e. any path from A to B goes through C. Such a C is said to be a separator. A sequence of subgraphs that cannot be further decomposed are the prime components of a graph. A graph is decomposable if every prime component is complete.

The graph G can be represented by a perfect ordering of its prime components and separators. An ordering of components $P_i \in \mathcal{P}$ and separators $S_i \in \mathcal{S}$, $(P_1, S_2, P_2, S_3, \ldots, P_k)$, is said to be perfect if for every $i = 2, 3, \ldots, k$ the running intersection property (Lauritzen, 1996, page 15) is fulfilled, meaning that there exists a $j < i$ such that

$$S_i = P_i \cap H_{i-1} \subset P_j,$$

where

$$H_{i-1} = \bigcup_{j=1}^{i-1} P_j.$$

A junction tree for G is a tree representation of the prime components. A tree with a set of vertices equal to the set of prime components of G is said to be a junction tree if, for any two prime components P_i and P_j and any P on the unique path between P_i and P_j, $P_i \cap P_j \subset P$. A set of vertices shared by two adjacent nodes of the junction tree is complete

and defines the separator of the two subgraphs induced by the nodes. This representation plays a critical role in our simulation method, as it does in graphical modelling generally. Figure 1 shows an example of a graph and its junction tree. Efficient ways of generating the junction tree for any graph are discussed in Jones et al. (2005).

2.2 *Gaussian graphical models*

A Gaussian graphical model, or covariance selection model as named by Dempster (1972), defines a set of pairwise conditional independence relationships on a p-dimensional normally distributed random quantity X. With a non singular, positive-definite covariance matrix Σ, giving precision matrix $\Omega = \Sigma^{-1}$ with entries ω_{ij}, the univariate elements x_i and x_j of X are conditionally independent if and only if $\omega_{ij} = 0$. If $G = (V, E)$ is an undirected graph representing the joint distribution of X, $\omega_{ij} = 0$ for all pairs $(i, j) \notin E$. The canonical parameter Ω belongs to $M(G)$, the set of all positive-definite symmetric matrices with elements equal to zero for all $(i, j) \notin E$.

The density of X factorises as

$$p(X|\Sigma, G) = \frac{\prod_{P \in \mathcal{P}} p(X_P|\Sigma_P)}{\prod_{S \in \mathcal{S}} p(X_S|\Sigma_S)}, \tag{1}$$

a ratio of products of densities where X_P and X_S indicate subsets of variables

in prime components and separators respectively. Given G, this distribution is defined completely by the component-marginal covariance matrices Σ_P, subject to the consistency condition that submatrices in the intersecting, i.e. separating, components are identical, as in Dawid & Lauritzen (1993); that is, if $S = P_1 \cap P_2$ the elements of Σ_S are common in Σ_{P_1} and Σ_{P_2}.

2.3 *The hyper-inverse Wishart distribution*

In order to implement a conjugate Bayesian analysis of decomposable Gaussian graphical models Dawid & Lauritzen (1993) defined a family of probability distributions called the hyper-inverse Wishart. If $\Omega \in M(G)$, the hyper-inverse Wishart

$$\Sigma \sim \mathrm{HIW}_G(b, D) \tag{2}$$

has a degree-of-freedom parameter b and location matrix D. This distribution is the unique hyper-Markov distribution for Σ with consistent clique-marginals that are inverse Wishart; to be specific, for each $P \in \mathcal{P}$, $\Sigma_P \sim \mathrm{IW}(b, D_C)$ with density

$$p(\Sigma_P | b, D_P) \propto |\Sigma_P|^{-(b+2|P|)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma_P^{-1} D_P)\right\}, \tag{3}$$

where D_P is the positive-definite symmetric diagonal block of D corresponding to Σ_P. The full hyper-inverse Wishart joint density factorises on the

junction tree, as

$$p(\Sigma|b, D) = \frac{\prod_{P\in\mathcal{P}} p(\Sigma_P|b, D_P)}{\prod_{S\in\mathcal{S}} p(\Sigma_S|b, D_S)} \tag{4}$$

The key practical extension of the above structure to unrestricted graphs, including non-decomposable cases when some of the prime components are incomplete, is the local hyper-inverse Wishart model in which the same basic form and density decomposition hold, but with modification to the component densities on incomplete components, as in Jones et al. (2005). On a prime component P that is not complete, the component prior density $p(\Sigma_P|b, D_C)$ is obtained as follows: start with the usual inverse Wishart $\Sigma_P \sim \mathrm{IW}(b, D_P)$ to deduce the Wishart distribution for $\Omega_P = \Sigma_P^{-1}$; condition the implied Wishart density by constraints that set off-diagonal elements of Ω_P to zero consistent with G; then deduce the implied density of Σ_P by change of variables. The core representation of equation (4) holds with this modification.

3 Simulation Method

3.1 *General framework*

The sampling strategy is based on the compositional form of the joint distribution over the sequence of subgraphs defined by the junction tree.

Let $G = (V, E)$ be a graph on p nodes and assume a Gaussian graphical model with $\Omega = \Sigma^{-1} \in M(G)$. Suppose that $\Sigma \sim \text{HIW}_G(b, D)$. By generating the junction tree of G, the prime components are perfectly ordered as $\{P_1, S_2, P_2, \ldots, P_k\}$ and the joint density (4) can be written as

$$p(\Sigma|b, D) = p(\Sigma_{P_1}) \prod_{i=2}^{k} p(\Sigma_{P_i}|\Sigma_{S_i}). \tag{5}$$

Equation (5) indicates that, starting from Σ_{P_1}, there is a clear sequence of conditional distributions to be simulated in order to obtain a draw from $p(\Sigma|b, D)$ via composition. We simply need to identify the sequence of conditional distributions and a method to sample them.

3.2 *Decomposable models*

In decomposable models in which all prime components are complete, i.e. cliques, conditioning results for inverse-Wishart random variables enable sampling from each of the elements in the composition directly.

For a perfect ordering of cliques $\{C_1, C_2, \ldots, C_k\}$ we use the traditional notation $R_i = C_i \setminus H_{i-1} = C_i \setminus S_i$ and write Σ_{C_i} and D_{C_i} in their conformably partitioned forms

$$\Sigma_{C_i} = \begin{pmatrix} \Sigma_{S_i} & \Sigma_{S_i,R_i} \\ \Sigma_{R_i,S_i} & \Sigma_{R_i} \end{pmatrix}, \qquad D_{C_i} = \begin{pmatrix} D_{S_i} & D_{S_i,R_i} \\ D_{R_i,S_i} & D_{R_i} \end{pmatrix},$$

where $\Sigma_{S_i,R_i} = \Sigma'_{R_i,S_i}$. Also, let

$$\begin{aligned} \Sigma_{R_i.S_i} &= \Sigma_{R_i} - \Sigma_{R_i,S_i}\Sigma_{S_i}^{-1}\Sigma_{S_i,R_i}, \\ D_{R_i.S_i} &= D_{R_i} - D_{R_i,S_i}D_{S_i}^{-1}D_{S_i,R_i}. \end{aligned}$$

The sampling scheme is defined as follows:

(i) sample $\Sigma_{C_1} \sim \mathrm{IW}(b, D_{C_1})$, which gives values to the submatrix Σ_{S_2};

(ii) for $i = 2, \ldots, k$, sample

$$\begin{aligned} \Sigma_{R_i.S_i} &\sim \mathrm{IW}(b + |R_i|, D_{R_i.S_i}), \\ U_i &\sim \mathrm{N}(D_{R_i,S_i}D_{S_i}^{-1}, \Sigma_{R_i.S_i} \otimes D_{S_i}^{-1}), \end{aligned}$$

and then directly compute the implied values of $\Sigma_{R_i,S_i} = U_i\Sigma_{S_i}$ and $\Sigma_{R_i} = \Sigma_{R_i.S_i} + \Sigma_{R_i,S_i}\Sigma_{S_i}^{-1}\Sigma_{S_i,R_i}$.

This sequence completes the sampling of all elements in the intersecting block components of Σ on the junction tree. It remains to fill in the implied values

of the elements of Σ in the positions where $\omega_{ij} = 0$. This is done via the standard completion operation described in a general context in Massam & Neher (1998); that is, given the perfect ordering of cliques and separators, and defining $A_{i-1} = H_{i-1} \setminus S_i$ for each i, we directly evaluate the required elements as

$$\Sigma_{R_i,A_{i-1}} = \Sigma_{R_i,S_i}\Sigma_{S_i}^{-1}\Sigma_{S_i,A_{i-1}}. \tag{6}$$

3.3 *Non-decomposable models*

In non-decomposable models we use the same junction tree representation for compositional sampling, thereby breaking the problem into a series of conditional simulations. The steps are precisely as described above for prime components that are complete. The key difference, and computational difficulties, arise when we visit a prime component of the junction tree that is not complete; for such a component the standard conditioning results for the inverse-Wishart, see step (ii) in §3.2, do not apply. The challenge is then to identify a way of sampling from the appropriate conditional distribution of the elements of Σ in that component conditional on the set of values of its preceding separator.

Here we can use and extend the general theory of Atay-Kayis & Massam

(2005) that expresses a global hyper-inverse Wishart distribution, and defines a sampler for it, through the Cholesky decomposition of Ω. The key points here are, first, to use this only in each incomplete prime component within the overall compositional sampler, thereby allowing for efficient computation and scaling to large graphical models by exploiting local computation, and, secondly, to extend the theory to derive samples from the conditional distributions of hyper-inverse Wishart matrices given separating parameters. The details are as follows.

For any incomplete prime component P, first consider the Cholesky method for sampling a defined distribution $\Sigma_P \sim \mathrm{HIW}_P(b, D_P)$ on that component alone, following Atay-Kayis & Massam (2005). This method generalises properties of the Bartlett decomposition to restricted Wishart matrices based on the fact that for any matrix $x = z'z \in M(G)$ the Cholesky decomposition z is completely defined by its 'free' elements z_{ij}, $(i,j) \in E$; the remaining elements z_{ij}, $(i,j) \notin E$, are functions of the free elements and can be directly determined by the completion operation defined in Lemma 2 of Atay-Kayis & Massam (2005). With this generalisation, if $x^{-1} \sim \mathrm{HIW}_G(b, I)$ a sample of x can be simply obtained by sampling the free elements of z from independent normal and chi-squared random variates followed by the evaluation

of the non-free elements. Now, for an incomplete prime component P with $\Sigma_P \sim \text{HIW}_P(b, D_P)$, write $D_P^{-1} = T'T$ for the Cholesky decomposition of the hyper-inverse Wishart parameter matrix. Then, for $\Omega_P = \Sigma_P^{-1}$, write the Cholesky decomposition as $\Omega_P = \Phi'\Phi$, and define $\Psi = \Phi T^{-1}$. The structure of the subgraph P implies certain constraints on the elements of Ψ; see Atay-Kayis & Massam (2005) and Jones et al. (2005). The free elements are those ψ_{ij} such that (i, j) is an edge in P, and these can be simulated directly from independent chi-squared and normal random variates; see below. Then Ψ will be completed by direct, deterministic evaluation of the remaining, constrained elements. The details are as follows:

Step 1. Compute the Cholesky decomposition T of D_P^{-1}.

Step 2. Define $t_{\langle ij]} = t_{ij}/t_{jj}$.

Step 3. Create the $p \times p$ upper triangular matrix A with $a_{ii} = 0$ and, for $i \neq j$, $a_{ij} = 1$ if (i, j) is an edge in P, $a_{ij} = 0$ otherwise.

Step 4. Compute ν_i as the number of 1's in the ith row of A.

Step 5. Sample the free variables Ψ_{ij} for edges (i, j) in P: for $i = 1, \ldots, p$, $\Psi_{ii} = \sqrt{u_i}$, where $u_i \sim \chi^2_{b+\nu_i}$; for $i \neq j$ and $a_{ij} = 1$, $\Psi_{ij} \sim N(0, 1)$.

For edges (i, j) not in P, compute Ψ_{ij} as follows:

$$\Psi_{1j} = -\sum_{k=1}^{j-1} \Psi_{1k} t_{\langle kj]},$$

and, for $i > 1$,

$$\Psi_{ij} = \sum_{k=i}^{j-1} \Psi_{ik} t_{\langle kj]} - \sum_{r=1}^{i-1} \left(\frac{\Psi_{ri} + \sum_{l=r}^{i-1} \Psi_{rl} t_{\langle li]}}{\Psi_{ii}} \right) \left(\Psi_{rj} + \sum_{l=r}^{j-1} \Psi_{rl} t_{\langle kj]} \right).$$

Step 6. Finally, set $\Phi = \Psi T$ and compute $\Omega_P = \Phi' \Phi$ and then $\Sigma_P = \Omega_P^{-1}$.

The modification we need is that we want to sample from $p(\Sigma_P | \Sigma_S)$, where S represents the nodes in P that lie in the preceding separator in the junction tree, so that Σ_S is an upper left block of Σ_P as in §3.2. The changes are in fact almost trivial: we simply note that conditioning is equivalent to fixing the values of the elements in the initial rows of Φ, and therefore of Ψ, corresponding to the separator S, and skipping the corresponding steps in the sequence of computations above. Also, in fact, the elements of Φ corresponding to S can be obtained from the Cholesky decomposition of the Σ elements in the preceding prime component, so that the corresponding elements of Ψ can be immediately computed and plugged in step 5 above.

After sampling Σ_P, we continue moving down the junction tree, working with both complete, i.e. cliques, or incomplete prime components until all

the block components of the full Σ are completed. Then, again as described in §3.2 for decomposable models, the completion operation comes into play to fill in the remaining elements of Σ.

3.4 *Additional features*

A key feature and, for scaling to higher-dimensions, a critical advantage of the presented method, is that no matrix calculation exceeds the cardinality of the largest prime component, so that the largest inversion or decomposition will be of a matrix of such dimension.

Also, it should be evident that inferences about precision matrices, as in the example in the next section, are easily obtained by simple calculations based on sampled values of the variance matrices; following Lauritzen (1996, page 136),

$$\Omega = \sum_{P \in \mathcal{P}} \left(\Sigma_P^{-1}\right)^0 - \sum_{S \in \mathcal{S}} \left(\Sigma_S^{-1}\right)^0, \tag{7}$$

where K^0 denotes an extension of the matrix K with zeros so as to give it the appropriate dimensions.

4 Example

4.1 *A financial portfolio example*

The example concerns posterior inference about an 11-dimensional covariance matrix based on the graph G in Fig. 1, linking international currency exchange rates relative to the U.S. dollar. The graph is consistent with a series of $n = 100$ consecutive daily returns from the mid-1990s. The graph was generated by exploring the posterior distribution over graphical models using the method of Jones et al. (2005); this particular graph represents a posterior mode from that search, i.e. the most probable graph discovered in the Markov chain Monte Carlo sampling over graphs. The graph is economically interpretable in terms of pre-2000 trading relationships, in the way in which the structure links economic trading partners and groups. For example, the graph groups together the tightly related mainland European Union currencies into one large clique, C_3 in Fig. 1; it ties the U.K. into that clique as a key trading partner and E.U. member but one whose currency ties to the U.S. dollar were more substantially influenced by idiosyncratic British-U.S.A. factors than were by those of the central E.U. countries; and it links New Zealand and Australia together as a tight clique linked to both the dominant

E.U. and the U.K. with whom these two economies have preferential trading relationships.

The graph also happens to be decomposable; hence, under a specified hyper-inverse Wishart prior $\Sigma \sim \text{HIW}_G(b_0, D_0)$, the implied posterior is the decomposable hyper-inverse Wishart form $(\Sigma|n, S) \sim \text{HIW}_G(b, D)$ with $b = b_0 + n$ and $D = D_0 + S$, where $n = 100$ and S is the sample variance matrix of the centred and scaled returns. The prior parameters chosen are relatively non informative, with $b = 3$ and $D = I$; see Jones et al. (2005) for discussion of prior specifications. We make comparisons below with a parallel analysis on the full graph, under the usual full inverse Wishart distribution with no conditional independence constraint, thereby ignoring econometric structuring and also the parsimony that is embodied in G. The difference in log-marginal likelihood of G to the full graph is 102.6, which indicates that the current $n = 100$ observations very strongly support the structured graph relative to the full graph, even if numbers of parameters and the issue of parsimony are ignored.

The simulation method was applied to generate 1000 samples from the posterior hyper-inverse Wishart distribution. Figure 2 displays an image of the theoretically exact value of $E(\Omega|n, S)$ and compares it to the image of

the Monte Carlo estimate, the latter being just the sample mean of the 1000 simulated precision matrices. The comparison can be investigated in more detail but the graphs suffice to demonstrate the efficacy of the simulation.

Of central practical importance in financial times series and portfolio management are functions of variance matrices of residual returns that define optimal portfolio reallocations in sequential decision making about investments on items such as exchange rates; see Aguilar & West (2000), Quintana et al. (2003) and Simpson & Wilkinson (2002), for example. This serves as a very nice and practically linked example of inference about functions of variance-covariance parameters and the use of simulation of structured models of variance matrices. If y represents the returns at the next time-point, and a is a vector of 11 weights representing proportional allocation of funds invested in each of the 11 currencies, then the constrained, $1'a = 1$, portfolio minimising standard deviation as a measure of risk is given by the choice $a = \Omega 1/(1'\Omega 1)$, derived in Aguilar & West (2000), for example. The corresponding risk level is the standard deviation of $a'y$, equals to $1/\sqrt{(1'\Omega 1)}$. Hence posterior samples of Ω produce, by direct computation, posterior samples for the optimal portfolio weights and related minimised risk. Figure 3 summarises these posterior samples from the hyper-inverse Wishart posterior

on the structured graph G using, as a benchmark comparison, parallel analysis on the full, unconstrained graph that would typically be used. Figure 3 shows that the levels of variation of the optimal portfolio weights across currencies are smaller than under the full model, implying a more stable investment portfolio of a kind that is desirable on economic and business grounds; see Ledoit & Wolf (2004). The first boxplot of Fig. 4 takes this further, presenting the posterior distribution for the ratio of standard deviations, i.e risks, of these optimal portfolios under the full graph relative to that under the graph G. The optimal risk level is inferred as likely to be smaller, and practically significantly smaller, under the graph G. This forcefully suggests that a structured, parsimonious graphical model can indeed aid in reducing uncertainty and variation in portfolio weights, and thereby reduce investment risk. Additional examples of graphical structure in portfolio problems appear in Carvalho & West (2007a,b).

To take this further we combine variance matrix parameter learning with learning about the graphical model using results from the Markov chain Monte Carlo search over graphs too. From that search, the 20 most probable graphs identified appear to have posterior probabilities substantially exceeding those of other discovered graphs, so that uncertainty about the graph

structure may be approximately represented by these 20 graphs; the graph G is the posterior modal graph. Under a formal model averaging strategy, the uncertainty about graphs feeds through to the posterior distribution for the portfolio weights and variances, and these can be compared with the portfolios from both the graph G and the full graph already described. The computations then simply use the hyper-inverse Wishart simulator for the posteriors conditional on each of the sampled graphs, and average results with respect to the evaluated posterior probabilities of those graphs. The results appear in the second and third boxplots of Fig. 4. Evidently, the projected portfolio risk under this 'Bayesian model averaged' strategy exceeds that under the strategy that conditions on G, apparently naturally induced by diversity in some aspects of the underlying graphical model structure that induces more variation in portfolio weights. As with the modal graph G, the model averaged graph beats the full graph in the sense of having smaller risk for a fixed target return, as well as representing inferences based on graphs that fit the data very substantially better than the full graph.

4.2 *Scaling experiments*

A range of simulation studies to evaluate scalability has been performed on standard desktop computing platforms. We have empirically verified that our hyper-inverse Wishart simulation method is efficient in problems involving up to one thousand variables. One of the contexts for experimentation we have used concerns fitting models in several tens and low hundreds of dimensions to gene expression data taken from Jones et al. (2005). That reference describes various analyses of graphical models under hyper-inverse Wishart priors in which the sparsity of graphs, in terms of the distribution of numbers of edges, is controlled by prior distributions of the form $\Pr(\text{edge in}) = k/p$, independently over edges, for some small number $k << p$. Jones et al. (2005) develop, and provide software for, stochastic search and Markov chain Monte Carlo methods to explore posterior distributions over graphs. Using that method, we fit such models to subsets of gene expression data from a set of samples on up to 1000 genes, selecting subsets of increasing dimension. For randomly selected graphs from that posterior analysis, we can then simulate the implied hyper-inverse Wishart distribution for the variance matrix on those graphs. This experiment thus provides insight into how the computational burden of the hyper-inverse Wishart sampler changes with dimension

on relevant graphs in this real data context.

Evidently, computation time increases with the complexity of the graph, in terms of the sizes of larger cliques in decomposable graphs and the nature and dimension of larger prime components in non-decomposable cases. Figure 5 displays some results from this experiment. The analysis described above was run repeatedly, first with graphs generated under priors in which edges are included independently with probability $2/p$ and then, separately, $4/p$, to generate graphs with differing complexity, as measured by degrees of sparsity in terms of numbers of edges. In this series of experiments, and in others we have evaluated, there is an approximately linear increase in cpu time for smaller numbers of nodes, and the experiments bear out the view that the computational burden will increase at linear or less than linear rates. This is understandable since the number of cliques in larger decomposable graphs with similar degrees of sparsity will increase roughly linearly. We comment on decomposable cases in the following discussion section. For this evaluation, Table 1 provides some details of the structure of the graphs simulated for this experiment, in terms of numbers of cliques and edges.

5 Discussion

Theoretical investigations of the scalability of the method are of interest but seem very challenging. It will be of particular interest to investigate further how different aspects of sparsity in terms of numbers of edges, or size and 'density' of prime components, influence the computational burden with dimension, as in denser graphs, with larger components, matrix manipulations of higher-order are required. In non-decomposable graphs, the computational demands are affected by the complexity of the structure of prime components as well the distribution of component size. Jones et al. (2005) discuss scalability of computations in trying to estimate marginal likelihoods for non-decomposable graphs and their experiences are germane here too. In non-decomposable graphs we encounter substantially diverse structures, i.e. very large and sparse incomplete components as well as more dense large incomplete components, and the computational burdens are rather unpredictable; they are, however, predictably more substantial than for decomposable graphs in general. Advances here will rely on advances in applied probability over random graphs to generate insights into structure and complexity of sparse graphs as dimension increases. There are also evident connections with computational questions in other related areas of multivariate

modelling with Gaussian graphical structures, including directed graphs and other models, such as in Wilkinson & Yeung (2002), Wilkinson & Yeung (2004) and Yeung & Wilkinson (2002). A concern for computational scalability seems likely to force an even closer focus on questions of modelling and prior specification, and especially on the issue of sparsity of graphical structures as dimension increases, in these contexts as in undirected graphical models.

Acknowledgement

The authors are grateful for the constructive comments of the editor and referees on the original version of this paper. The authors acknowledge the support of grants from the U.S. National Science Foundation. Software implementing the hyper-inverse Wishart simulation method and the exchange rate data set used in the portfolio study example are freely available at the web site `http://www.isds.duke.edu/~mw`.

References

Aguilar, O. & West, M. (2000). Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *J. Bus. Econ. Statist.*

18, 338–57.

Atay-Kayis, A. & Massam, H. (2005). The marginal likelihood for decomposable and non-decomposable graphical Gaussian models. *Biometrika* **92**, 317–35.

Carvalho, C. M. & West, M. (2007a). Dynamic matrix-variate graphical models. *Bayesian Analysis* **-**, to appear.

Carvalho, C. M. & West, M. (2007b). Dynamic matrix-variate graphical models - A synopsis. In *Bayesian Statistics VIII*, Eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith & M. West, Oxford University Press, to appear.

Dawid, A. P. & Lauritzen, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–317.

Dempster, A. (1972). Covariance selection. *Biometrics* **28**, 157–75.

Dobra, A., Jones, B., Hans, C., Nevins, J. & West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Mult. Anal.* **90**, 196–212.

Giudici, P. & Green, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86**, 785–801.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C. & West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.* **20**, 388–400.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

Ledoit, O. & Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *J. Portfolio Manag.* **30**, 110–9.

Massam, H. & Neher, E. (1998). Estimation and testing for lattice conditional independence models on euclidean jordan algebras. *Ann. Statist.* **26**, 1051–82.

Quintana, J., Lourdes, V., Aguilar, O. & Liu, J. (2003). Global gambling. In *Bayesian Statistics VII*, Eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith & M. West, Oxford University Press, pp. 349–68.

Roverato, A. (2000). Cholesky decomposition of a hyper-inverse Wishart matrix. *Biometrika* **87**, 99–112.

Simpson, A. & Wilkinson, D. (2002). Computationally intensive techniques for a fully Bayesian, decision theoretic approach to financial forecasting and portfolio selection. *Comp. Sci. Statist.* **33**, 472–83.

Wilkinson, D. & Yeung, S. (2002). Conditional simulation from highly structured Gaussian systems, with application to blocking-MCMC for the Bayesian analysis of very large linear models. *Statist. Comp.* **12**, 287–300.

Wilkinson, D. & Yeung, S. (2004). A sparse matrix approach to Bayesian computation in large linear models. *Comp. Statist. Data Anal.* **44**, 493–516.

Wong, F., Carter, C. & Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90**, 809–30.

Yeung, S. & Wilkinson, D. (2002). Adaptive Metropolis-Hastings samplers for the Bayesian analysis of large linear Gaussian systems. *Comp. Sci. Statist.* **33**, 128–38.

Upper:

Nodes	10	30	50	100	150	1000
Cliques	7	27	47	89	130	1130
Edges	11	31	51	118	204	2020

Lower:

Nodes	10	30	50	100	150	1000
Cliques	2	18	30	70	99	998
Edges	25	89	160	276	572	1170

Table 1: Structure of simulated graphs for cpu benchmark studies. The table gives the median number of cliques and edges in the 100 generated graphs for each case, i.e. number of nodes, under the two different sparsity priors, namely the upper, dotted line, and lower, solid line, examples in Fig. 5.

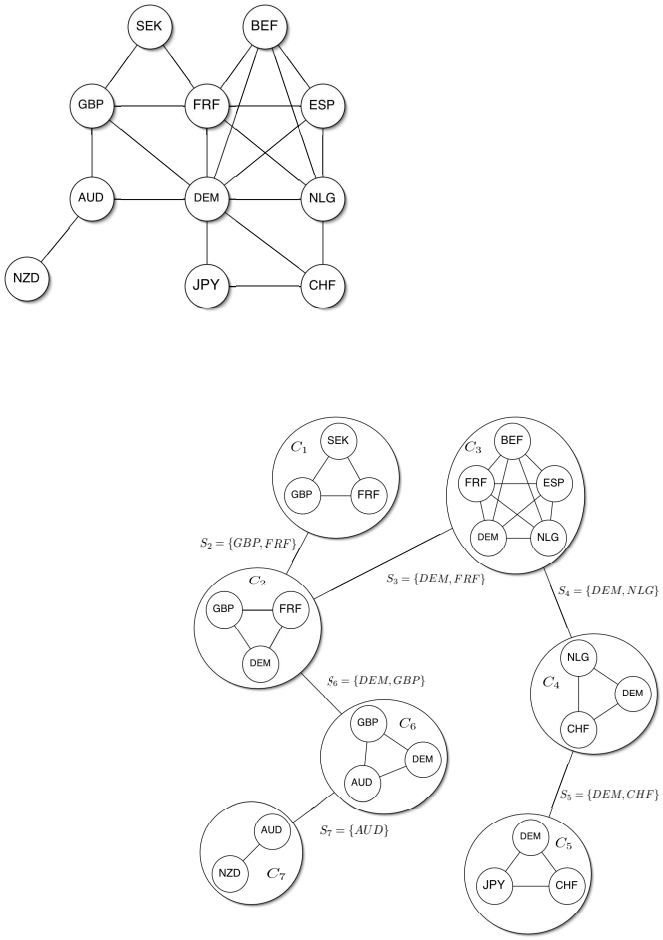

Figure 1: Exchange rate example. The currencies are as follows: New Zealand Dollar (NZD), Australian Dollar (AUS), Japanese Yen (JPY), Swedish Krone (SEK), British Pound (GBP), Spanish Peseta (ESP), Belgian Franc (BEF), French Franc (FRF), Swiss Franc (CHF), Dutch Guilder (NLG) and German Mark (DEM).

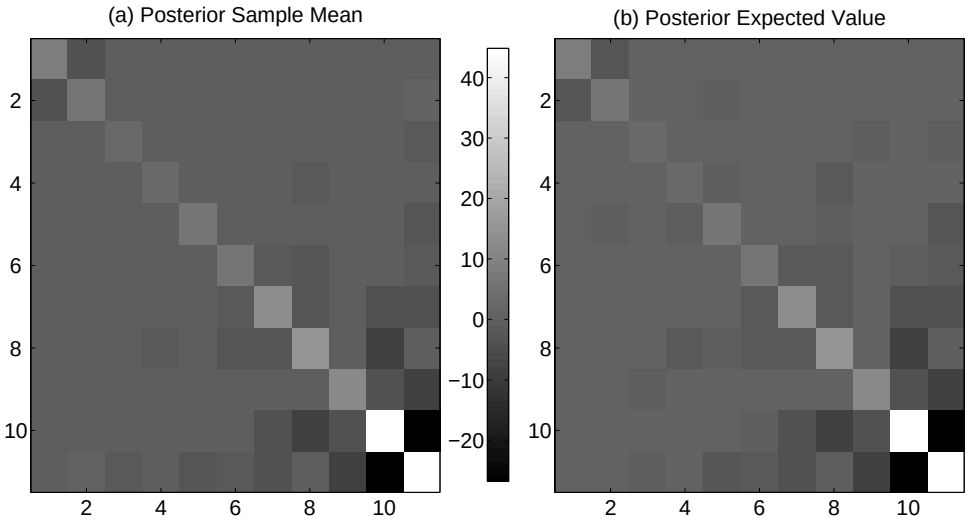

Figure 2: Exchange rate example. Grey-scale images of (a) the Markov chain Monte Carlo estimate of the posterior mean of Ω, and (b) the theoretically exact posterior mean of Ω.

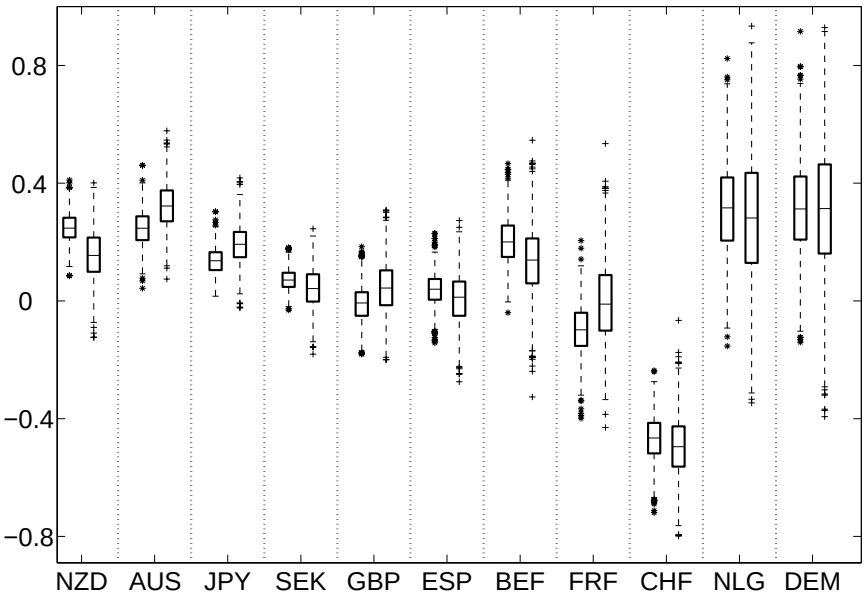

Figure 3: Exchange rate example. Boxplot summaries of posterior distributions of optimal portfolio weights a under G, first plot of each pair, and the competing full graph.

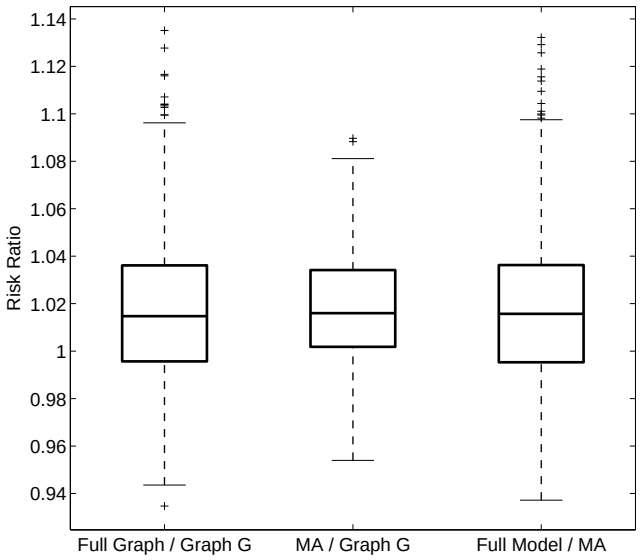

Figure 4: Exchange rate example. Posterior distributions for the ratios of standard deviations of the optimal portfolios under three strategies: the full graph relative to that under the graph G, the Bayesian model average (MA) over graphs relative to that on the graph G, and the full graph relative to the model average (MA).

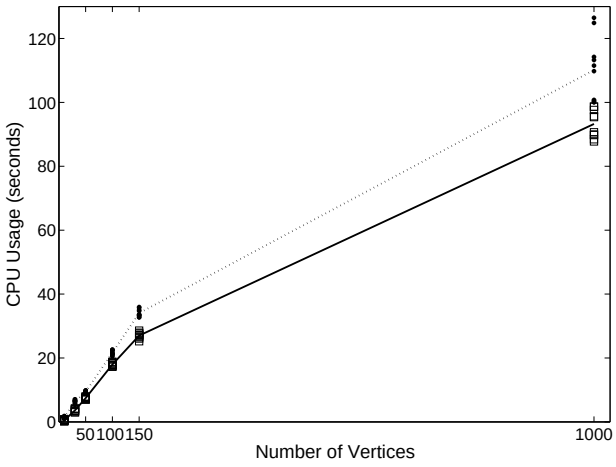

Figure 5: Computation time as a function of the size of graph. The graph shows the increase in cpu time to simulate the hyper-inverse Wishart distribution 100 times on a decomposable graph, and how the time changes as a function of the dimension, i.e. number of vertices. Graphs were generated randomly from posterior distributions over graphs using the Metropolis-Hastings algorithm described in Jones et al. (2005) and with subsets of data from the gene expression data used in their example. The figure here has upper, dotted, and lower, solid, lines represent differing degrees of sparsity: the upper cases correspond to graphs in which edges occur with prior probability $2/p$, and the lower those with probability $4/p$, where p is the number of vertices. The points and squares represent cpu times for specific simulated graphs.