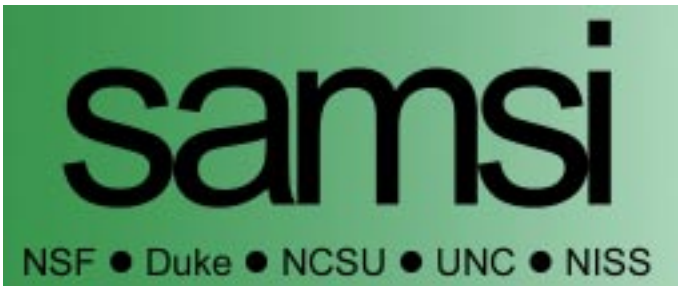


# Efficient Estimation of Covariance Selection Models

Frederick Wong, Christopher K. Carter
and Robert Kohn

Technical Report #2003-12
March 10, 2003

This material was based upon work supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

# Efficient Estimation of Covariance Selection Models

By Frederick Wong
Australian Graduate School of Management,
UNSW, Sydney, NSW 2052, Australia
email: fwong@agsm.edu.au
Christopher K. Carter
Institute of Statistics & Decision Sciences,
Box 90251, Duke University, Durham, NC 27708-0251,USA
email: chrisc@stat.duke.edu
and Robert Kohn
Australian Graduate School of Management,
UNSW, Sydney, NSW 2052, Australia
email: R.Kohn@unsw.edu.au

March 10, 2003

**Summary**

A Bayesian method is proposed for estimating an inverse covariance matrix from Gaussian data. The method is based on a prior that allows the off-diagonal elements of the inverse covariance matrix to be zero, and in many applications results in a parsimonious parameterisation of the covariance matrix. No assumption is made about the structure of the corresponding graphical model, so the method applies to both non-decomposable and decomposable graphs. All the parameters are estimated by model averaging using an efficient Metropolis-Hastings sampling scheme. A simulation study demonstrates that the method produces statistically efficient estimators of the covariance matrix, when the inverse covariance matrix is sparse. The methodology is illustrated by applying it to three examples that are high-dimensional relative to the sample size.

*Some key words*: Bayesian estimation; Gaussian graphical model; Multivariate regression; Model averaging; Partial correlation.

## 1 Introduction

Estimation of a covariance matrix is an important problem in multivariate analysis (Mardia et al., 1979) and in modelling longitudinal data (Lindsey, 1993), but it can be difficult, especially when the dimension of the covariance matrix, $p$, is large relative to the sample size, $n$; see for example Yang & Berger (1994) and Dempster (1969), who pointed out that estimators based on scalar multiples of the sample covariance matrix tend to distort the eigenstructure of the true covariance matrix, unless $p/n$ is small. The two main difficulties are that the number of unknown elements in the covariance matrix increases quadratically with $p$, and that it is difficult to deal directly with individual elements of the covariance matrix because it is necessary to keep the estimated matrix positive definite.

A number of approaches have been suggested for estimating a covariance matrix efficiently. Early work is by C. Stein, in several unpublished papers cited by Yang & Berger (1994), and Efron & Morris (1976). Yang & Berger (1994) used a Bayesian approach based on a spectral decomposition of the covariance matrix. Leonard & Hsu (1992) and Chiu et al. (1996) modelled the matrix logarithm of the covariance matrix. Pourahmadi (1999, 2000) estimated the covariance matrix by parameterising the Cholesky decomposition of its inverse. Smith & Kohn (2002) used a prior that allows for zero elements in the strict lower triangle of the Cholesky decomposition of the inverse of the covariance matrix to obtain a parsimonious representation of the covariance matrix. Although the Cholesky decomposition applies to a general covariance matrix, it is most useful and interpretable for longitudinal data. Barnard et al. (2000) modelled the covariance matrix in terms of standard deviations and correlations, proposed several shrinkage estimators and discussed suitable priors. Further results and simulation comparisons are given in Daniels & Kass (1999).

Dempster (1972) proposed to estimate the covariance matrix efficiently and parsimoniously by identifying zeros in its inverse, and called models estimated in this way covariance selection models. His idea was that in many problems the inverse covariance matrix has a large number of zeros in its off-diagonal elements and these should be exploited in the estimation. There is a natural interpretation of such zeros: if the $i, j$th element of the inverse is zero then the partial correlation between the $i$th and $j$th variables is zero; see Whittaker (1990) and the references therein.

The above result shows that a covariance selection model can be regarded as a Gaussian graphical

model; see Lauritzen (1996) and the references therein. Giudici & Green (1999) gave a Bayesian approach for estimating the structure of a decomposable Gaussian graphical model. For the case of non-decomposable Gaussian models, Roverato (2002) and Dellaportas et al. (2001) give Bayesian approaches that require importance sampling to estimate normalising constants for each possible graph. Since the number of possible graphs is $2^{p(p-1)/2}$ their methods are restricted to low-dimensional applications.

Our approach to covariance selection is Bayesian, using a prior that allows elements of the inverse to be zero. To make it easier to place a prior on the inverse covariance matrix we factorise it as a product of the inverse partial variances and the matrix of partial correlations, which is related to the way that Barnard et al. (2000) modelled the covariance matrix. Our prior for the partial correlation matrix has normalising constants for each possible graph size, rather than for each possible graph, which gives a tractable number of normalising constants that are calculated and stored offline. The methodology applies to non-decomposable as well as decomposable graphical models.

The computation is carried out using a reversible jump Metropolis-Hastings method, with the partial correlations generated one at a time in order to enforce efficiently the constraint that the covariance matrix is positive definite. We use Gaussian proposal densities in the Metropolis-Hastings method and show that they closely approximate the true conditional densities when the sample size $n$ is moderate or large. This makes our sampling method more efficient than the methods proposed by Giudici & Green (1999) and Barnard et al. (2000). The relative efficiency of our sampling scheme also means that we can estimate larger covariance matrices than most previous approaches; our largest example has $p = 81$.

A simulation study compares the statistical efficiency of the covariance selection method in our article with that of maximum likelihood and the Bayesian approach of Yang & Berger (1994). The simulation study considers several configurations of the inverse covariance matrix and two different loss functions. The results suggest that the covariance selection approach in our article compares favourably with the other two estimators when the inverse is sparse. The simulation study also looks at the bias of the estimators of the extreme eigenvalues.

The methodology in the article is illustrated using three examples that are high-dimensional relative to the sample size. The first two examples consider covariance selection in a multivariate regression setting. The first example estimates a multivariate stock return model, where the individual returns are adjusted for the market return. The second example is based on longitudinal data from biometry.

The third example carries out covariance selection for a first-order Markov random field model. In all three examples, the estimated inverse covariance matrix is highly parsimonious.

## 2 Estimation of the inverse covariance matrix

### 2·1 *Decomposition of the inverse covariance matrix*

To simplify the discussion, the observations in §2 are assumed to have a zero mean. Suppose the observations are generated as

$$y_t \sim N(0, \Sigma), \qquad t = 1, \ldots, n, \tag{2·1}$$

independently over $t$, where $y_t$ is a $p \times 1$ vector and $\Sigma$ is a $p \times p$ covariance matrix. Let $\Omega = \Sigma^{-1}$ be the inverse covariance matrix and write $\Omega = T \times C \times T$, where $T$ is a diagonal matrix and $C$ is a correlation matrix, i.e. $C_{ii} = 1$. Note that $T$ has entries $T_i = \Omega_{ii}^{1/2}$, so that $T_i^2$ is the inverse of the partial variance of $y_{i,t}$. Note also that the partial correlation coefficients $\rho^{ij}$ are given by

$$\rho^{ij} = -\Omega_{ij} / (\Omega_{ii}\Omega_{jj})^{1/2} = -C_{ij}, \tag{2·2}$$

so that $C$ contains the negative of the partial correlation coefficients.

Let $Y = (y_1', \ldots, y_n')'$. From (2·1) the likelihood of $\Omega$ is

$$\begin{aligned} p(Y|\Omega) &= \{\det(2\pi\Omega^{-1})\}^{-n/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{n} y_t'\Omega y_t\right) \\ &\propto (\det\Omega)^{n/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\,(\Omega S_y)\right\} \\ &= (\det T)^n (\det C)^{n/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\,(TCTS_y)\right\}, \end{aligned} \tag{2·3}$$

where $S_y = \sum_{t=1}^{n} y_t y_t'$.

To carry out a Bayesian analysis we place a prior on $\Sigma$ or, equivalently, on $\Omega$, by putting priors on $\Omega_{ii}, i = 1, \ldots, p$, and the elements $\{C_{ij}, i < j\}$, of $C$. We assume that in the prior the elements $\{\Omega_{ii}, i = 1, \ldots, p\}$ are independent and identically distributed, and are independent of the elements of $C$. We call this prior on $\Sigma$ a covariance selection prior because it allows the off-diagonal elements of $\Omega$ to be identically zero.

### 2·2 *Prior for the partial precisions* $\Omega_{ii}$

The prior for $\Omega_{ii}$ is a gamma distribution with parameters $\alpha$ and $\beta$, so that

$$p(\Omega_{ii}) \propto \Omega_{ii}^{\alpha-1} \exp(-\beta\Omega_{ii}),$$

implying that the prior for $T_i$ is

$$\begin{aligned} p(T_i) &\propto p(\Omega_{ii}) \frac{d\Omega_{ii}}{dT_i} \\ &\propto T_i^{2\alpha-1} \exp(-\beta T_i^2). \end{aligned} \tag{2·4}$$

To make the prior uninformative, $\alpha$ and $\beta$ are chosen to be small. In the applications in §§4 and 5, we take $\alpha = 10^{-10}$ and $\beta = 10^{-8}$.

### 2·3 *Prior for the partial correlation matrix*

We introduce the following notation to specify the prior for the $C_{ij}, i < j$. Let $J_{ij}$ be the event defined by

$$J_{ij} = \begin{cases} 0 & \text{if } C_{ij} = 0 \\ 1 & \text{otherwise} \end{cases}$$

and let $J$ be the set of $J_{ij}, i < j$. Let

$$S(J) = \text{the number of the } J_{ij} \text{ that are equal to 1, } i < j,$$

and let $r = p(p-1)/2$, so that $0 \le S(J) \le r$. Let $\mathcal{C}_p$ be the set of $p \times p$ correlation matrices. Let

$$V(J^*) = \int_{C \in \mathcal{C}_p : J(C) = J^*} \left( \prod_{i<j, J_{ij}=1} dC_{ij} \right)$$

be the volume of the positive definite region for $C$, given the constraints imposed by $J$, and let

$$\overline{V}(l) = \binom{r}{l}^{-1} \left( \sum_{J : S(J) = l} V(J) \right)$$

be the average volume for regions with dimension $l$.

The prior for $C$ can be expressed in hierarchical form as

$$p(dC|J) = V(J)^{-1} dC_{\{J=1\}} I(C \in \mathcal{C}_p), \tag{2·5}$$

$$p\{J|S(J) = l\} = \binom{r}{l}^{-1} \frac{V(J)}{\overline{V}(l)}, \tag{2·6}$$

$$p\{S(J) = l|\psi\} = \binom{r}{l} \psi^l (1-\psi)^{r-l}, \tag{2·7}$$

where $0 \leq \psi \leq 1$, $I(C \in \mathcal{C}_p) = 1$ if $C$ is a correlation matrix and is 0 otherwise and $C_{\{J=1\}} = \{C_{ij} : C_{ij} \not\equiv 0\}$. The parameter $\psi$ is the probability that $J_{ij} = 1$ and its prior can be specified by the user. However, in this article we take $p(\psi) = 1$.

Appendix 1 describes an approximation for $\overline{V}(l)$, which we find gives good results for values of $l$ ranging from very small to very large.

*Remark 1.* Note that in (2·6) $p(J|S(J) = l)$ is not uniform, but is proportional to $V(J)/\overline{V}(l)$. If the prior $p(J|S(J) = l)$ is taken as uniform, then it is necessary to evaluate the volumes $V(J)$ for all configurations $J$. This would be computationally intractable because most volumes need to be evaluated using simulation and there are $2^{p(p-1)/2}$ such volumes for a given $p$.

To generate $C_{ij}$ it is convenient to rewrite the prior for $C$ as

$$\begin{aligned} p(dC|\psi) &= p[dC|J(C), S\{J(C)\}, \psi]p[J(C)|S\{J(C)\}, \psi]p(S(J(C))|\psi) \\ &= \frac{I(C \in \mathcal{C}_p)}{\overline{V}[S\{J(C)\}]} \prod_{i<j} \{I(0 \in dC_{ij})(1-\psi) + I(0 \notin dC_{ij})\psi dC_{ij}\} \ . \end{aligned} \tag{2·8}$$

## 2·4 *Sampling scheme*

Let $\theta = (T_i, i = 1, \ldots, p, C_{ij}, i < j)$ be the vector of unknown parameters. Then we generate a sample from the posterior distribution $p(d\theta|Y)$ by using the Markov chain Monte Carlo sampling scheme based on the following Metropolis-Hastings proposals:

$$q\left(T_i|Y, T_{\{-i\}}, C\right) \qquad \text{for } i = 1, \ldots, p \tag{2·9}$$

$$q\left(dC_{ij}|Y, T, C_{\{-ij\}}\right) \qquad \text{for all pairs } i < j \ . \tag{2·10}$$

In (2·9) and (2·10) the parameter $\psi$ is integrated out and not generated. The $T_i$ are generated one at a time using a Gaussian proposal. The $C_{ij}$ are generated one at a time using a Metropolis-Hastings proposal that allows $C_{ij}$ to be identically zero, and that uses a Gaussian proposal for the continuous part of the conditional density. An important part of this Metropolis-Hastings proposal is the result in part (b) of Lemma 2 below, which states that $\det(C)$ is a quadratic in $C_{ij}$, which restricts $C_{ij}$ to an interval when it is combined with the constraint that $\det(C) > 0$.

## 2·5 *Generating the $T_i$*

From (2·3) and (2·4), the conditional density of $T_i$ is

$$\begin{aligned} p(T_i|Y, T_{\{-i\}}, C) &\propto p(Y|T, C)p(T_i) \\ &\propto T_i^n \exp\left[-\frac{1}{2}\left\{T_i^2(S_y)_{ii} + 2T_i\sum_{j\neq i}^{p}(S_y)_{ij}C_{ij}T_j\right\}\right] T_i^{2\alpha-1}\exp(-\beta T_i^2) \\ &\propto T_i^{n_\alpha}\exp(-aT_i^2 - 2bT_i), \end{aligned} \tag{2·11}$$

where $n_\alpha = n + 2\alpha - 1$, $a = (S_y)_{ii}/2 + \beta$ and $b = \frac{1}{2}\sum_{j\neq i}(S_y)_{ij}C_{ij}T_j$. The coefficient $b$ can be computed efficiently as

$$b = \frac{1}{2}\left\{\sum_{j=1}^{p}(S_y)_{ij}C_{ij}T_j - (S_y)_{ii}C_{ii}T_i\right\}.$$

**Lemma 1** . *The conditional density of $T_i$ in the expression above tends to normality as $n \to \infty$.* The proof is in Appendix 2.

The conditional density $T_i$ is approximated by a Gaussian density using a second order Taylor series expansion about the mode of the log conditional density. Consider the equation

$$f(x) = n_\alpha \log(x) - ax^2 - 2bx.$$

Then, $f'(x) = n_\alpha/x - 2ax - 2b$, which has the two real roots, the positive one of which gives the mode of the conditional density of $T_i$, namely

$$\hat{T}_i = \frac{1}{2a}\left\{-b + \sqrt{(b^2 + 2an_\alpha)}\right\}.$$

The second derivative of $f(x)$ is

$$f''(x) = -n_\alpha/x^2 - 2a$$

so $f''(x) < 0$ for all $x$. Let $\sigma_T^2 = -1/f''(\hat{T}_i)$. The function $f(T_i)$ is approximated by its second order Taylor series expansion about $\hat{T}_i$, which is

$$f_a(T_i) = f(\hat{T}_i) - \frac{1}{2\sigma_T^2}(T_i - \hat{T}_i)^2.$$

A Gaussian density with mean $\hat{T}_i$ and variance $\sigma_T^2$ is used as the Metropolis-Hastings proposal density $q(T_i|Y, T_{\{-i\}}, C)$ in (2·9). The acceptance probability is then calculated using (2·11).

### 2·6 *Generating the partial correlation matrix terms* $C_{ij}$

This section first discusses the conditional distribution $p(dC_{ij}|Y,T,C_{\{-ij\}})$ before describing the Metropolis-Hastings step in (2·10). A reversible jump method similar to that of Green (1995) is used because the conditional distribution $p(dC_{ij}|Y,T,C_{\{-ij\}})$ is a mixture of a continuous and a degenerate distribution.

From (2·3), the conditional distribution of $C_{ij}$ is

$$\begin{aligned} p(dC_{ij}|Y,T,C_{\{-ij\}}) &\propto p(Y|T,C)p(dC_{ij}|C_{\{-ij\}}) \\ &\propto (\det C)^{n/2}\exp\left\{-\frac{1}{2}\operatorname{tr}(TCTS_y)\right\} \end{aligned} \tag{2·12}$$

The following four lemmas simplify the expression for the conditional distribution. The proofs of all four lemmas are in Appendix 2. The first lemma assumes that $i = p$ and $j = p-1$. There is no loss of generality in making this assumption because the likelihood and the prior are invariant to permutations of the indices. The lemma uses the Cholesky decomposition of $C$, which can be updated efficiently after permuting the indices using Givens rotations; see for example the functions qrinsert and qrdelete in Matlab.

**Lemma 2** . *Suppose that $i = p$ and $j = p-1$, and let $C = R'R$ be the Cholesky decomposition of $C$ with the matrix $R$ upper triangular. Then*

$$(i) \qquad \operatorname{tr}(TCTS_y) = 2C_{ij}T_iT_j(S_y)_{ij}\,,$$

*plus terms that do not depend on $C_{ij}$;*

$$(ii) \qquad \det C \propto c - (C_{ij}-a)^2/b^2$$

*where $a, b$ and $c$ do not depend on $C_{ij}$ and are given by*

$$a = \sum_{j=1}^{p-2} R_{j,p-1}R_{j,p}, \qquad b = R_{p-1,p-1}, \qquad c = C_{pp} - \sum_{j=1}^{p-2} R_{j,p}^2;$$

*(iii) The conditional distribution of $C_{ij}$ is given by*

$$p(dC_{ij}|Y,T,C_{\{-ij\}}) \propto \left\{1 - \frac{(C_{ij}-a)^2}{cb^2}\right\}^{n/2} \exp(-\gamma C_{ij})p\left(dC_{ij}|C_{\{-ij\}}\right)$$

*where $\gamma = T_iT_j\left(S_y\right)_{ij}$.*

Note that a fast way of computing $c$ in part (ii) is to note that $c = R^2_{p-1,p} + R^2_{p,p}$ which is independent of $C_{ij}$.

Barnard et al. (2000) show that $\det C$ is a quadratic in $C_{ij}$, but do not provide a fast way of calculating the quadratic.

The second lemma gives an explicit expression for the distribution $p(dC_{ij}|C_{\{-ij\}})$ as a mixture of a uniform and a degenerate distribution.

**Lemma 3** .

$$p(dC_{ij}|C_{\{-ij\}}) = I(|C_{ij} - a| < b\sqrt{c})\, \frac{I(C_{ij} = 0) + (dC_{ij}) h\left(J_{\{-ij\}}\right)}{I(a < b\sqrt{c}) + 2(b\sqrt{c}) h\left(J_{\{-ij\}}\right)} ,$$

*where* $S\left(J_{\{-ij\}}\right)$ *is the number of* $J_{\{-ij\}}$ *that are equal to 1 and*

$$h\left(J_{\{-ij\}}\right) = \frac{S\left(J_{\{-ij\}}\right) + 1}{r - S\left(J_{\{-ij\}}\right)} ,$$

*which is approximately the empirical ratio of 1's to 0's.*

Let

$$g(C_{ij}) = \left\{1 - \frac{(C_{ij} - a)^2}{cb^2}\right\}^{n/2} \exp(-\gamma C_{ij}).$$

Combining Lemmas 2 and 3 gives the third lemma.

**Lemma 4** .

$$p(dC_{ij}|Y, T, C_{\{-ij\}}) \quad \propto \quad I(|C_{ij} - a| < b\sqrt{c})\, g(C_{ij}) \left\{I(C_{ij} = 0) + dC_{ij} h(J_{\{-ij\}})\right\}$$

The fourth lemma gives a normal approximation to the continuous component of $p(dC_{ij}|Y, T, C_{\{-ij\}})$. It is convenient to transform $C_{ij}$ to $v = \sqrt{n}(C_{ij} - a)/\left(b\sqrt{c}\right)$.

**Lemma 5** . *(i) The conditional density of* $v$ *is given by*

$$p(v|Y, T, C_{\{-ij\}}, J_{ij} = 1) \propto I(|v| < \sqrt{n})(1 - v^2/n)^{n/2} \exp(-\delta v) ,$$

*where* $\delta = \gamma b\sqrt{(c/n)}$.
*(ii) The mode of* $p(v|Y, T, C_{\{-ij\}}, J_{ij} = 1)$ *is*

$$\hat{v} = \frac{n - \sqrt{(4n\delta^2 + n^2)}}{2\delta}.$$

*(iii) Let*

$$f(v) = \frac{n}{2}\log\left(1 - \frac{v^2}{n}\right) - \delta v.$$

*Then $f(v)$ is the log of the conditional density of $v$, up to additive constants,*

$$f'(v) = \frac{-v}{1 - v^2/n} - \delta \qquad \textit{and} \qquad f''(v) = -\frac{(1 + v^2/n)}{(1 - v^2/n)^2},$$

*so that $f''(v) < 0$ for all $v$.*

*(iv) As $n \to \infty$, the conditional density of $v$ tends to that of a normal distribution with mean $\hat{v}$ and variance $\sigma_v^2 = -1/f''(\hat{v})$.*

To construct the Metropolis-Hastings proposal in (2·10), we use a normal approximation to the continuous component of $p(dC_{ij}|Y, T, C_{\{-ij\}})$. Let

$$f_a(v) = f(\widehat{v}) - \frac{1}{2\sigma_v^2}(v - \widehat{v})^2.$$

By Lemma 5 parts (ii) and (iii), $f_a(v)$ is the second order Taylor series approximation to $f(v)$, which should be accurate by part (iv). Let

$$g_a(C_{ij}) = \exp\left[f_a\left\{\frac{\sqrt{n}\,(C_{ij} - a)}{b\sqrt{c}}\right\}\right].$$

The Metropolis-Hastings proposal for generating $C_{ij}$ is given by

$$\begin{aligned} q(dC_{ij}|Y, T, C_{\{-ij\}}) \quad &\propto \quad I(|C_{ij} - a| < b\sqrt{c}) \\ &\times \quad \left\{I(C_{ij} = 0)g(0) + g_a(C_{ij})dC_{ij}h\left(J_{\{-ij\}}\right)\right\}. \end{aligned} \tag{2·13}$$

To describe the acceptance probabilities it is convenient to define the densities of $p(dC_{ij}|Y, T, C_{\{-ij\}})$ and $q(dC_{ij}|Y, T, C_{\{-ij\}})$ with respect to the measure

$$\xi(dC_{ij}|C_{-ij}) = I(C_{ij} = 0)g(0) + dC_{ij}h\left(J_{\{-ij\}}\right).$$

The densities are given by

$$p(C_{ij}|Y, T, C_{\{-ij\}}) \quad \propto \quad I(|C_{ij} - a| < b\sqrt{c})\ \{I(C_{ij} = 0) + I(C_{ij} \neq 0)g(C_{ij})\} \tag{2·14}$$

$$q(C_{ij}|Y, T, C_{\{-ij\}}) \quad \propto \quad I(|C_{ij} - a| < b\sqrt{c})\ \{I(C_{ij} = 0) + I(C_{ij} \neq 0)g_a(C_{ij})\}\ . \tag{2·15}$$

Then the Metropolis-Hastings acceptance probability for the move $C_{ij}^{\mathrm{c}} \longrightarrow C_{ij}^{\mathrm{pr}}$ is given by

$$\alpha\left(C_{ij}^{\mathrm{c}} \to C_{ij}^{\mathrm{pr}}\right) = \min\left\{1, \frac{p(C_{ij}^{\mathrm{pr}}|Y, T, C_{\{-ij\}})}{p(C_{ij}^{\mathrm{c}}|Y, T, C_{\{-ij\}})} \times \frac{q(C_{ij}^{\mathrm{c}}|Y, T, C_{\{-ij\}})}{q(C_{ij}^{\mathrm{pr}}|Y, T, C_{\{-ij\}})}\right\}.$$

From (2·14) and (2·15), this probability can be evaluated as follows

$$\begin{aligned}
\alpha\left(0 \to 0\right) &= 1 \\
\alpha\left(0 \to C_{ij}^{\mathrm{pr}} \neq 0\right) &= \min\left\{1, \frac{g\left(C_{ij}^{\mathrm{pr}}\right)}{g_a\left(C_{ij}^{\mathrm{pr}}\right)}\right\} \\
\alpha\left(C_{ij}^{\mathrm{c}} \neq 0 \to\right) &= \min\left\{1, \frac{g_a\left(C_{ij}^{\mathrm{c}}\right)}{g\left(C_{ij}^{\mathrm{c}}\right)}\right\} \\
\alpha\left(C_{ij}^{\mathrm{c}} \neq 0 \to C_{ij}^{\mathrm{pr}} \neq 0\right) &= \min\left\{1, \frac{g\left(C_{ij}^{\mathrm{pr}}\right) g_a\left(C_{ij}^{\mathrm{c}}\right)}{g\left(C_{ij}^{\mathrm{c}}\right) g_a\left(C_{ij}^{\mathrm{pr}}\right)}\right\} .
\end{aligned}$$

The Metropolis-Hastings proposal given above is an example of reversible jump Markov chain Monte Carlo (Green, 1995) because the spaces defined by $J_{ij} = 0$ and $J_{ij} = 1$ have different dimensions.

Once $C_{ij}$ is generated and accepted, it is necessary to update the Cholesky decomposition. This is done as follows. Let $C_{ij}^{\mathrm{N}}$ be the new value of $C_{ij}$, let $C_{ij}^{\mathrm{c}}$ be the current value, and let $\Delta = C_{ij}^{\mathrm{N}} - C_{ij}^{\mathrm{c}}$ be the change in $C_{ij}$. If $\Delta = 0$ then there is no need to update the Cholesky factor $R$. If $\Delta > 0$, let $w$ be a $p \times 1$ vector, which is 0 except for entries $\sqrt{\Delta}$ in the $i$th and $j$th positions, let $w_1$ be a $p \times 1$ vector which is 0 except for any entry $\sqrt{\Delta}$ in the $i$th position and let $w_2$ be a $p \times 1$ vector which is 0 except for any entry $\sqrt{\Delta}$ in the $j$th position. Then

$$C^{\mathrm{N}} = \{(C^{\mathrm{c}} + ww') - w_1 w_1\} - w_2 w_2'$$

and the Cholesky can be updated efficiently using the cholupdate function in Matlab, or the corresponding code in Fortran. If $\Delta < 0$ then $w, w_1$ and $w_2$ are defined similarly, except that $\sqrt{\Delta}$ is replaced by $\sqrt{(-\Delta)}$. Now we write

$$C^{\mathrm{N}} = \{(C^{\mathrm{c}} + w_1 w_1') + w_2 w_2\} - ww'.$$

## 3 The Linear Regression Model

This section adds a linear regression mean to the model (2·1). Suppose that $y_{i,t}$ is the $i$th component of the $t$th observation, so that

$$y_{i,t} = x_{i,t}'\beta_i + e_{i,t}, \qquad t = 1, \ldots n, \quad i = 1, \ldots p,$$

where $x_{k,t}$ is a $m \times 1$ vector of regressors and $\beta_k$ is $m \times 1$ vector of coefficients. By stacking the $p$ equations we obtain

$$y_t = X_t\beta + e_t, \qquad e_t \sim N(0, \Sigma), \qquad t = 1, 2, \ldots n, \tag{3·1}$$

where $e_t = (e_{1,t}, \ldots, e_{p,t})'$, $\beta = \left(\beta_1', \ldots \beta_p'\right)'$, and $X_t = \text{diag}(x_{1,t}', \ldots, x_{p,t}')$ is a $p \times mp$ matrix; diag() means the block-diagonal matrix formed from the $x_{k,t}'$. Thus, we can write

$$Y = X\beta + e \tag{3·2}$$

where $X = (X_1', \ldots, X_n')'$ and $e \sim N(0, I \times \Sigma)$.

For a given value of $\Sigma$, let $\Omega = \Sigma^{-1}$ and let $\hat{\beta} = \{X'(I \otimes \Omega)X\}^{-1} X'(I \otimes \Omega)Y$ be the generalised least squares estimator of $\beta$. We take the prior for $\beta$, conditional on $\Sigma$, as

$$\beta|\Sigma \sim N(\hat{\beta}, c^{-1}I)$$

where $c$ is very small. In our application we take $c = 10^{-10}$; that is, the prior is centred at the generalised least squares estimate, but is diffuse.

With this prior, the distribution of $\beta$, conditional on $Y$ and $\Sigma$, is

$$\beta|Y, \Sigma \sim N[\hat{\beta}, \{X'(I \otimes \Omega)X + cI\}^{-1}]. \tag{3·3}$$

In the full sampling scheme, the matrices $C$ and $T$ are generated as in § 2, conditional on $\beta$, and $\beta$ is generated conditional on $C$ and $T$ as in (3·3).

Generating $\beta$ from its conditional distribution requires us to compute the matrix $X'(I \otimes \Omega)X$ and the vector $X'(I \otimes \Omega)Y$. Direct computation of these terms is slow when the sample size $n$ is large because $X$ is $np \times mp$ and $Y$ is $np \times 1$. Appendix 3 outlines a fast method of evaluating these terms that is independent of sample size.

## 4 Comparison with other Estimators

This section gives a simulation study comparing several Bayes estimators of $\Sigma$ using priors from §2 with the maximum likelihood estimator and with several Bayes estimators of $\Sigma$ using the prior in Yang & Berger (1994).

Yang & Berger (1994) used two loss functions, $L_1$ and $L_2$, defined by

$$\begin{aligned} L_1(\widehat{\Sigma}, \Sigma) &= \text{tr}(\widehat{\Sigma}\Sigma^{-1}) - \log|\widehat{\Sigma}\Sigma^{-1}| - p, \\ L_2(\widehat{\Sigma}, \Sigma) &= \text{tr}\{(\widehat{\Sigma}\Sigma^{-1} - I)^2\}, \end{aligned}$$

where $\widehat{\Sigma}$ denotes an arbitrary estimator of a $p \times p$ matrix $\Sigma$ and $I$ is the $p \times p$ identity matrix. The loss function $L_1$ was used by C. Stein in a Berkeley technical report cited by Yang & Berger (1994).

Both $L_1$ and $L_2$ are non-negative and are equal to zero when $\widehat{\Sigma} = \Sigma$, so small values of $L_1$ or $L_2$ indicate good performance.

Yang & Berger (1994) show that the Bayes estimator for $L_1$ is

$$\widehat{\Sigma}_{L_1} = \{E(\Sigma^{-1}|Y)\}^{-1}$$

and the Bayes estimator for $L_2$ is

$$\text{vec}(\widehat{\Sigma}_{L_2}) = \{E(\Sigma^{-1} \otimes \Sigma^{-1}|Y)\}^{-1}\text{vec}\{E(\Sigma^{-1}|Y)\}.$$

Both $\widehat{\Sigma}_{L_1}$ and $\widehat{\Sigma}_{L_2}$ can be estimated from a Markov chain Monte Carlo sample from the posterior distribution of $\Sigma$.

We also consider the posterior mean $\widehat{\Sigma}_{PM} = E(\Sigma|Y)$, which is the Bayes estimator for squared error loss.

Yang & Berger (1994) use the following reference prior for $\Sigma$:

$$p_R(\Sigma) \propto |\Sigma|^{-1} \prod_{i<j} (d_i - d_j)^{-1},$$

where $d_i, d_j$ are eigenvalues of $\Sigma$ with $d_i \geq d_j$. With this reference prior, the posterior distribution of $\Sigma$ does not have a standard form and must be sampled using Markov chain Monte Carlo; see Yang & Berger (1994) for details.

Six inverse covariance matrix structures were used in the simulation study. They were, in decreasing order of sparsity, the identity, tridiagonal, loop, block-diagonal, 'overlapping block' and the full matrices. The loop matrix has the form of a tridiagonal matrix plus two nonzero elements $\Omega_{1p} = \Omega_{p1}$. The 'overlapping block' matrix is formed by placing 2 full sub-blocks $A$ and $B$ in the main diagonal, where $A$ is $a$ by $a$, and making $A_{a,a}$ and $B_{1,1}$ overlap. Although its structure is close to that of a block-diagonal matrix, its inverse, $\Sigma$, is full. In the simulations, we chose matrices of dimension $p = 10$. In each set of simulations, the matrix $\Sigma^{-1}$ was kept the same. In the case of full, block-diagonal and 'overlapping block' matrices $\Sigma^{-1}$, all the block matrices were generated from a Wishart distribution with 10 degrees of freedom, scaled down by 10, i.e. $0.1 \times W(p, 10)$. In the block-diagonal case, $4 \times 4$ and $6 \times 6$ blocks were used. In the 'overlapping block' case, the $5 \times 5$ block overlapped with the $6 \times 6$ block. The element $\Omega_{5,5}$ was chosen to be the average of the overlapping elements of these two sub blocks. In the tridiagonal case, $\Sigma^{-1}$ was obtained by extracting the tridiagonal band from the full matrix generated from the same Wishart distribution. To construct the loop matrix we

took the tridiagonal matrix and then set the (1,10) and (10,1) elements to 0.3, to make the size of these elements comparable to the nonzero off-diagonal elements in the tridiagonal matrix. For the tridiagonal matrix, the largest absolute value of the off-diagonal elements is 0.388 and the largest diagonal element is 1.275. Figure 1 gives the image plots of all six matrices that are used in the simulation. In the image plots, the lightest colour corresponds to elements of the matrix that have minimum absolute value and the darkest colour corresponds to elements of the matrix that have maximum absolute value. Shades of grey correspond to interpolated values in between the minimum and the maximum.

——————— Figure 1 about here ———————

We considered six estimators of $\Sigma$. The first three are Bayes estimators based on the covariance selection prior for the loss functions $L_1$, denoted by $\widehat{\Sigma}_{ESL_1}$, $L_2$, denoted by $\widehat{\Sigma}_{ESL_2}$, and squared error loss, denoted by $\widehat{\Sigma}_{ESPM}$, so that $\widehat{\Sigma}_{ESPM}$ is the posterior mean. The fourth estimator is the maximum likelihood estimator, $\widehat{\Sigma}_{ML}$. The fifth and sixth estimators are the Bayes estimators based on the prior from Yang & Berger (1994), $\widehat{\Sigma}_{YBL_1}$ for the loss function $L_1$, and $\widehat{\Sigma}_{YBL_2}$ for $L_2$. All the comparisons are based on the $L_1$ and $L_2$ loss functions.

One hundred replications were carried out for each choice of $\Sigma^{-1}$ using a sample size of $n = 100$. For each replication, there were 10,000 warm-up iterations, while 30,000 iterations were used for inference. For each set of 100 replications, all the other estimators were compared to the Bayes estimator $\widehat{\Sigma}_{ESL_1}$ when using the $L_1$ distance and to $\widehat{\Sigma}_{ESL_2}$ when using the $L_2$ distance. This was done by taking differences in the log scale; for example, for the $L_1$ distance, we computed

$$\log\{L_1(\widehat{\Sigma}, \Sigma)\} - \log\{L_1(\widehat{\Sigma}_{ESL_1}, \Sigma)\}, \tag{4·1}$$

where $\widehat{\Sigma}$ stands for any estimator of $\Sigma$ and $\widehat{\Sigma}_{ESL_1}$ is as described above. Thus, a boxplot that compares $\widehat{\Sigma}$ and $\widehat{\Sigma}_{ESL_1}$ has the following interpretation: values above zero correspond to cases where $\widehat{\Sigma}$ has worse performance than $\widehat{\Sigma}_{ESL_1}$ and values below zero correspond to the opposite case. Similar pairwise comparisons were made using the loss function $L_2$.

Figure 2 presents the results and shows that when the inverse is sparse the estimators based on the covariance selection prior outperform the maximum likelihood estimator and the corresponding estimators based on the Yang and Berger prior. For the full matrix, the estimators based on the covariance selection prior performed similarly to the maximum likelihood estimator and the estimators based on the Yang and Berger prior. The figure also shows that the Bayes estimator for the $L_1$,

respectively $L_2$, loss function generally outperforms the posterior mean estimator in terms of $L_1$, respectively $L_2$, loss.

———————— Figure 2 about here ————————-

An important issue in covariance matrix estimation is the bias of the estimators of the extreme eigenvalues of the covariance matrix. It is known from the work of C. Stein, cited by Yang & Berger (1994) and Dempster (1969), that the extreme eigenvalues of the maximum likelihood estimator are biased, with the smallest eigenvalues biased downwards and the largest biased upwards. Various authors including C. Stein in unpublished work cited by Yang & Berger (1994), Haff (1991) and Yang & Berger (1994) have attempted to correct for this bias.

We now report on a small simulation study that compared the bias in the minimum and maximum eigenvalues of the maximum likelihood estimator, and the estimators based on the covariance selection prior and the reference prior used by Yang & Berger (1994) . Although the study is limited, it suggests that the Bayes estimators using the covariance selection prior does not produce extremely distorted eigenvalues, nor do the estimators perform badly relative to the reference prior in Yang & Berger (1994).

There are three Bayes estimators corresponding to the covariance selection prior and the Yang & Berger (1994) prior: the posterior mean and the Bayes estimators for the $L_1$ and $L_2$ loss functions. The study considered four $10 \times 10$ covariance matrices. The first is an identity matrix representing a covariance matrix with equal eigenvalues or more generally with eigenvalues that are very close to each other. The second is a diagonal matrix with eigenvalues increasing in powers of 2 from $2 = 2^1$ to $2^{10} = 1024$, representing a sparse and ill-conditioned inverse covariance matrix. The third is a covariance matrix whose inverse is full and has nine eigenvalues that are equal to one and one eigenvalue that is equal to nine. The matrix was constructed as follows. First a matrix was generated from a Wishart with 10 degrees of freedom and then scaled by $1/10$. This matrix was then factored as $QDQ'$, where $Q$ is an orthonormal matrix and $D$ is a diagonal matrix, and the diagonal elements of $D$ were then replaced by nine ones and a nine. The fourth matrix has an inverse that is full. It was generated from a Wishart with ten degrees of freedom and then scaled by $1/10$. The study used $n = 100$ observations and fifty replications for each estimator, with the constructed covariance matrix held constant for all 50 replications.

Figure 3 presents boxplots of the smallest and largest eigenvalues for each estimator for each

type of covariance matrix. The solid horizontal lines show the true values of the eigenvalues. The figure shows that for the identity matrix, all estimators underestimate the minimum eigenvalue and overestimate the maximum eigenvalue. We obtained similar results when all the eigenvalues were close to each other. For the ill-conditioned diagonal matrix, the estimators based on the covariance selection prior outperform those based on the Yang & Berger (1994) prior. For the third matrix, most of the estimators seem to underestimate both the maximum and minimum eigenvalues. For the fourth matrix, the minimum eigenvalue seems to be overestimated and the maximum eigenvalue seems to be underestimated.

We conclude that except for the first covariance matrix, there is no consistent pattern of bias for the estimators of the extreme eigenvalues. Furthermore, the estimators based on the covariance selection prior seem to do as well as the estimators based on the Yang & Berger (1994) prior.

———————— Figure 3 about here ————————-

## 5 Examples

### 5·1 *Multivariate Capital Asset Pricing Model*

The first example considers the estimation of a multivariate Capital Asset Pricing Model, which states that the excess stock returns from individual firms are linearly related to the excess market return, without an intercept (Sharpe, 1964; Campbell et al., 1997). The errors represent the variability of each excess stock return that is unrelated to the variability in the market return. Efficient estimates of the covariance matrix are important in this application because they help market practitioners form optimal stock portfolios that minimise variance for a given return or maximise return for a given variance.

The model takes the form

$$y_t = \beta x_t + e_t \tag{5·1}$$

(Gibbons, 1982), where $y_t$ is the $p \times 1$ vector of monthly excess returns at time $t$, $\beta$ is the $p \times 1$ vector of regression coefficients, $x_t$ is the excess market return at time $t$, and $e_t \sim N(0, \Sigma)$. The $e_t$ are assumed independent.

We use monthly excess returns for the $p = 81$ firms on the S&P 100 that traded continuously

between November 1976 and November 1996. The excess stock return is defined as the nominal return minus the risk-free rate; following standard practice, we use the one month Treasury bill rate as a proxy for the risk-free rate. The market return is the value-weighted market return on the NASDQ/AMEX/NYSE exchanges for any given month. The excess market return is defined similarly to the excess stock return. All the data are from the Center for Research in Securities data base.

Figure 4 summarises the results of the estimation. Figure 4(a) shows an 'image' of the estimates of the probabilities $\mathrm{pr}(J_{ij} = 1|Y)$, with the shading being lightest for 0 and darkest for 1. Here $J_{ij} = 0$ if $\Omega_{ij} = 0$ and $J_{ij} = 1$ if $\Omega_{ij} \neq 0$. Although possibly not obvious from the plot, the matrix $\widehat{\Omega}$ is sparse, with the estimated proportion of nonzero elements in the off-diagonal part of $\Omega$ being $\sum_{k=1}^{M} S(J^{[k]})/Mr = 0.0446$, where $S(J^{[k]})$ is the number of $\gamma_{ij} = 1$ in the $k$th iteration, for $k = 1, 2, \ldots, M$, in the $M = 10,000$ sampling periods, and $r = p(p-1)/2$ is the number of elements in the strict lower triangle of $\Omega$. Panels (b) and (c) show the 'image' plots of the absolute value of each element of $\widehat{\Sigma}^{-1}_{ESL_1}$ and $\widehat{\Sigma}^{-1}_{ML}$. Although both estimates show a similar structure, $\widehat{\Sigma}^{-1}_{ESL_1}$ clearly has more near-zero values.

———————— Figure 4 about here ————————-

The model was estimated using 10,000 iterations for both burn-in and inference, and this gave agreement to three significant figures of the estimates when $\Omega$ was initialised from different starting positions. The acceptance rate of the Metropolis-Hastings step was 0.97. We used an SGI Origin 2000 computer with a 250 MHZ R10000 IP27 processor, which took 8 hours to estimate the model. The program was written in Fortran 90.

We also studied the 10,000 iterates of the loglikelihood and selected elements of $\Omega$ during the inference period. The elements of $\Omega$ that were studied were a diagonal element, a non-diagonal element that was close to zero and a non-diagonal element that was definitely nonzero. In all cases, the time sequence plots showed a stable variation about a mean, while the corresponding plots of the autocorrelations died out after 10 lags at most. The same results were also obtained when the sampler was initialised at different starting positions. The results suggest that for this example, the sampler converged and mixed well.

### 5·2 *Repeated measures longitudinal data*

This section applies covariance selection to data from a repeated measures longitudinal study in which a sequence of measurements is taken on each of a number of subjects. Most of the literature models the within-subject correlation using a simple parametric model, such as a first-order autoregressive model, while the measurements on different subjects are assumed independent; see for example Lindsey (1993) and Diggle et al. (1994).

Our application involves longitudinal data obtained from an experiment reported by Kenward (1987) that compared two treatments for controlling internal parasites in cattle. In the experiment 60 animals were randomly assigned to one of the two treatment groups, each of size 30. The animals were put out to pasture at the start of the grazing season and the weight of each animal was then recorded 11 times to the nearest kilogram. The first 10 measurements were taken at two-weekly intervals, and the final measurement was made after a one-week interval. Thus, the 11 observations were taken on the days $0, 14, 28, \ldots, 112, 126$ and 133, after the start of the experiment, with the data given in Table 1 of Kenward (1987).

For each of the groups, $g = 1, 2$, we modelled the mean as a quadratic trend,

$$f_g(t) = \beta_0 + \beta_{1g}t + \beta_{2g}t^2\,, \qquad g = 1, 2,$$

so that at the start of the experiment $f_1(0) = f_2(0) = \beta_0$.

The model was estimated using one million iterations for both burn-in and inference and took 15 hours on the SGI computer. This number of iterations gave agreement of the estimates to three significant figures when $\Omega$ was initialised from different starting positions. We checked that, for this dataset, using 10,000 iterations gave agreement to two significant digits when the sampler was initialised from different positions. The acceptance rate of the Metropolis-Hastings step was 0.99. Figures 5 and 6 summarise the estimation results. Figure 5 gives 'image' plots of the posterior mean estimates of the $J$ matrix, and the absolute value of each element of $\widehat{\Sigma}^{-1}_{ESL_1}$ and $\widehat{\Sigma}^{-1}_{ML}$. The figure shows that the estimated $J$ and $\Omega$ matrices appear banded, with two bands on either side of the diagonal, whereas $\widehat{\Sigma}^{-1}_{ML}$ appears fuller. The estimated structure of $\Omega$ is consistent with the results in Kenward (1987), and is called by Kenward a second-order antedependence model. Such a second-order antedependence model can be viewed as a generalisation of a second-order autoregressive process and has a natural interpretation for longitudinal data.

———————— Figure 5 about here ————————-

Similarly to the Capital Asset Pricing Model data, we studied the properties of the iterates of the loglikelihood and selected elements of the $\Omega$ matrix. The iterate plots and the corresponding autocorrelation plots suggested that for this example the sampling scheme converged and mixed well.

Using the SGI Origin 2000 computer it took 8 hours to estimate the model using iterations each for burn-in and sampling.

Figure 6 plots the estimates of the posterior means of the two trend functions together with $\pm 2$ posterior standard deviations bands about the first trend function. The figure also plots the estimate of the difference in the trend functions together with $\pm 2$ posterior standard deviations bands about the estimated difference. The figure suggests that the two trend functions are close to each and that there is no difference in the diet effects.

———————— Figure 6 about here ————————-

### 5·3 *First-order Markov random field*

This section reports on covariance selection for data generated by a $9 \times 9$ first-order Markov random field model. Each internal point has four neighbours, consisting of adjacent horizontal and vertical neighbours. Each corner point has two neighbours, and each edge point that is not a corner point has three neighbours. The resulting concentration matrix $\Omega$ is $81 \times 81$ with diagonal elements equal to 1 and the nonzero off-diagonal elements equal to 0.2. Our reason for studying the first-order Markov random field model is that it is an example of a highly sparse non-decomposable graphical model, and that it is straightforward to check that the closest decomposable models have substantially more nonzero elements, although they are still highly sparse. A similar example was considered by Giudici & Green (1999), who looked at a $5 \times 5$ first-order Markov random field.

One hundred zero-mean observations were generated from the Markov random field, and covariance selection was carried out using 10,000 iterations for both burn-in and inference. With this number of iterations the estimates of $\Omega$ agreed to three significant digits when $\Omega$ was initialised from different positions. The acceptance rate of the Metropolis-Hastings step was 0.99 and the computer time on the SGI machine was 8 hours. Figure 7 summarises the estimation results and gives 'image' plots of the posterior mean estimates of the $J$ matrix, and the absolute value of each element of $\widehat{\Sigma}^{-1}_{ESL_1}$ and

$\widehat{\Sigma}_{ML}^{-1}$. The figure shows that $\widehat{\Sigma}_{ML}^{-1}$ seems full, whereas the image plots of the estimates of the $J$ and $\Omega$ matrices reproduce the Markov random field structure.

———————— Figure 7 about here ————————

Similarly to the analysis of the other two datasets, we studied the properties of the iterates of the loglikelihood and selected elements of the $\Omega$ matrix. The iterate plots and the corresponding autocorrelation plots suggested that for this example the sampling scheme converged and mixed well.

ACKNOWLEDGEMENT

We would like to thank the editor, an associate editor and two referees whose suggestions improved the quality of the paper.

## Appendix 1

*Estimating normalising constants*

This appendix describes our method of estimating the average volumes $\overline{V}(l)$ of the positive definite regions of $C$ that have $l$ nonzero entries. Further details can be obtained from the authors. To estimate the average volumes, we work with a more general prior specification than (2·8), because it enables us to check the accuracy of the volume approximations.

*A more general prior for $C$.* We consider the following prior specification for $C$, which reduces to the prior (2·5)-(2·7) if $k(\psi) = 1$ and $\Psi(l) = \overline{V}(l)$. In the notation in § 2,

$$\begin{aligned} p(dC|J) &= V(J)^{-1} dC_{\{J=1\}} I(C \in \mathcal{C}_p)\,, \\ p\{J|S(J)=l\} &= \binom{r}{l}^{-1} \frac{V(J)}{\overline{V}(l)}\,, \\ p\{S(J)=l|\psi\} &= \frac{k(\psi)\overline{V}(l)}{\Psi(l)} \binom{r}{l} \psi^l (1-\psi)^{r-l}\,. \end{aligned} \qquad \text{(A1·1)}$$

The following lemma is used later to check the quality of the approximations to the average volumes $\overline{V}(l)$.

**Lemma 6** .

$$\Psi(l) = \overline{V}(l) \text{ for } l = 0, \ldots, r \Longleftrightarrow k(\psi) \propto 1.$$

*Proof.* Summing (A1·1) over $l$ gives

$$k(\psi)^{-1} = \sum_{l=0}^{r} \frac{\overline{V}(l)}{\Psi(l)} \binom{r}{l} (1-\psi)^{r-l}\psi^l$$

and the result follows because the polynomials $(1-\psi)^{r-l}\psi^l$, for $l = 0, \ldots, r$, are linearly independent.

We estimate the ratio $k(\psi')/k(\psi)$ of two neighbouring points $\psi'$ and $\psi$ using importance sampling. Integrating $p(dC|\psi)$ over the positive definite region for $C$ gives

$$\frac{1}{k(\psi)} = \int_{C\in\mathcal{C}_p} \frac{1}{\Psi(S)} p^*(dC|\psi) \ ,$$

so that

$$\begin{aligned} \frac{k(\psi')}{k(\psi)} &= \int_{C\in\mathcal{C}_p} \frac{p^*(dC|\psi)}{p^*(dC|\psi')} \left[\frac{k(\psi')}{\Psi\{S(J(C))\}} p^*(dC|\psi')\right] \\ &= \int_{C\in\mathcal{C}_p} \left(\frac{1-\psi}{1-\psi'}\right)^{r-S\{J(C)\}} \left(\frac{\psi}{\psi'}\right)^{S\{J(C)\}} p(dC|\psi') \,, \end{aligned}$$

which can be estimated using importance sampling by generating a sample from $p(dC|\psi')$ and averaging the terms

$$\left(\frac{1-\psi}{1-\psi'}\right)^{r-S\{J(C)\}} \left(\frac{\psi}{\psi'}\right)^{S\{J(C)\}}$$

over the sample values of $C$.

*Estimating the average volume* $\overline{V}(l)$. We now describe how to calculate $V(J^*)$ for block-diagonal partial correlation matrices. We denote the multivariate gamma function by

$$\Gamma_p(\alpha) = \int_{A>0} (\det A)^{\alpha-(p+1)/2} \exp\{-\operatorname{tr}(A)\}\, dA \qquad \text{(A1·2)}$$

$$= \pi^{p(p-1)/4} \prod_{i=1}^{p} \Gamma\{\alpha - (i-1)/2\} \ , \qquad \text{(A1·3)}$$

where $A$ is a $p \times p$ matrix and the condition $A > 0$ means that $A$ is positive definite, and hence symmetric. Equation (A1·3) follows from Theorem 2.1.12 in Muirhead (1982).

We denote the volume of the positive definite region $\mathcal{C}_p$ for $p \times p$ unconstrained correlation matrices by $V(\mathcal{C}_p)$. The following lemma gives a formula for $V(\mathcal{C}_p)$.

**Lemma 7** .

$$V(\mathcal{C}_p) = \Gamma_p\left(\frac{p+1}{2}\right) / \Gamma\left(\frac{p+1}{2}\right)^p \ . \qquad \text{(A1·4)}$$

*Proof.* Change variables in (A1·2) using the decomposition

$$A = TCT\,,$$

where $T$ is a diagonal matrix and $C$ is a correlation matrix. Barnard et al. (2000) show that

$$dA = 2^p \left( \prod_{i=1}^{p} T_{ii} \right)^p \; dT \; dC,$$

so that

$$\begin{aligned} \Gamma_p \left( \frac{p+1}{2} \right) &= \int_{T>0, C \in \mathcal{C}_p} 2^p \left( \Pi_{i=1}^p T_{ii} \right)^p \exp\left\{ -\operatorname{tr}(T^2) \right\} \; dT \; dC \\ &= V(\mathcal{C}_p) \left\{ 2 \int_0^\infty t^p \exp(-t^2) dt \right\}^p \\ &= V(\mathcal{C}_p) \Gamma \left( \frac{p+1}{2} \right)^p , \end{aligned}$$

as required.

Suppose that $J^*$ specifies the constraints that $C$ is block diagonal and let

$$C = \operatorname{diag}(C_{k_1}, \ldots, C_{k_m}) \;,$$

where $p = \sum_{i=1}^m k_i$ and the submatrices $C_{k_1}$, ..., $C_{k_m}$ are unconstrained correlation matrices. Then

$$V(J^*) = \prod_{i=1}^{m} V(\mathcal{C}_{k_i}) \,.$$

Our approach to estimating $\overline{V}(l)$ is to use (A1·4) to evaluate $V(J)$ for block-diagonal matrices with $S(J) = l$. The number of possible block-diagonal matrices grows exponentially with $p$, so it is not possible to consider all block-diagonal matrices for $p$ larger than about 40. In this case, we draw a random sample of block-diagonal matrices instead. We then use nonparametric regression to estimate $\overline{V}(l)$ as a function of $l$. We denote this estimator by $\overline{V}_{\mathrm{reg}}(l)$.

We can also calculate a regression curve for $\overline{V}(l)$ by interpolating $V(\mathcal{C}_i)$ for $i = 1, \ldots, p$. We denote this estimator by $\overline{V}_{\mathrm{LB}}(l)$, because it is likely to be close to the lower bound of $\overline{V}(l)$.

From Lemma 6, $\Psi(l) = \overline{V}(l)$ if and only if $k(\psi) \propto 1$, so we can check the accuracy of our results by estimating $k(\psi)$ using the importance sampling method described earler in this appendix. We found that a $0.6 : 0.4$ interpolation of $\overline{V}_{\mathrm{reg}}(l)$ and $\overline{V}_{\mathrm{LB}}(l)$ gave an approximately constant value of $k(\psi)$ and we used this as the value of $\overline{V}(l)$ in (2·8) for the computations in §§4 and 5.

## Appendix 2

*Proofs of lemmas*

*Proof of Lemma* 1. We note that up to normalising constants the density $T_i^{n+2\alpha-1}\exp\left(-\frac{1}{2}T_i\right)$ is close to a chi-squared distribution with $2n$ degrees of freedom, and hence, for large $n$, is close to normal. The density $\exp\left(-aT_i^2 - bT_i + \frac{1}{2}T_i\right)$ is Gaussian and hence the conditional density of $T_i$ is Gaussian for $n$ large because it is a product of Gaussians.

*Proof of Lemma 2.* Part (i) is straightforward. To show part (ii), note that with $i = p$ and $j = p-1$,

$$\det C = \prod_{k=1}^{p} R_{kk}^2 \propto R_{pp}^2 \ ,$$

because $R_{kk}$ does not depend on $C_{p-1,p}$ for $k < p$. Now,

$$C_{pp} = \sum_{j=1}^{p-2} R_{j,p}^2 + R_{p-1,p}^2 + R_{p,p}^2 \ ,$$

$$C_{p-1,p} = \sum_{j=1}^{p-2} R_{j,p-1}R_{j,p} + R_{p-1,p-1}R_{p-1,p}.$$

Hence

$$R_{p-1,p} = \frac{C_{p-1,p} - a}{b} \quad \text{and} \quad R_{pp}^2 = c - \left(\frac{C_{p-1,p} - a}{b}\right)^2 .$$

Part (iii) is straightforward.

*Proof of Lemma 3* . From (2·8) and Lemma 2 part (ii)

$$\begin{aligned} p(dC_{ij}|C_{\{-ij\}}) &= p(dC_{ij}|dC_{\{-ij\}}) \\ &\propto p(dC) \\ &= \int_0^1 p(dC|\psi)p(\psi)d\psi \\ &\propto \int_0^1 I\left(|C_{ij} - a| < b\sqrt{c}\right)\left\{I\left(0 \in dC_{ij}\right)(1-\psi) + dC_{ij}\psi\right\} \\ &\times (1-\psi)^{r-1-S\left(J_{\{-ij\}}\right)}\psi^{S\left(J_{\{-ij\}}\right)}d\psi. \end{aligned}$$

The result follows after normalising the last line.

*Proof of Lemma 4.* The result follows from Lemma 2 part (iii) and Lemma 3.

*Proof of Lemma 5.* Part (i) follows from Lemma 4 part (iii). Parts (ii) and (iii) are straightforward. Part (iv) follows from part (i) and the fact that $(1 - v^2/n)^{n/2} \to \exp(-v^2/2)$ as $n \to \infty$.

## Appendix 3

*Fast generation of $\beta$*

We now show how to evaluate efficiently the $mp \times mp$ matrix $A = X'(I \otimes \Omega)X$ which is required in each iteration when generating $\beta$ as in § 3. The vector $X'(I \otimes \Omega)Y$ is evaluated similarly.

It is straightforward to check that

$$\begin{aligned} A(i,j) &= \sum_{t=1}^{n}\sum_{s=1}^{p}\sum_{r=1}^{p} X_t(i,r)'\Omega(r,s)X_t(s,j) \\ &= \sum_{t=1}^{n}\sum_{s=1}^{p}\sum_{r=1}^{p} X_t(r,i)X_t(s,j)\Omega(r,s) \\ &= \sum_{s=1}^{p}\sum_{r=1}^{p}\left[\sum_{t=1}^{n} X\{(t-1)p+r,i\}X\{(t-1)p+s,j\}\right]\Omega(r,s)\ . \end{aligned}$$

Let $G(i,j,r,s) = \sum_{t=1}^{n} X\{(t-1)p+r,i\}X\{(t-1)p+s,j\}$. We pre-compute the four-dimensional array $G$ of size $m \times m \times p \times p$ once prior to the Markov chain Monte Carlo simulation. Then, in each iteration of the simulation, we compute $A$ using $A(i,j) = \sum_{s=1}^{p}\sum_{r=1}^{p} G(i,j,r,s)\Omega(r,s)$. Thus, the computation required to generate $\beta$ is independent of the sample size $n$.

## References


Barnard, J., McCulloch, R., & Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10**, 1281–311.

Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The Econometrics of Financial Markets.* Princeton University Press, New Jersey.

Chiu, T., Leonard, T., & Tsui, K. (1996). The matrix-logarithm covariance model. *J. Am. Statist. Assoc.* **81**, 310–20.

Daniels, M. & Kass, R. (1999). Nonconjugate Bayesian estimation of covariance matrices. *J. Am. Statist. Assoc.* **94**, 1254–63.

Dellaportas, P., Giudici, P., & Roberts, G. (2001). Bayesian inference for nondecomposable graphical gaussian models. *Sankhya, Series A* To appear.

Dempster, A. P. (1969). *Elements of Continuous Multivariate Analysis.* Addison-Wesley, Reading, MA.

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

Diggle, P., Liang, K., & Zeger, S. (1994). *Analysis of Longitudinal Data.* Clarendon Press, Oxford.

Efron, B. & Morris, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann. Statist.* **4**, 22–32.

Gibbons, M. (1982). Multivariate tests of financial models: A new approach. *J. Finan. Econ.* **14**, 217–36.

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

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **84**, 711–32.

Haff, L. R. (1991). The variational form of certain Bayes estimators. *Ann. Statist.* **19**, 1163–90.

Kenward, P. J. (1987). A method of comparing profiles of repeated measurements. *Appl. Statist.* **36**, 296–308.

Lauritzen, S. (1996). *Graphical Models.* Oxford University Press, Oxford.

Leonard, T. & Hsu, J. S. J. (1992). Bayesian inference for a covariance matrix. *Ann. Statist.* **20**, 1669–96.

Lindsey, J. K. (1993). *Modelling Frequency and Count Data.* Clarendon Press, Oxford.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, London.

Muirhead, R. (1982). *Aspects of Multivariate Statistical Theory.* Wiley, New York.

Pourahmadi, M. (1999). Joint mean-covariance models with application to logitudinal data: Unconstrained parameterisation. *Biometrika* **86**, 677–90.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87**, 425–35.

Roverato, A. (2002). Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scand. J. Statist.* **29**, 391–411.

Sharpe, W. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *J. Finan.* **19**, 425–42.

Smith, M. & Kohn, R. (2002). Bayesian parsimonious covariance matrix estimation for longitudinal data. *J. Am. Statist. Assoc.* **87**, 1141–53.

Whittaker, J. (1990). *Graphical Models in Applied Mathematical Analysis.* Wiley, New York.

Yang, R. & Berger, J. O. (1994). Estimation of a covariance catrix using the reference prior. *Ann. Statist.* **22**, 1195–211.

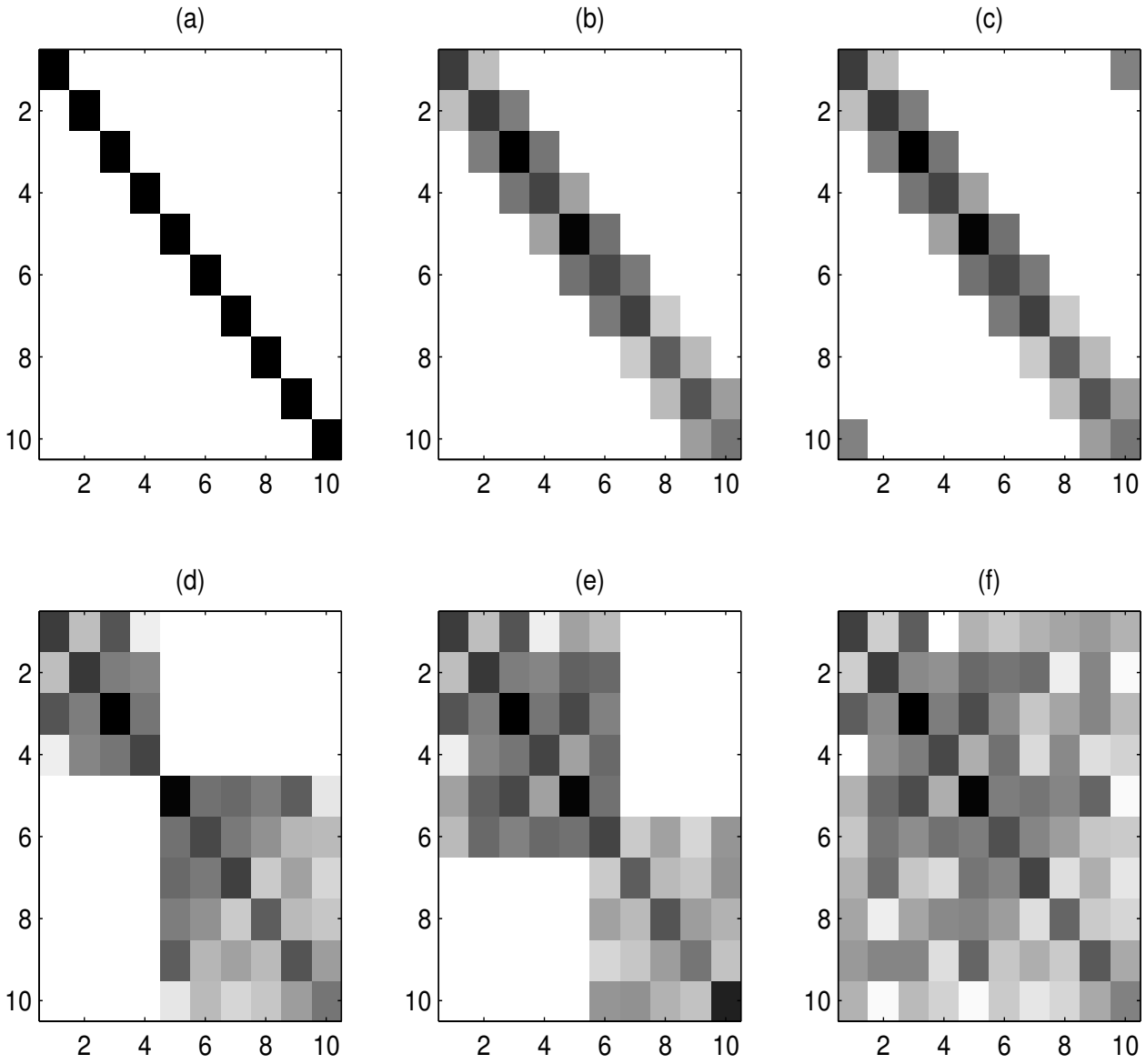


Figure 1: Image plots of the six inverse covariance matrices used in the simulation. Panels (a)-(f) represent the identity matrix, the tridiagonal matrix, the loop matrix, the block-diagonal matrix, the overlapping blocks matrix and the full matrix.

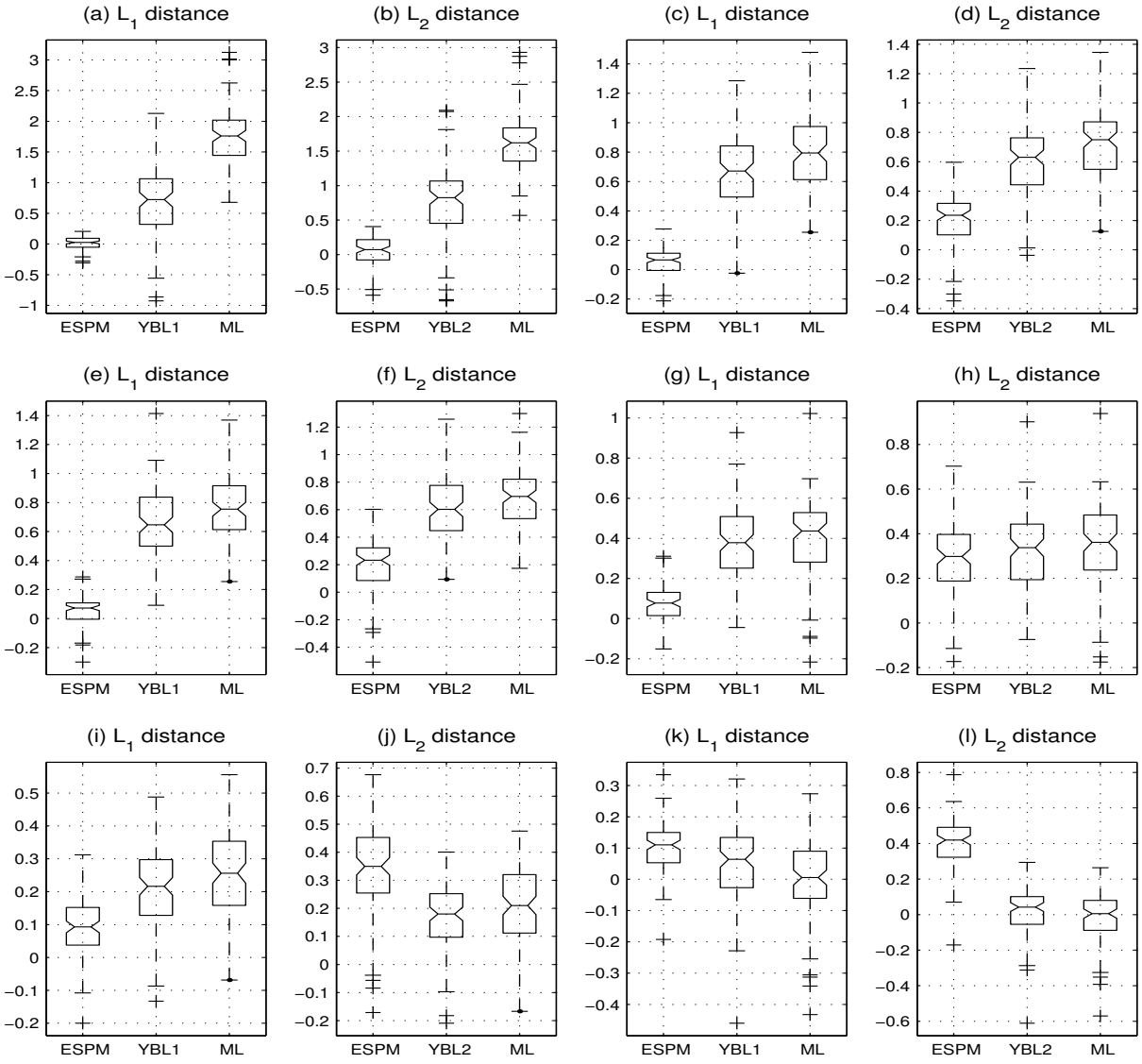


Figure 2: Comparison of the Bayes estimator of $\Sigma$ for the $L_1$ (or $L_2$) loss function based on the covariance selection prior $\widehat{\Sigma}_{ESL_1}$ (or $\widehat{\Sigma}_{ESL_2}$), with the Bayes estimator based on the Yang and Berger prior $\widehat{\Sigma}_{YBL_1}$ (or $\widehat{\Sigma}_{YBL_2}$) and the maximum likelihood estimator $\widehat{\Sigma}_{ML}$. Panels (a) and (b), $\Sigma^{-1}$ is the identity matrix; panels (c) and (d), $\Sigma^{-1}$ is tridiagonal; panels (e) and (f), $\Sigma^{-1}$ is a loop matrix; panels (g) and (h), $\Sigma^{-1}$ is a block-diagonal matrix; panels (i) and (j), $\Sigma^{-1}$ is an overlapping blocks matrix; panels (k) and (l), $\Sigma^{-1}$ is a full matrix.

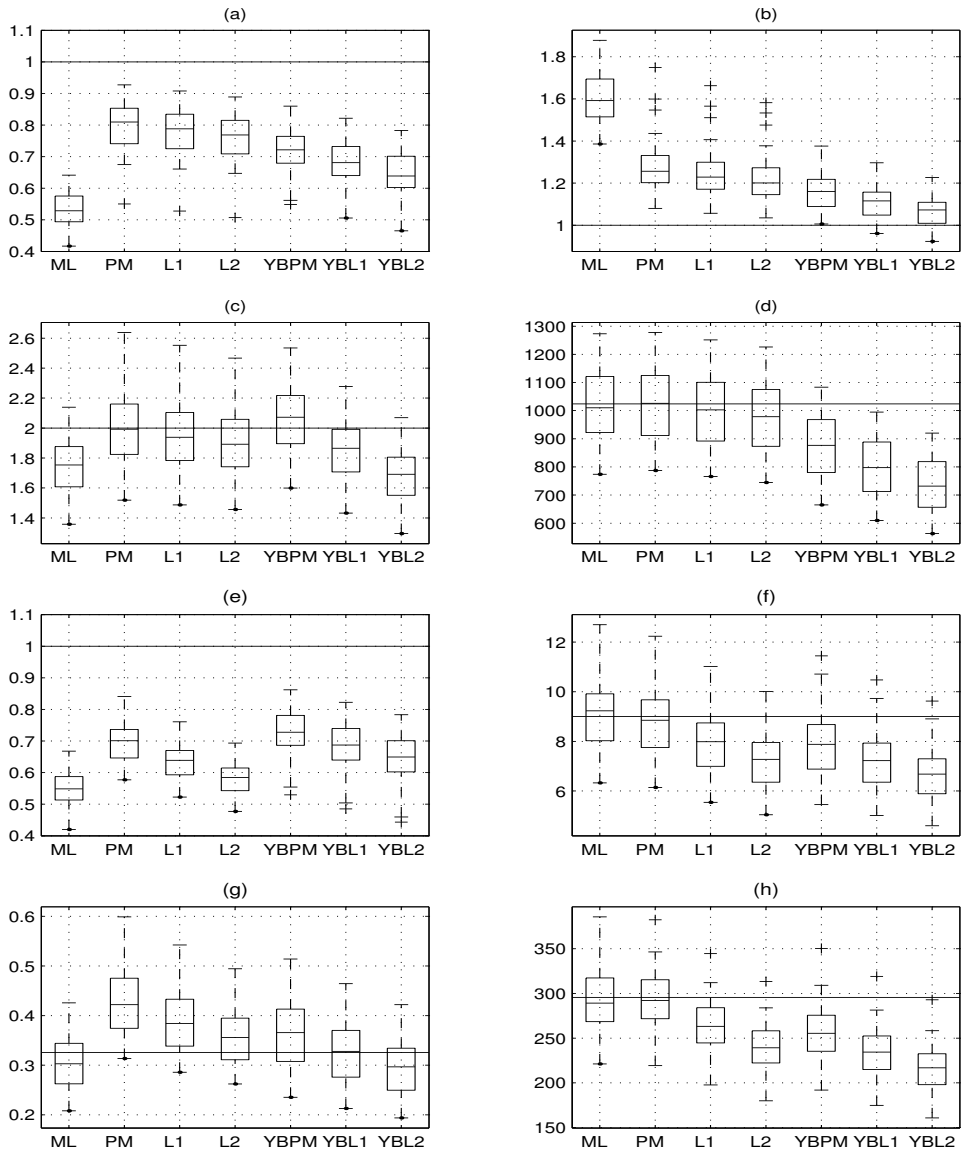


Figure 3: Boxplots of the smallest (left panel) and largest eigenvalue (right panel) estimators. Panels (a) and (b): identity matrix; (c) and (d) ill-conditioned diagonal matrix; (e) and (f) full matrix with nine diagonal elements equal to one and one equal to nine; (g) and (h) full matrix. The solid horizontal lines indicate the true values of the eigenvalues. ML: maximum likelihood estimator. PM, L1 and L2 are the posterior mean, and L1 and L2 Bayes estimators using the covariance selection prior. YBPM, YBL1 and YBL2 are the corresponding estimators for the Yang and Berger prior.

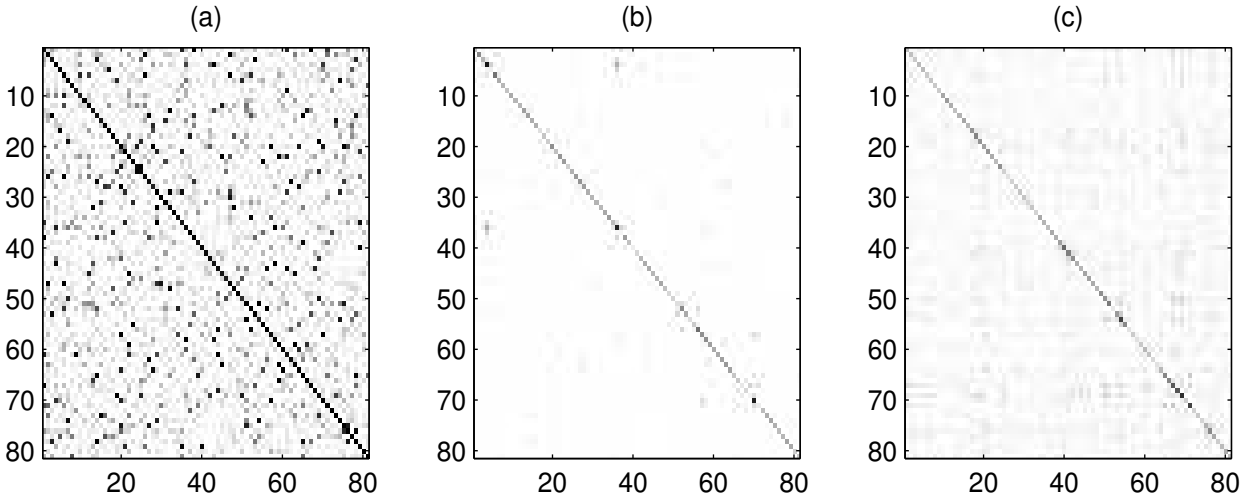


Figure 4: Capital Asset Pricing Model: (a) Shaded 'image' of the estimated probability of $J_{ij} = 1$, from 0 (lightest) to 1 (darkest), showing the sparsity of $\Omega$. Panels (b) and (c) are the shaded 'images' of $\widehat{\Sigma}^{-1}_{ESL_1}$ and $\widehat{\Sigma}^{-1}_{ML}$ respectively. Absolute values of individual elements are considered, with lightest corresponding to zero.

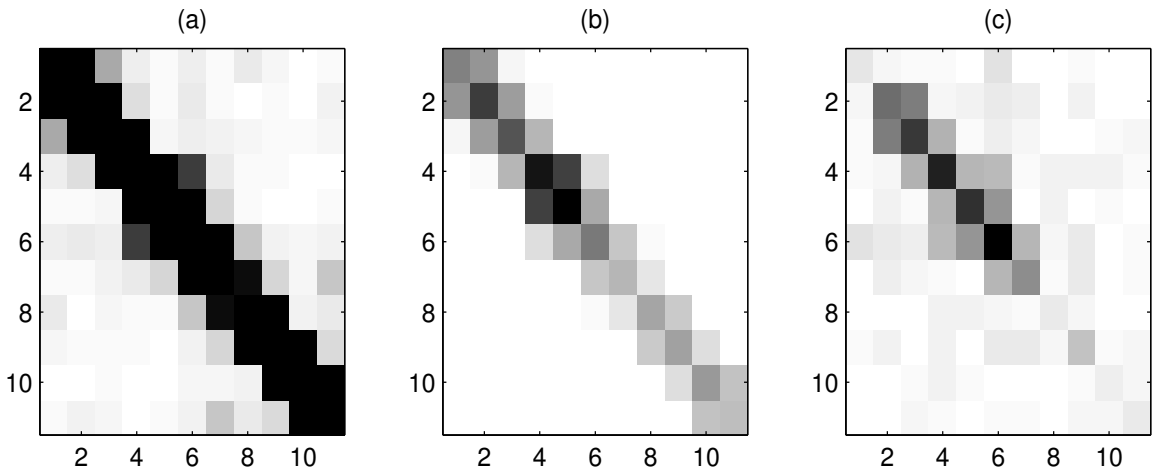


Figure 5: Diet experiment: (a) Shaded 'image' of the estimated probability of $J_{ij} = 1$, from 0 (lightest) to 1 (darkest), showing the sparsity of $\Omega$. Panels (b) and (c) are the shaded 'images' of $\widehat{\Sigma}^{-1}_{ESL_1}$ and $\widehat{\Sigma}^{-1}_{ML}$ respectively. Absolute values of individual elements are considered, with lightest corresponding to zero.

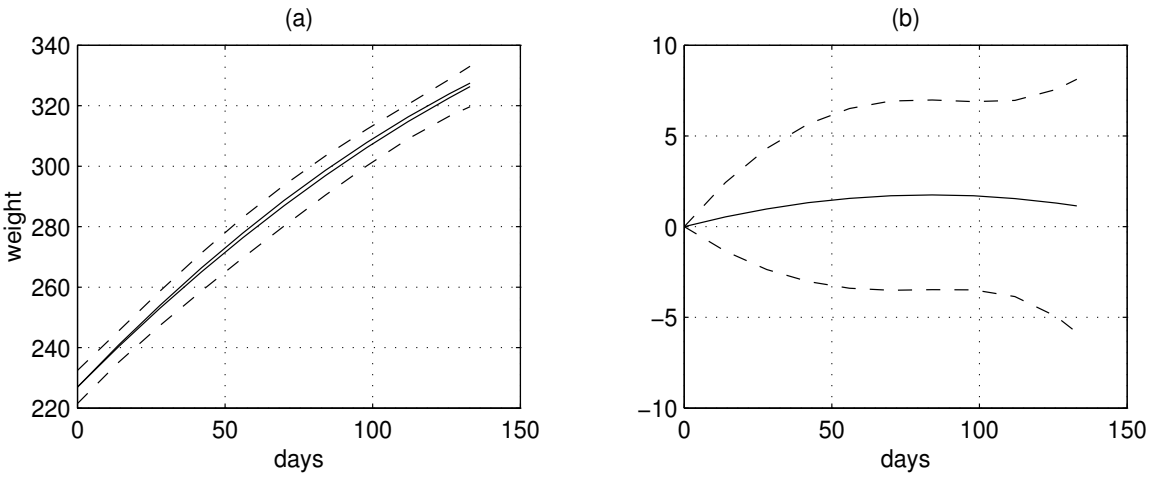


Figure 6: Diet experiment: Panel (a) plots the estimated trend functions together with $\pm 2$ posterior standard deviation bands for the first trend function. Panel (b) plots the estimated difference between the two trend functions as well as the $\pm 2$ standard deviations bands of the difference.

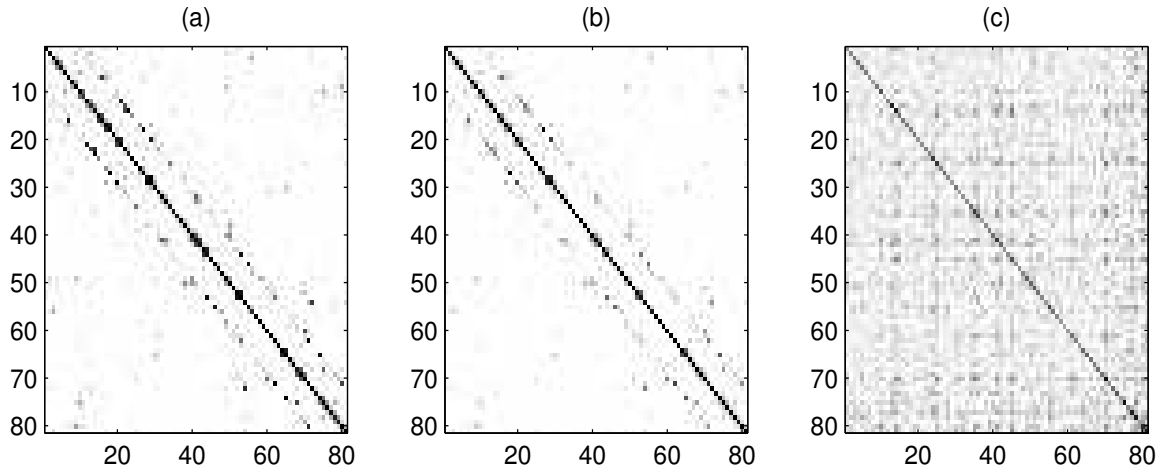


Figure 7: First-order Markov random field: (a) Shaded 'image' of the estimated probability of $J_{ij} = 1$, from 0 (lightest) to 1 (darkest), showing the sparsity of $\Omega$. Panels (b) and (c) are the shaded 'images' of $\widehat{\Sigma}^{-1}_{ESL_1}$ and $\widehat{\Sigma}^{-1}_{ML}$ respectively. Absolute values of individual elements are considered, with lightest corresponding to zero.