Parsimonious Covariance Matrix Estimation for Longitudinal Data

Michael SMITH and Robert KOHN

This article proposes a data-driven method to identify parsimony in the covariance matrix of longitudinal data and to exploit any such parsimony to produce a statistically efficient estimator of the covariance matrix. The approach parameterizes the covariance matrix through the Cholesky decomposition of its inverse. For longitudinal data, this is a one-step-ahead predictive representation, and the Cholesky factor is likely to have off-diagonal elements that are zero or close to zero. A hierarchical Bayesian model is used to identify any such zeros in the Cholesky factor, similar to approaches that have been successful in Bayesian variable selection. The model is estimated using a Markov chain Monte Carlo sampling scheme that is computationally efficient and can be applied to covariance matrices of high dimension. It is demonstrated through simulations that the proposed method compares favorably in terms of statistical efficiency with a highly regarded competing approach. The estimator is applied to three real examples in which the dimension of the covariance matrix is large relative to the sample size. The first two examples are from biometry and electricity demand modeling and are longitudinal. The third example is from finance and highlights the potential of our method for estimating cross-sectional covariance matrices.

KEY WORDS: Bayesian model; Cholesky decomposition; High-dimensional covariance matrix; Markov chain Monte Carlo; Model averaging.

1. INTRODUCTION

This article develops a model that allows a parsimonious, but flexible representation of the covariance matrix of multivariate Gaussian longitudinal data. The model is used to efficiently estimate the covariance matrix, even when the dimension is large relative to the sample size. This is an important issue because the number of free elements in a covariance matrix increases quadratically with its dimension, and the unrestricted maximum likelihood estimator of the covariance matrix has long been known to be statistically inefficient (Stein 1956). When the covariance matrix is that of the errors in a regression model, this inefficiency also may lead to inefficient estimators of the regression coefficients.

We parameterize the longitudinal covariance matrix using the Cholesky factorization (Golub and van Loan 1996, p. 142) of its inverse. That is, if $e \sim N(0, \Sigma)$, then our approach uses the decomposition

$$\Sigma^{-1} = BDB', \tag{1}$$

where B is a lower-triangular matrix with 1s on the diagonal and D is a diagonal matrix. The lower-triangular elements of the factor B can be interpreted as the coefficients of the one-step-ahead predictive equation $B'e = \zeta$, for e in reverse time order, where $\zeta \sim N(0, D^{-1})$. It is therefore likely that many of the strict lower-triangular elements of B are zero or close to zero. For example, if e is generated by a zero-mean autoregressive model, then B is a lower-triangular band matrix.

To flexibly identify zeros in the lower triangle of B, we place a hierarchical prior on B similar to that used in the Bayesian variable selection literature (see George and McCulloch 1997 for a discussion of this literature). The hierarchical structure has the advantage in that not only can potential zeros in the Cholesky factor B be identified, but also estimates of the parameters can be calculated that account for the "model uncertainty" associated with determining which elements in the lower triangle of B are zero. The entire Bayesian model is estimated using a Markov chain Monte Carlo (MCMC) sampling scheme, which is shown to be practical to implement even when the covariance matrix is of high dimension. Parameter inference is obtained by taking a weighted average over the different models for B, with the weights being the posterior model probabilities (see, e.g., Raftery, Madigan, and Hoeting 1997).

Our work is related to that on Gaussian covariance selection models, also called Gaussian graphical models, that studies covariance matrices with patterns of zeros in the inverse (see Dempster 1972; Whittaker 1990). However, parsimony in the Cholesky factor B does not imply equivalent parsimony in the graphical model, and vice versa. For example, if the strict lower triangle of B is zero except for its first and fifth subdiagonals, then the corresponding graphical model will not be as parsimonious because its second, third, and fourth subdiagonals will also have some non-zero elements. Giudici and Green (1999) provided a Bayesian framework and a sampling scheme for estimating decomposable graphical models. Their method is restricted to decomposable models, and the intricacy of their sampling approach suggests that it is restricted to covariance matrices of reasonably small dimension. Wermuth (1980) showed that if Σ corresponds to a decomposable graphical model, then for a perfect ordering of the elements of e, the lower triangle of B has the same pattern of zeros as Σ^{-1} (see also Roverato 2000). But the shortcomings of this result are that the graphical model needs to be decomposable and a perfect ordering is unknown if Σ is unknown. Nevertheless, the results of Wermuth (1980) suggest that some cross-sectional

Michael Smith is Associate Professor, Econometrics and Business Statistics, Faculty of Economics and Business, University of Sydney, NSW 2006, Australia (E-mail: *mikes@econ.usyd.edu.au*). Robert Kohn is Professsor, Australian Graduate School of Management, University of New South Wales, Sydney, NSW 2052, Australia (E-mail: *robertk@agsm.unsw.edu.au*). This work was partially supported by grants from the Australian Research Council. The authors thank an anonymous referee and the associate editor for helping clarify the results in the article. They also thank Chris Carter, Denzil Fiebig, and Tom Smith for useful comments; Garry Twite for help in obtaining the data for the CAPM example; and staff at Pacific Power for supplying the electricity demand and temperature data.

Journal of the American Statistical Association
December 2002, Vol. 97, No. 460, Theory and Methods
DOI 10.1198/016214502388618942

covariance matrices can also be represented by a parsimonious Cholesky factor B in the decomposition (1).

A number of authors have considered the problem of efficiently estimating a covariance matrix for both longitudinal data and more general covariance matrices (see Pinheiro and Bates 1996 for a partial summary). The work of Pourhamadi (1999, 2000) is closest to ours. He used the modified Cholesky decomposition $\Sigma^{-1} = B'DB$, with B and D as defined in (1), and modeled the strict lower-triangular elements of B as linear functions of explanatory variables. The approach of Pourhamadi (1999, 2000) subsumes many of the models used for a covariance matrix in longitudinal data, including autoregressive and antedependence models (see, e.g., Gabriel 1962; Macchiavelli and Arnold 1994). Pourhamadi (1999, 2000) stressed interpretation of the Cholesky decomposition for longitudinal data, something also stressed here. However, Pourhamadi's approach differs from ours because it is parametric and does not attempt to formally identify any structural zeros. He also estimated his model by maximum likelihood and did not average over models. In contrast, the Bayesian approach provides a framework for averaging over parsimonious configurations of B, and inference is based on the finite-sample posterior distribution.

An alternative factorization of the covariance matrix used by a number of authors is the spectral decomposition, $\Sigma = O'\Lambda O$, where O is the orthonormal matrix of eigenvectors and Λ is the diagonal matrix of corresponding eigenvalues. The matrix O is further decomposed into a product of Givens rotation matrices, so that Σ is parameterized in terms of its eigenvalues and Givens angles. (see, e.g., Yang and Berger 1994; Pinherio and Bates 1996; Daniels and Kass 1999; and references therein). Yang and Berger (1994) placed a reference prior on the Givens angles and eigenvalues and used it to carry out Bayesian inference on Σ. They showed empirically that the performance of the resulting Bayes estimators of Σ and its inverse under different loss functions compares favorably with several alternative estimators of Σ. The approach of Yang and Berger (1994) is flexible in that it does not assume any specific parametric form for the covariance matrix, and is more general than our approach because it applies equally to covariance matrices arising from cross-sectional data as well as those from longitudinal data. However, it does not identify any specific parsimonious structure in the covariance matrix or its inverse, which is an objective of our approach for covariance matrices arising in longitudinal data.

We demonstrate by simulation that for longitudinal data, our proposed method can efficiently estimate the covariance matrix and its inverse when the Cholesky factor is sparse or has elements close to zero. The performance is shown to be competitive with the estimator suggested by Yang and Berger (1994), which is recognized as one of the most statistically efficient covariance matrix estimators in the current literature.

The practical value of our estimator is demonstrated by its application to three real examples in which the dimension of the covariance matrix is relatively high. The first example is an analysis from a longitudinal repeated-measures experiment on the live weight of a sample of cows from Diggle, Liang, and Zeger (1994, p. 100), where measurements are made on each cow at 23 unequally spaced time points. The second example uses an econometric model for the half-hourly demand for electricity. The 48 intradaily observations are assumed to be correlated within a day, whereas, given temperature as an independent variable, each day is considered an independent repeated measure. The last example considers a multivariate capital asset pricing model, in which the dependent variable vector consists of returns on the 89 stocks on the Standard and Poor 100 index that traded continuously over the period November 1986–November 1996. The first two examples are longitudinal. The third example is cross-sectional, although the Cholesky factor B and the inverse covariance matrix Σ^{-1} are still found to be parsimonious, suggesting the applicability of our method to some cross-sectional covariance structures as motivated by Wermuth (1980).

The article is organized as follows. Section 2 presents the Bayesian hierarchical model. Section 3 discusses Bayesian inference and presents the sampling scheme used to generate samples from the posterior distribution. Section 4 presents the results of a simulation study that compares our method with that of Yang and Berger (1994) and the unrestricted maximum likelihood estimator, and Section 5 applies the methodology to the three real examples. Section 6 summarizes the results of the article. An Appendix presents the details of the Markov chain Monte Carlo sampling scheme.

2. THE MODEL AND PRIOR

2.1 Parsimonious Covariance Matrix Decomposition

Suppose that $e_1, \ldots, e_n$ are n vectors of dimension m distributed independently $e_i \sim N(0, \Sigma)$. This section and the next consider modeling and estimating the $m \times m$ covariance matrix Σ. To simplify the exposition, the zero-mean case is considered, with the non-zero-mean case considered in the real examples in Section 5.

Here we are concerned primarily with the case where the number, $m(m+1)/2$, of unknown parameters in Σ is reasonably large compared to the total number of scalar observations nm. In this case it is hard to obtain reliable estimates of Σ without imposing some restrictions on its form. However, in many examples it is difficult to determine these restrictions a priori, and thus a data-driven method that allows for such restrictions is useful.

We factor the inverse of the covariance matrix $\Sigma^{-1} = \{\sigma^{i,j}\}$ into a full-rank $m \times m$ matrix $B = \{b_{i,j}\}$ and a diagonal matrix $D = \text{diag}(d_1, \ldots, d_m)$ as in (1). To allow parsimony in the representation, each of the lower triangular elements of B can be exactly zero with positive probability. This is achieved by introducing binary indicator variables $\gamma_{i,j}$, so that

$$b_{i,j} = 0 \quad \text{iff} \quad \gamma_{i,j} = 0$$

and

$$b_{i,j} \neq 0 \quad \text{iff} \quad \gamma_{i,j} = 1$$

for the elements $j = 1, \ldots, m-1, \quad i > j$. Therefore, the form of the Cholesky factor B is known only conditional on the "model parameter" $\gamma = \{\gamma_{i,j} | j = 1, \ldots, m-1; i > j\}$.

We note that many longitudinal covariance matrices are parameterized parsimoniously in terms of the Cholesky factor B. For example, if Σ is the covariance of an autoregressive process, then B is a band matrix.

2.2 Likelihood

When the mean of the regression is zero, the likelihood of the parameters (B, D, γ) is the density of $e = (e_1', \ldots, e_n')'$ given (B, D, γ), that is,

$$\begin{aligned} p(e|B, D, \gamma) &= (2\pi)^{-nm/2}|D|^{n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n} e_i' BDB' e_i\right\} \\ &= (2\pi)^{-nm/2}\prod_{i=1}^{m}(d_i)^{n/2}\exp\left\{-\frac{1}{2}\sum_{k=1}^{m} d_k b_k' A b_k\right\}, \quad (2) \end{aligned}$$

where $b_k = (b_{1,k}, \ldots, b_{m,k})'$ is the kth column of B and $A = \sum_{i=1}^{n} e_i e_i'$. The matrix A is positive definite almost surely if $m \le n$.

The upper-triangular elements of B are equal to zero, whereas the diagonal elements are equal to one. Moreover, given a particular model γ, some lower-triangular elements of B are also exactly equal to zero. If the vector of the remaining unconstrained elements of b_k is denoted as $\beta_k = \{b_{i,k} | i > k, \gamma_{i,k} = 1\}$, then the quadratic form in the likelihood in (2) can be written in terms of a quadratic function of β_k, so that

$$b_k' A b_k = \begin{cases} a_{k,k} + 2\beta_k' a_k + \beta_k' A_k \beta_k & \text{for } k = 1, \ldots, m-1 \\ a_{m,m} & \text{for } k = m. \end{cases}$$

Here the vector $a_k = \{a_{i,k} | i > k, \gamma_{i,k} = 1\}$ and the matrix $A_k = \{a_{i,j} | i > k, j > k, \gamma_{i,j} = 1\}$. We denote the number of unconstrained elements in the kth column (i.e., the dimension of β_k) as q_k. The total number of unconstrained elements in B corresponding to model γ is denoted as $q_\gamma = \sum_{k=1}^{m-1} q_k$. Note that A_k is strictly positive definite if $0 < q_k < n$.

Thus the likelihood can be expressed as

$$\begin{aligned} p(e|B_\gamma, D, \gamma) &= (2\pi)^{-nm/2}\prod_{k=1}^{m}(d_k)^{n/2} \\ &\quad \times \exp\left(-\frac{d_k}{2}\{S_k(\gamma) + (\beta_k - m_k)' A_k (\beta_k - m_k)\}\right), \quad (3) \end{aligned}$$

where $m_k = -A_k^{-1} a_k$ and $S_k(\gamma) = a_{k,k} - a_k' A_k^{-1} a_k$.

2.3 Prior Specification

We begin with the conditional prior $p(B|\gamma, D)$, which is similar to the priors used to account for variable and subset uncertainty in linear regression (see George and McCulloch 1997; Kohn, Smith, and Chan 2001 for a discussion of the literature). Given γ, some of the lower-triangular elements of B are fixed to be zero, and a prior is required for the unconstrained elements, which we denote as B_γ. We construct a fractional conditional prior for B_γ by setting

$$p(B_\gamma|\gamma, D) \propto p(e|B, D, \gamma)^{1/n}. \quad (4)$$

The rationale for such a prior is that it is similar to the likelihood but provides only $1/n$th of the weight provided by the likelihood. It follows from (3) that

$$p(\beta_1, \ldots, \beta_m|D, \gamma) = \prod_{k=1}^{m-1} p(\beta_k|D, \gamma) \quad \text{with}$$

$$\beta_k|D, \gamma \sim N(m_k, \Omega_k),$$

where m_k is as defined earlier and $\Omega_k = \frac{n}{d_k} A_k^{-1}$.

The indicator variables $\gamma_{i,j}$, where $j = 1, \ldots, m-1$ and $i > j$, are taken a priori independent, with $p(\gamma_{i,j} = 1|\omega) = \omega$. This implies that for given ω, the prior expected number of nonzero elements in the strict lower triangle of B is $m(m-1)\omega/2$. We regard ω as an unknown parameter and assume that it has a uniform prior on [0, 1]. To make the simulation in Section 3 more efficient, the parameter ω is integrated out of the analysis, so that

$$\begin{aligned} p(\gamma) &= \int p(\gamma|\omega)p(\omega)\,d\omega \\ &= \int_0^1 \omega^{q_\gamma}(1-\omega)^{(r-q_\gamma)}\,d\omega = \text{beta}(q_\gamma + 1, r - q_\gamma + 1). \quad (5) \end{aligned}$$

In (5), beta$(\cdot, \cdot)$ is the beta function and $r = m(m-1)/2$ is the total number of elements in the strict lower triangle of B.

The prior for the diagonal elements $d_1, \ldots, d_m$ of D is

$$p(d_1, \ldots, d_m|\xi, \kappa) = \prod_{k=1}^{m} p(d_k|\xi, \kappa), \quad \text{where}$$

$$p(d_k|\xi, \kappa) = \frac{d_k^{\xi/\kappa - 1}\exp(-d_k/\kappa)}{\Gamma(\xi/\kappa)\kappa^{\xi/\kappa}}.$$

That is, the d_k are independent, with each d_k having a gamma distribution with parameters ξ/κ and κ, so the prior mean is ξ and the prior variance is $\xi\kappa$. It can be shown that the joint posterior distribution (B, D, γ) is insensitive to small perturbations in ξ and κ, and that the improper prior $p(d_k) \propto 1/d_k$ leads to a proper posterior.

We set $\xi = 100$ and $\kappa = 1{,}000$ in all of our empirical work, which makes the prior for the d_i proper but uninformative for all of the examples in this article. Furthermore, for any other application, it is straightforward for the user to choose values of ξ and κ that make the prior for the d_i uninformative but consistent with the likelihood.

We also looked at a shrinkage prior for the d_i to see whether estimation of the covariance matrix could be further improved. In particular, we considered the hierarchical shrinkage prior that assumes that the $\log(d_i)$ are independent and $N(\mu_d, \tau_d)$, and the prior for μ_d and τ_d is $p(\mu_d, \tau_d) \propto 1/\tau_d$. We note that it is straightforward to sample the d_i from the first prior because it is conjugate, as explained in Section 3.2, whereas the log-normal prior requires a Metropolis–Hastings step for generation of the d_i.

Comparing the performance of the two priors on the simulated examples in Section 4, we found that they performed similarly, except that the shrinkage prior resulted in a more efficient estimator when $\Sigma = I$, because it allowed the diagonal elements of D to shrink to a common value close to 1. Sections 4 and 5 report the results for only the first prior, because the results using the second prior were similar, but using the second prior is computationally less efficient.

3. INFERENCE AND SIMULATION METHOD

3.1 Bayesian Inference

Suppose that $\gamma^{[j]}, B^{[j]}$ and $D^{[j]}, j = 1, \ldots, J$, is a sample from the joint posterior distribution $p(\gamma, B, D|e)$. In this section we show how to construct Bayes estimators of unknown parameters based on this sample. We first consider posterior mean estimators, which correspond to a squared error loss function.

3.1.1 Posterior Mean Estimators. The posterior mean of B is $E(B|e)$, which can be expressed as

$$E(B|e) = \sum_{\gamma} E(B|\gamma, e)p(\gamma|e). \tag{6}$$

The estimator (6) is often called a Bayesian model average estimator because it is a weighted average of the posterior means of B, conditional on the configuration or model γ, with the weights being the posterior probabilities $p(\gamma|e)$ (see Raftery et al. 1997).

Evaluating $E(B|e)$ analytically is computationally difficult because it requires integrating out the model parameter γ that has $2^r = 2^{m(m-1)/2}$ possible values. Thus we estimate it by

$$\widehat{B} = J^{-1} \sum_{j=1}^{J} E(B|\gamma^{[j]}, D^{[j]}, e). \tag{7}$$

In calculating the conditional expectation in (7), the upper-triangular elements of B are always zero and the diagonal elements are always one, whereas the elements $b_{i,k}|\gamma_{i,k} = 0$ are fixed exactly to zero. The posterior conditional distribution of the remaining elements B_γ is calculated in Section A.1 of the Appendix and shows that $\beta_1, \ldots, \beta_{m-1}$ are independently distributed with conditional posterior mean $E(\beta_k|D, \gamma, e) = -A_k^{-1}a_k$, where the matrix A_k and the vector a_k are as defined in Section 2.2. Similarly, the mixture estimator of the posterior mean $E(D|e)$ is given by

$$\widehat{D} = J^{-1} \sum_{j=1}^{J} E(D|\gamma^{[j]}, B^{[j]}, e).$$

Section A.2 shows that the diagonal elements of D are independently conditionally distributed gamma, so that it is straightforward to compute the conditional expected value of D.

Histogram estimates of the posterior means $E(\Sigma|e)$ and $E(\Sigma^{-1}|e)$ can be constructed based on the iterates $D^{[j]}$ and $B^{[j]}, j = 1, \ldots, J$. For example, the histogram estimate of the posterior mean of Σ^{-1} is

$$\widehat{\Sigma^{-1}} = J^{-1} \sum_{j=1}^{J} B^{[j]} D^{[j]} (B')^{[j]}.$$

It is also possible to get Monte Carlo estimates of the marginal posterior probability intervals for all parameters. If θ is a scalar quantity of interest, then the lower and upper bounds of a $100(1-\alpha)\%$ marginal posterior probability interval for θ can be estimated by counting off the $J\alpha/2$ lowest and highest generated values $\theta^{[j]} \sim \theta|e$. For example, Section 5 carries out Bayesian inference on the partial correlations of Σ, which are simple functions of the elements of Σ^{-1}.

The expected proportion of non-zero lower triangular elements in B is given by the posterior mean $E(\omega|e)$, which is estimated by

$$\hat{\omega} = \frac{1}{J} \sum_{j=1}^{J} q_{\gamma^{[j]}}/r.$$

Note that ω is a measure of the level of parsimony in Σ. If $\omega = 0$, then $\Sigma = D^{-1}$ and the elements of e_i are uncorrelated. If $\omega = 1$, then B has all non-zeros on the lower triangle, and there are $m(m+1)/2$ free parameters to estimate in the decomposition of Σ^{-1}.

In some applications, such as that in Section 5.2, it is useful to estimate $p(\gamma_{i,k} = 1|e)$, which is the marginal posterior probability that $b_{i,k}$ is non-zero. We estimate this probability using the histogram estimate

$$\hat{p}(\gamma_{i,k} = 1|e) = \frac{1}{J} \sum_{j=1}^{J} \gamma_{i,k}^{[j]}.$$

3.1.2 Bayes Estimators for Other Loss Functions. It is also common to calculate Bayes estimators of the covariance matrix with respect to various alternative loss functions. In particular, Yang and Berger (1994) studied the loss functions

$$L_1(\widehat{\Sigma}, \Sigma) = \operatorname{tr}(\widehat{\Sigma}\Sigma^{-1}) - \log|\widehat{\Sigma}\Sigma^{-1}| - m$$

and

$$L_2(\widehat{\Sigma}, \Sigma) = \operatorname{tr}(\widehat{\Sigma}\Sigma^{-1} - I)^2$$

and showed that the Bayes estimators for Σ for these two loss functions are

$$\delta_1 = (E[\Sigma^{-1}|e])^{-1}$$

and

$$\operatorname{vec}(\delta_2) = (E[\Sigma^{-1} \otimes \Sigma^{-1}|e])^{-1} \operatorname{vec}(E[\Sigma^{-1}|e]).$$

The posterior expected values in the formulas for δ_1 and δ_2 can be estimated by calculating their histogram estimates from the Monte Carlo sample.

The L_1 and L_2 loss functions for Σ^{-1} are $L_1(\widehat{\Sigma}^{-1}, \Sigma^{-1})$ and $L_2(\widehat{\Sigma}^{-1}, \Sigma^{-1})$, and the Bayes estimators are $\delta_3 = (E[\Sigma|e])^{-1}$ and $\operatorname{vec}(\delta_4) = (E[\Sigma \otimes \Sigma|e])^{-1} \operatorname{vec}(E[\Sigma|e])$. Yang and Berger (1994) have given a more detailed discussion of inference for covariance matrices using differing loss functions.

3.2 Markov Chain Monte Carlo Sampling

This section shows how to generate a sample from the joint posterior distribution of (Σ, B, D) using the following MCMC sampling scheme, with ω integrated out as in (5).

3.2.1 Sampling Scheme. The sampling scheme comprises the following three steps:

1. Generate from $B|\gamma, D, e$.
2. Generate from $D|B, \gamma, e$.
3. Generate from $\gamma_{i,j}|\gamma_{\setminus i,j}, D, e$, for $j = 1, \ldots, m-1$, $i > j$.

In step 3, the notation $\gamma_{\setminus i,j}$ means γ with the element $\gamma_{i,j}$ excluded. Steps 1–3 are repeated for a warmup period, at the end of which it is assumed that the iterates are generated from the joint posterior distribution. A further J iterates are used as a Monte Carlo sample for inference.

Generating the free elements of B from the conditional distribution in step 1 is straightforward, because each of the columns of unconstrained elements (β_k) is independent and has a multivariate normal distribution. Generating from the conditional distribution in step 2 is also straightforward, because the diagonal elements of D are independently distributed with a gamma distribution. Calculation of these posterior distributions is given in Sections A.1 and A.2 of the Appendix.

We note here that the computational efficiency of our approach for generating B_γ and D is due primarily to the form of the likelihood when the Cholesky factorization is used, rather than the prior. The likelihood is Gaussian in B_γ, with the columns of B_γ independent of one another. Thus whatever prior is imposed on B_γ, the likelihood can be used to form a Gaussian proposal density for B_γ and the Metropolis–Hastings method can then be used to correct for the prior. Similarly, the likelihood is a gamma distribution in the d_i, with the d_i independent of each other. Hence the likelihood can again be used as the proposal density for the d_i, with the Metropolis–Hastings method used to correct for the prior. That is, it is the form of the Cholesky decomposition in the likelihood, not the prior, that creates the computational simplicity.

Step 3 generates an iterate of γ one element at a time, using the priors discussed in Section 2.3. This generation can be carried out using a variety of MCMC sampling methods, including Gibbs sampling as done by Smith and Kohn (1996). However, we use the sampling step outlined here because it is more efficient computationally than a Gibbs sampler. The gain in computational efficiency is important, because $r = m(m-1)/2$ indicators must be generated in each iteration, and r becomes large as m increases.

3.2.2 Generating $\gamma_{i,j}$. This step uses the decomposition of the conditional posterior into a product of the likelihood and a conditional prior density,

$$\underbrace{p(\gamma_{i,j}|\gamma_{\setminus i,j}, D, e)}_{\pi^*(\gamma_{i,j})} \propto \underbrace{p(e|\gamma, D)}_{l(\gamma_{i,j})} \underbrace{p(\gamma_{i,j}|\gamma_{\setminus i,j})}_{\pi(\gamma_{i,j})}, \tag{8}$$

where π^*, l, and π are a shorthand notation for the posterior, likelihood, and conditional priors. Sections A.3 and A.4 in the Appendix show how to calculate these quantities. The transition kernel Q that we use to generate $\gamma_{i,j}$ is

$$Q(\gamma^c_{i,j} = 1 \to \gamma^g_{i,j} = 0) = \pi(\gamma_{i,j} = 0)\frac{l(\gamma_{i,j} = 0)}{l(\gamma_{i,j} = 1) + l(\gamma_{i,j} = 0)}$$

and

$$Q(\gamma^c_{i,j} = 0 \to \gamma^g_{i,j} = 1) = \pi(\gamma_{i,j} = 1)\frac{l(\gamma_{i,j} = 1)}{l(\gamma_{i,j} = 1) + l(\gamma_{i,j} = 0)},$$

where $\gamma^c_{i,j}$ is the current value of $\gamma_{i,j}$ and $\gamma^g_{i,j}$ is the generated $\gamma_{i,j}$ value. Kohn et al. (2001) showed that this transition kernel satisfies the detailed balance criterion and has π^* as its invariant distribution. The following algorithm shows how to generate γ_{ij} using this transition kernel in a way that is more efficient computationally than generating directly from the conditional density $\pi^*(\gamma_{i,j})$.

Algorithm for generating $\gamma_{i,j}$

1. Generate u from a uniform distribution on (0, 1).
2. a. If $\gamma^c_{i,j} = 1$ and $u > \pi(\gamma_{i,j} = 0)$, then set $\gamma^g_{i,j} = 1$.
   b. If $\gamma^c_{i,j} = 1$ and $u < \pi(\gamma_{i,j} = 0)$, then generate $\gamma^g_{i,j}$ from the density

$$\Pr(\gamma_{i,j} = 0) = \frac{l(\gamma_{i,j} = 0)}{l(\gamma_{i,j} = 1) + l(\gamma_{i,j} = 0)}.$$

   c. If $\gamma^c_{i,j} = 0$ and $u > \pi(\gamma_{i,j} = 1)$, then set $\gamma^g_{i,j} = 0$.
   d. If $\gamma^c_{i,j} = 0$ and $u < \pi(\gamma_{i,j} = 1)$, then generate $\gamma^g_{i,j}$ from the density

$$\Pr(\gamma_{i,j} = 1) = \frac{l(\gamma_{i,j} = 1)}{l(\gamma_{i,j} = 1) + l(\gamma_{i,j} = 0)}.$$

Note that in cases (a) and (c), it is only necessary to compute $\pi(\gamma_{i,j})$, which is trivial. In examples where most of the lower triangular elements of B are identified as zero, $\pi(\gamma_{i,j} = 1)$ is usually close to zero, and case (c) will occur most frequently. Conversely, in examples where most of the lower-triangular elements are identified as non-zero, $\pi(\gamma_{i,j} = 0)$ will be close to zero, so case (a) will occur most frequently. In either case, substantial computational savings are obtained over generating directly from the posterior, which requires evaluation of $l(\gamma_{i,j})$ for all r indicators. (For a full discussion of this and related sampling schemes for efficiently generating binary variables, see Kohn et al. 2001.)

4. SIMULATION STUDY

This section reports the results of a simulation study comparing the performance of the estimator discussed in this article (which we label SK) with that of Yang and Berger (1994) (which we label YB94) and the unrestricted maximum likelihood estimator (MLE). The method of Yang and Berger (1994) is based on the reference prior of Chang and Eaves (1990) and results in a posterior density that shrinks the eigenstructure of Σ. This posterior density does not have a recognizable closed form and was estimated by Yang and Berger (1994) using a random-walk Metropolis–Hastings method.

We considered four example covariance structures that correspond to important cases in applied practice. Examples 1 and 2 are diagonal covariance matrices chosen because they are used in Yang and Berger (1994) and represent highly parsimonious covariance matrices. Examples 3 and 4 are the covariance structures of autoregressive and moving average models of lag length 1. For simplicity, the mean of the data is assumed to be 0.

Example 1 (Homoscedastic System). Here $B = I$ and $D = I$, so the variables are independent with $\Sigma = I$.

Example 2 (Heteroscedastic System). Here $B = I$ and $D = \text{diag}(\frac{1}{m}, \frac{1}{m-1}, \ldots, 1)$, so $\Sigma = \text{diag}(m, m-1, \ldots, 1)$.

Example 3 (Autoregressive Covariance). We used the factorization of the inverse covariance matrix of a first-order autoregressive process with autoregressive coefficient $\theta = .8$. That is, $b_{i+1,i} = -\theta$ for $i = 1, \ldots, m-1$; $b_{i,j} = 0$ for $j = 1, \ldots, m-2$ and $i > j+1$; $d_m = 100/(1-\theta^2)$; and $d_i = 100$ for $i = 2, \ldots, m$.

Example 4 (Moving Average Covariance). We used the factorization of the inverse covariance matrix of a first-order moving average process with coefficient $\theta = .8$. The lower-triangular factor B of this covariance structure is full, so that $b_{i,j} \neq 0$ for all $i > j$. The elements on the ith band of B are of $O(-\theta^i)$.

We used three cases of (m, n) in the simulation: $(m = 5, n = 40)$, $(m = 15, n = 40)$, and $(m = 30, n = 100)$. We generated 100 datasets for each example in each of the three cases.

The matrices Σ and Σ^{-1} were estimated for the simulated data, using the Bayes estimators for squared error loss (posterior means) and the L_1 and L_2 loss functions discussed in Section 3.1. The estimators were calculated using the priors and sampling scheme suggested in this article, and also those suggested by Yang and Berger (1994). In addition, the MLEs of Σ and Σ^{-1},

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} e_i e_i' \qquad \text{and} \qquad \widehat{\Sigma^{-1}} = (\widehat{\Sigma})^{-1},$$

were also calculated.

The performance of the MLE and the posterior mean estimators was assessed by calculating the squared error loss function

$$D(\widehat{\Sigma}, \Sigma) = \frac{1}{m^2}\left(\sum_i \sum_j (\hat{\sigma}_{ij} - \sigma_{ij})^2\right)^{1/2}.$$

The performance of the Bayes estimators for L_1 and L_2 was assessed relative to these loss functions. Table 1 provides the median values of the squared error, L_1 and L_2 losses obtained in the simulations for each method. For the diagonal and tridiagonal inverse covariance matrices in examples 1–3, the SK estimator clearly outperforms the MLE and the YB94 estimator on all metrics because it automatically identifies the parsimonious Cholesky structure of (1) in these examples.

In example 4, the SK estimator is less efficient than either the YB94 estimator or the MLE in the first case $(n = 40, m = 5)$, because the Cholesky factor B is made up of non-zero elements. However, as the size of the matrix increases, as in the two cases $(n = 40, m = 15)$ and $(n = 100, m = 30)$, the relative performance of the SK estimator improves because the elements $b_{i,j}$ are of $O((-.8)^{i-j})$, and as m gets larger, a higher proportion of the lower-triangular elements are close to zero, so that a parsimonious representation on the scale of the factor B is a reasonable approximation. For this example, the median estimated values of the expected posterior proportion of non-zero strict lower-triangular elements, $\hat{\omega}$ (as defined in Sec. 3.1), were .83, .272 and .183 for the three cases.

It has long been recognized that the MLE of the eigenvalues of Σ is biased (see e.g., Stein 1956; Dempster 1969).

Table 1. Sample Medians of the Metrics D,L_1, and L_2 for the Estimation of Both Σ and Σ^{-1} by the Two Bayesian Methods (YB94 and SK) and the MLE

Metric estimator		Matrix distance measure D			L_1		L_2	
		YB94	SK	MLE	YB94	SK	YB94	SK
For estimation of Σ								
$n = 40$	Eg1	.022	.022	.034	.129	.119	.266	.232
$m = 5$	Eg2	.086	.061	.100	.277	.110	.490	.222
	Eg3	8.9×10^{-4}	9.3×10^{-4}	8.7×10^{-4}	.330	.263	.570	.493
	Eg4	.051	.064	.054	.330	.388	.573	.656
$n = 40$	Eg1	6.6×10^{-3}	4.1×10^{-4}	1.1×10^{-2}	1.353	.409	2.013	.800
$m = 15$	Eg2	.066	.034	.086	2.456	.356	3.654	.682
	Eg3	3.4×10^{-4}	2.6×10^{-4}	3.0×10^{-4}	2.468	.963	3.550	1.681
	Eg4	.014	.018	.018	2.966	3.854	4.365	7.993
$n = 100$	Eg1	2.5×10^{-3}	8.8×10^{-4}	3.4×10^{-3}	2.938	.291	NC	NC
$m = 30$	Eg2	.045	.015	.053	4.679	.310	NC	NC
	Eg3	1.1×10^{-4}	6.7×10^{-5}	9.6×10^{-5}	4.909	.656	NC	NC
	Eg4	5.0×10^{-3}	4.1×10^{-3}	5.5×10^{-3}	5.727	5.661	NC	NC
For estimation of Σ^{-1}								
$n = 40$	Eg1	.021	.020	.041	.128	.123	.264	.250
$m = 5$	Eg2	.015	.008	.017	.288	.107	.521	.215
	Eg3	4.387	4.017	5.528	.345	.260	.660	.473
	Eg4	.049	.056	.052	.360	.429	.599	.785
$n = 40$	Eg1	.012	.005	.026	1.593	.414	3.304	.807
$m = 15$	Eg2	3.3×10^{-3}	1.2×10^{-3}	5.9×10^{-3}	2.606	.362	4.733	.712
	Eg3	1.590	1.021	4.012	2.824	.961	4.958	1.732
	Eg4	.034	.048	.058	3.198	3.212	5.52	4.44
$n = 100$	Eg1	4.5×10^{-3}	8.8×10^{-4}	6.4×10^{-3}	3.980	.299	NC	NC
$m = 30$	Eg2	7.6×10^{-4}	1.7×10^{-4}	1.0×10^{-3}	5.594	.302	NC	NC
	Eg3	.596	.180	1.083	5.583	.681	NC	NC
	Eg4	.016	.021	.019	7.643	3.994	NC	NC

NOTE: The results are presented for each of the four examples and for each of the three sample size/matrix dimension cases. The L_2 estimator is not calculated for the case where $m = 30$, because it involves the inversion of a $m^2 \times m^2$ matrix.

Table 2. Median Values of the Eigenrange and the Condition Number From the Estimation of Σ by the MLE, the Posterior Mean YB94 and SK Estimators, and the True Matrix

	Case 1: $n = 40, m = 5$				Case 2: $n = 40, m = 15$				Case 3: $n = 100, m = 30$			
	Eg1	Eg2	Eg3	Eg4	Eg1	Eg2	Eg3	Eg4	Eg1	Eg2	Eg3	Eg4
Median range of eigenvalues of $\hat{\Sigma}$												
YB94	.570	3.96	.10	2.89	1.36	16.6	.18	3.83	1.52	34.9	.21	4.15
SK	.573	4.51	.10	3.49	.87	16.2	.20	5.02	.58	32.7	.23	4.40
MLE	1.034	4.98	.10	3.29	2.11	22.18	.19	4.96	2.00	44.8	.23	5.00
True	0	4.00	.10	2.77	0	14.0	.18	3.14	0	29.0	.21	3.18
Median condition number of $\hat{\Sigma}$												
YB94	1.31	2.07	5.21	3.14	2.17	3.84	7.95	6.12	2.40	6.12	10.1	9.79
SK	1.30	2.34	5.87	3.53	1.51	4.22	8.64	4.47	1.33	5.86	9.18	5.00
MLE	1.72	2.61	6.18	3.85	3.48	6.12	13.6	10.72	3.09	8.35	14.4	12.7
True	1.00	2.24	5.46	3.45	1.00	3.87	7.65	6.73	1.00	5.48	8.45	8.19
Percentage of accepted iterates in the Metropolis–Hastings sampler of the YB94 estimator												
	11.7%	15.4%	17.0%	16.5%	6.9%	8.7%	10.0%	9.7%	3.2%	4.5%	5.1%	4.7%

NOTE: The last row contains the average percentage of accepted iterates from the sampling scheme of Yang and Berger (1994) of the YB94 estimator during the simulations.

In particular, the estimated range of the eigenvalues tends to be larger than the true range. Various authors, including Yang and Berger (1994), have attempted to correct for this bias. In our simulations, we studied the estimated range of the eigenvalues. Rows 1–4 in Table 2 provide the median eigenrange for the MLE and the posterior mean estimators of Σ, along with the true eigenrange. The results confirm that the MLE distorts the eigenrange in these examples and that the YB94 estimator shrinks the eigenrange toward the true value. Similarly, in examples 1 and 2, the SK estimator also shrinks the eigenrange. Note, however, that in example 3 the SK estimator does not shrink the eigenrange, yet it is still a much more efficient estimator for this example than YB94. In example 4, the range of the eigenvalues of the SK estimator is biased upward similarly to the MLE, except in the third case when $m = 30$. In this case, the SK estimator is the most efficient without shrinking the eigenrange as much as the YB94 estimator.

We also studied the condition number of the estimators. Rows 5–7 of Table 2 contain the median condition number of the estimates, whereas row 8 contains the condition number of the true covariance matrix. Note that for these examples, the YB94 estimator tends to have a condition number close to that of the true matrix, the SK estimator tends to have a condition number lower than that of the true matrix, and the MLE tends to have a condition number greater than the true matrix.

When the data are longitudinal, as in the autoregressive model of example 3, correct identification of the structure of the Cholesky factor B in (1) is important, because it may suggest an appropriate model. We therefore investigated the proportion of correctly identified non-zero (P_1) and zero (P_0) strict lower-triangular elements of B. This was undertaken for example 3, which has both zero and non-zero elements in the strict lower triangle of B.

We considered the i, jth lower-triangular element that is truly non-zero to be correctly identified if the histogram estimate $\hat{p}(\gamma_{i,j} = 1|e) > .5$. We consider the i, jth lower-triangular element that is truly zero to be correctly identified if the histogram estimate $\hat{p}(\gamma_{i,j} = 1|e) < .5$. Otherwise, the lower-triangular element is incorrectly identified. For each estimate, we calculated P_0 and P_1. Table 3 summarizes the results over the 100 simulated datasets in each of the three cases and shows that the method is excellent at identifying the structure of the Cholesky factor (and therefore the correlation structure) for this example.

The simulations suggest the following conclusions. First, if a high proportion of the elements of the factor B are exactly zero (or approximately equal to zero, as in example 4 when $m = 30$), then we expect the SK estimator to be more statistically efficient than the YB94 estimator because it exploits the parsimony of the Cholesky parameterization. However, when this is not the case (as in example 4 with $m = 5$), there is unlikely to be any benefit in using the SK estimator. Second, our approach successfully distinguished between lower-triangular elements of B that are significantly different from zero and those that are not.

In implementing the SK estimator, the sampling periods used, quoted as the pair (burn-in sample, Monte Carlo sample), were (5000, 5000) for cases 1 and 2 and (10000, 10000) for case 3. Row 9 of Table 2 provides the percentage of accepted iterates in the random walk Metropolis–Hastings sampling method used by Yang and Berger (1994). The acceptance rates of the Yang and Berger method are low and decline as m increases, so that it is around 5% in case 3 when $m = 30$. Therefore, we used sampling periods of (10000, 15000) for cases 1 and 2 and (25000, 25000) for case 3. We found these sampling periods to be sufficient in the sense that the estimates we report were very similar to those obtained for longer sampling periods.

Table 3. Summary Statistics of the Percentage of the Strict Lower Triangular Elements of B Correctly Identified for Example 3, Calculated for the Percentage of Correctly Identified Zeros (P_0) and Correctly Identified Non-Zeros (P_1)

Case	$\bar{P}_1$	$\bar{P}_0$	Min(P_1)	Min(P_0)
1 ($m = 5, n = 40$)	100.0%	100.0%	100.0%	100.0%
2 ($m = 15, n = 40$)	99.5%	98.9%	92.9%	98.9%
3 ($m = 30, n = 100$)	100%	100%	100%	100%

5. APPLICATIONS

5.1 Introduction

This section presents three substantive applications drawn from the biometry, econometric, and finance literatures. The covariance matrix in each example is large relative to the available sample size and has an unknown form. Nevertheless, the resulting Bayesian estimates suggest that each covariance structure can be represented in a highly parsimonious manner when considering the decomposition in (1). Identifying and exploiting this parsimony results in more efficient estimates.

In this section we work with the following regression model because all three examples require a non-zero mean:

$$y = X\alpha + e. \tag{9}$$

In (9), y is the $mn \times 1$ observation vector, α is the vector of regression coefficients, and $e \sim N(0, I_n \otimes \Sigma)$ is the error vector. The prior for α is $p(\alpha) \propto$ constant, the conditional prior for B_γ is defined by (4), with $e = y - X\alpha$, whereas the priors for D and γ are the same as in Section 2.3. To obtain a sample from the posterior distribution, we must augment the sampling scheme in Section 3.2 by a step that generates α. The conditional posterior density of α is

$$\begin{aligned} p(\alpha|y, B_\gamma, D, \gamma) &\propto p(y|\alpha, B_\gamma, D, \gamma)p(B_\gamma|\gamma, \alpha, D) \\ &\propto p(y|\alpha, B_\gamma, D, \gamma)^{1+1/n}. \end{aligned}$$

It is straightforward to show that this conditional posterior density is normal with mean $(X'(I \otimes \Sigma^{-1})X)^{-1}X'(I \otimes \Sigma^{-1})y$ and covariance matrix $n(1+n)^{-1}(X'(I \otimes \Sigma^{-1})X)^{-1}$. It follows that the mixture estimate of α is

$$\begin{aligned} \hat{\alpha} &= \frac{1}{J}\sum_{k=1}^{J} E\left(\alpha|y, \Sigma^{[k]}\right) \\ &= \frac{1}{J}\sum_{k=1}^{J} \left(X'\left(I \otimes (\Sigma^{-1})^{[k]}\right)X\right)^{-1} X'\left(I \otimes (\Sigma^{-1})^{[k]}\right)y \end{aligned}$$

5.2 Repeated Measures in a Longitudinal Study

There is an extensive literature on analyzing longitudinal studies using repeated measures (for examples, see Diggle et al. 1994; Hand and Crowder 1990). In many of these studies a sequence of measurements is taken on a number of subjects, and the analysis assumes that the time series of observations on each subject are correlated, but the observations between subjects are independent. The time series covariance matrix is usually assumed to follow a simple parametric form.

Here we demonstrate the use of our method in identifying the time series covariance structure of the live weight of $n = 25$ cows measured at $m = 23$ unequally spaced points in time. The data, discussed by Diggle et al. (1994, p. 100), are the result of a longitudinal study with a 2×2 factorial design. These authors model the logarithm of the live weight with a different quadratic function for each of the four treatment categories in the factorial design, so that if cow i gets treatment k, then

$$\log(y_{t,i}) = \left(a_k + b_k T_t + c_k T_t^2\right) + e_{t,i} \quad \text{for } i = 1, \ldots, n,\ t = 1, \ldots, m.$$

Here $y_{t,i}$ is the live weight of the ith cow at the tth observation, T_t is the time from the start of the study of the tth observation, and $(a_k, b_k, c_k), k = 1, \ldots, 4$, are the regression coefficients. The model can be written in the form (9), where $y = (y_1', \ldots, y_n')'$, $y_i = (y_{1,i}, \ldots, y_{m,i})'$, $\alpha = (a_1, b_1, c_1, \ldots, a_4, b_4, c_4)'$, and X is the appropriate $nm \times 12$ design matrix.

The error structure is assumed to be independent across cows, but not across time. Diggle et al. (1994) assumed a parametric time series covariance structure for the errors. However, the advantage of our approach is that such parametric assumptions are unnecessary and we only assume that $e_i = (e_{1,i}, \ldots, e_{m,i})'$ are independently distributed $N(0, \Sigma)$. We use our procedure to estimate the unknown covariance matrix Σ parsimoniously.

Table 4 gives the estimated factor $\widehat{B}$. Only the lower-triangular elements that have estimated probability of being non-zero $> .25$ [i.e., $\hat{p}(\gamma_{i,j} = 1|y) > .25$] are reported. The factor B is identified as sparse and corresponds to a nonstationary antedependence times series model with time-varying parameters and a short, but varying, lag length.

5.3 Intraday Electricity Demand

There is a large literature on the econometric modeling of intraday electricity demand (see Smith 2000 for a recent summary). Electricity utilities use such models to forecast demand at an intraday level to guide their day-to-day operational decisions and determine any pricing policies. One popular method is to estimate separate equations for each intraday period and allow the errors to be correlated across equations (Fiebig, Bartels, and Aigner 1991). Such models are called "seemingly unrelated regressions" (Greene 1997), and we develop such a model for total New South Wales (NSW) electricity demand observed during the $n = 15$ working days of the 3-week period of March 8–28, 1993. Longer lengths of data are not viable for the model at (10), because the parameters in the model vary over season and are effectively static only over periods of 2–3 weeks (Harvey and Koopmans 1993).

We use half-hourly average demand data, so that there are $m = 48$ regression equations with an unknown 48×48 error covariance matrix Σ. Each regression features a constant μ_i, along with a temperature effect based on the average half-hourly temperature $T_{i,t}$ measured at the Sydney central business district, which is the central point of NSW electricity demand. The temperature effect is modeled as an inverted V with the point of inversion at 18.3 °C (65 °F), which is a commonly used nonlinear functional form for the relationship between electricity demand and temperature (see Engle, Granger, Rice, and Weiss 1986 for more details). The model considered is

$$y_{i,t} = \mu_i + \eta_i|T_{i,t} - 18.3| + e_{i,t}, \quad \text{for } i = 1, \ldots, 48, \quad t = 1, \ldots, 15, \tag{10}$$

Table 4. Estimated Cholesky Factor of Σ^{-1} for the Cow Live Weight Application

Column number																						
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*
—																						
−1.436	1																					
—	−.745	1																				
—	—	−.655	1																			
—	—	—	−.35	1																		
—	—	—	—	—	1																	
—	—	—	—	−.308	−.261	1																
—	—	—	—	—	−.263	−.578	1															
—	—	—	−.256	−.242	—	−.254	−1.09	1														
—	—	−.228	—	−.252	—	—	—	—	1													
—	—	—	—	—	—	—	—	−.766	−.546	1												
—	-	—	—	—	—	—	—	—	—	−.319	1											
—	—	—	—	—	—	—	—	—	—	—	−.434	1										
—	—	—	—	—	—	—	—	—	−.267	−.326	−.424	−.846	1									
—	—	—	—	—	—	—	—	—	—	—	—	—	−.67	1								
—	—	—	—	—	—	—	—	—	—	—	—	—	−.345	−.704	1							
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	−1.065	1						
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	−.914	1					
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	−.972	1				
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	−1.031	1			
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	−1.246	1		
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	−.981	1	
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	.157	—	−.879	1

NOTE: Only entries with estimated posterior probability of being non-zero >.25 [i.e., estimated $\hat{p}(\gamma_{i,j} = 1|y) > .25$] are given, with the others denoted by "—."

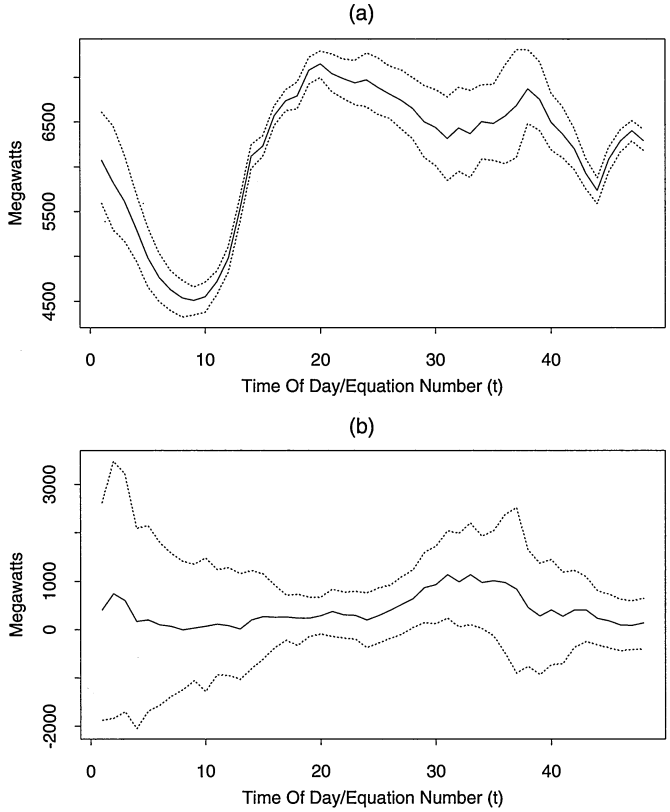

Figure 1. Posterior Mean Estimates (a) $\hat{\mu}_i$ and (b) $\hat{\eta}_i$ (———) Plotted Against the $i = 1, \ldots, 48$ Half-Hourly Intraday Periods. The dashed lines give the upper and lower estimated posterior 90% posterior confidence intervals.

where $y_{i,t}$ is NSW electricity demand at half hour i on day t. The regression model in (10) can be stacked and rewritten as (9), where $\alpha = (\mu_1, \eta_1, \ldots, \mu_m, \eta_m)'$, $y = (y_1', \ldots, y_n')'$, and X is the appropriate $(nm \times 2m)$ sparse design matrix. Figure 1 plots the estimates of the posterior means of μ_i and η_i against $i = 1, \ldots, 48$, together with the estimated 90% posterior probability intervals. Figure 1(a) provides an estimate of the so-called "demand profile" for working days (Harvey and Koopmans 1993), whereas 1(b) confirms the view in the literature that the response in demand to temperature changes depends on the time of day.

Figure 2(a) provides a graphical representation of the absolute value of the estimated correlation matrix derived from $\widehat{\Sigma}$.

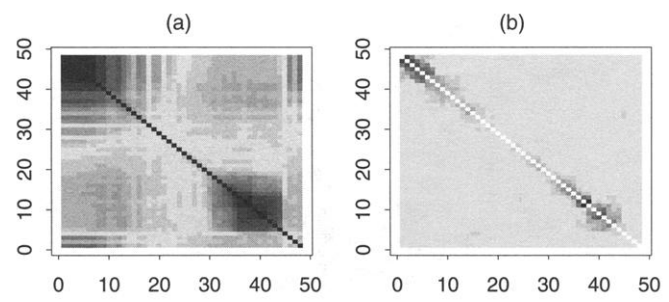

Figure 2. (a) Image Representation of the Absolute Value of the Estimated Correlation Matrix for the Estimated NSW Load Model, with Gray Scale From 0 (light) to 1.000 (dark), and (b) Image Representation of the Matrix of the Absolute Value of the Estimated Partial Correlations, with Gray Scale From 0 (light) to .287 (dark).

The off-diagonal elements of the correlation matrix are primarily positive and vary from $-.209$ to .971. There are large "chunks" of correlation between 0:30 to 7:00 (eqs. 1–14) and 18:00 to 22:30 (eqs. 36–45). These capture various day-specific demand levels, which are likely to be due to slight differences in working day demand profiles for different days of the week, as well as special events and television programming. However, the representation on this positive correlation is parsimoniously captured by the decomposition in (1), with $\hat{\omega} = .0278$, so that it is estimated that only 2.78% of the elements of the factor B are likely to be nonzero. Moreover, Figure 2(b) provides a representation of the matrix of the absolute value of the partial correlation estimates $|\hat{\rho}^{i,j}|$, where $\rho^{i,j}$ is the (i, j)th partial. These partial correlations are predominantly small, ranging from $-.005$ to .287, with the largest values clustered around the diagonal. We note that the partial correlations $\rho^{i,j}$ are given by

$$\rho^{i,j} = -\frac{\sigma^{i,j}}{\sqrt{\sigma^{i,i}\sigma^{j,j}}},$$

so that $\rho^{i,j} = 0$ if and only if $\sigma^{i,j} = 0$, where $\Sigma^{-1} = \{\sigma^{i,j}\}$.

We conclude the analysis of the electricity demand data by noting that the off-diagonal elements of B that are non-zero differ from iteration to iteration. Furthermore, in each iteration there are few such non-zero elements, so that $\hat{\omega}$ is small. This means that the modal estimates of B, D, and hence Σ differ substantially from the posterior mean estimates obtained by averaging over models. In contrast, in the previous example concerning the live weights of cows, the same elements of B are usually either zero or non-zero from iteration to iteration, because an antedependence model is appropriate for Σ. Therefore, in the live weight example, the modal estimates of B, D, and Σ are close to the model average estimates of their posterior means.

5.4 Estimation of a Multivariate Capital Asset Pricing Model

One of the most important developments in the area of finance is the capital asset pricing model (CAPM) (for a recent summary of this literature, see Campbell, Lo, and MacKinlay 1997). The CAPM shows that under some assumptions about the dynamics of stock returns, excess stock returns from individual firms are linearly related with excess market returns, without an intercept (Sharpe 1964). However, there is also known to be a strong correlation between the returns of individual firms over and above market returns due to industry factors (Fama and French 1997). Estimation of such correlation is particularly useful in allowing investors to form optimal portfolios of such assets. However, because the data used are typically sampled monthly, a limited number of observations are available for estimating the covariance matrix. Practitioners thus are usually limited to estimating a multivariate CAPM using only small numbers of stocks (Campbell et al. 1997). Nevertheless, our procedure allows for the estimation of a multivariate CAPM with a large number of stocks by identifying and exploiting the high degree of parsimony in the cross-sectional covariance of market-adjusted excess returns.

The data that we use consist of monthly excess returns for the $m = 89$ firms on the Standard and Poor 100 that traded

Table 5. Significant Partial Correlations, Along With Their 90% Posterior Probability Intervals

Tick and company name	*Primary industrial focus*	*Tick*	$\hat{\rho}^{i,j}$	*Confidence interval*
AEP: American Electric Power Inc.	Utility	SO	.661	(.566, .741)
AIG: American International Group Inc.	Finance	CI	.433	(.31, .551)
AIT: Ameritech Corp.	Telecommunications	BEL	.702	(.62, .775)
AMP: AMP Inc.	Electronics	MSFT	.344	(.216, .462)
ARC: Atlantic Richfield Co.	Petroleum	SLB, MOB	.371, .448	(.204, .497), (.328, .556)
BAC: Bankamerica Corp.	Banking	WFC	.411	(.286, .53)
BCC: Boise Cascade Corp.	Office products	CHA, WY	.455, .396	(.338, .569), (.294, .493)
BEL: Bell Atlantic Corp.	Telecommunications	AIT	.702	(.62, .775)
BMY: Bristol Myers Squibb Co.	Pharmaceuticals	MRK	.341	(.184, .472)
CHA: Champion International Corp.	Office products and resources	BCC	.455	(.338, .569)
CI: CIGNA Corp.	Finance	AIG	.433	(.31, .551)
ETR: Entergy Corp. New	Utility	UCM	.442	(.322, .555)
F: Ford Motor Co. Del.	Vehicle manufacture	GM	.427	(.313, .529)
GD: General Dynamics Corp.	Defense industries	HAL	−.262	(−.365, −.15)
GM: General Motors Corp.	Vehicle manufacture	JNJ, F	−.334, .427	(−.447, −.211), (.313, .529)
HAL: Halliburton Co.	Petroleum and construction	SLB, GD	.516, −.262	(.418, .606), (−.365, −.15)
INTC: Intel Corp.	Semiconductors	MSFT	.407	(.286, .524)
JNJ: Johnson & Johnson	Pharmaceutical	GM	−.334	(−.447, −.211)
LTD: Limited Inc.	Retail	WMT	.317	(.177, .44)
MOB: Mobil Corp.	Petroleum	ARC, XON	.448, .44	(.328, .556), (.327, .545)
MRK: Merck & Co. Inc.	Pharmaceuticals	BMY	.341	(.184, .472)
MSFT: Microsoft Corp.	Systems	INTC, AMP	.407, .344	(.286, .524), (.216, .462)
NSM: NL Semiconductor Corp.	Semiconductors	TXN	.447	(.331, .557)
SLB: Schlumberger Ltd.	Petroleum	ARC, HAL	.371, .516	(.204, .497), (.418, .606)
SO: Southern Co.	Utility	AEP	.661	(.566, .741)
TXN: Texas Instruments Inc.	Semiconductors and electronics	NSM	.447	(.331, .557)
UCM: Unicom Corp Holding Co.	Utility	ETR	.442	(.322, .555)
WFC: Wells Fargo & Co. New	Retail banking	BAC	.411	(.286, .53)
WMT: Wal-Mart Stores Inc.	Retail	LTD	.317	(.177, .44)
WY: Weyerhaeuser Co.	Construction and resources	BCC	.396	(.294, .493)
XON: Exxon Corp.	Petroleum	MOB	.44	(.327, .545)

NOTE: The first column gives the "tick," name, and primary industrial focus of the firm. The second column gives the tick of the firms with which it is signficantly partially correlated. The third column features the estimated partial correlations of these pairings, and the fourth column lists the respective 90% confidence intervals for the partial correlations.

continuously between November 1986 and November 1996. Excess returns are defined as the nominal return minus the risk-free rate, and, following standard financial practice, we use 1-month U.S. Treasury bills as a proxy for the risk-free rate. The sources of our data are as follows:

- *Risk-free rate*: f_t is the 1-month Treasury-bill rate at time t, which is the commonly used proxy for the risk-free rate of return [source: Center for Research in Security Prices (CRSP)].
- *Excess firm returns*: $y_{i,t} = r_{i,t} - f_t$, where $r_{i,t}$ is the return at time t for firm i (source: CRSP).
- *Excess market returns*: $x_t = m_t - f_t$, where m_t is the value-weighted market return on the NASDAQ/AMEX/NYSE exchanges at time t (source: CRSP).

The multivariate CAPM model first used by Gibbons (1982) is

$$y_t = \alpha x_t + e_t \quad \text{for } t = 1, \ldots, 120, \tag{11}$$

where $y_t = (y_{1,t}, \ldots, y_{89,t})'$ are the monthly excess returns on the 89 firms at time t, $\alpha = (\alpha_1, \ldots, \alpha_{89})'$ are regression coefficients, x_t is the excess market return at time t, and the error vectors $e_t = (e_{1,t}, \ldots, e_{89,t})'$ are independently distributed $e_t \sim N(0, \Sigma)$, with Σ an 89×89 matrix. The regression equations in (11) can be stacked together and rewritten in the form of (9), where $y = (y_1', \ldots, y_n')'$ and X is an $(nm \times m)$ matrix constructed by stacking the diagonal matrices $\text{diag}(x_t, x_t, \ldots, x_t)$ for $t = 1, \ldots, n$.

We infer from the data that there is a great deal of parsimony in the covariance structure with $\hat{\omega} = .0225$, suggesting that only 2.2% of the $r = 3{,}916$ lower triangular elements of B are effectively non-zero. We also computed the 90% posterior probability intervals for the 3,916 distinct partial correlations $\rho^{i,j}$ and found that only 19 of the partials were identified as non-zero. Table 5 provides the estimates of these partial correlations $\hat{\rho}^{i,j}$ and estimated probability intervals. These largely correspond to industry groupings; for example, Mircosoft is positively partially correlated with Intel; General Motors, with the other vehicle manufacturer, Ford; and Mobil, with the other petroleum producers, Exxon and Atlantic Richfield. For comparison, we estimated the multivariate CAPM in (11) without modeling the matrix Σ^{-1} parsimoniously (i.e., where $\gamma_{i,j} = 1$ was fixed). In this case, the number of significant partial correlations, as identified by their estimated posterior probability, was much higher at 1,546.

6. CONCLUSION

Our article obtains statistically efficient estimators of the covariance matrix for longitudinal data. It does so by factoring the inverse of the covariance matrix using the Cholesky decomposition and using a hierarchical Bayesian model for the factor B. Parsimony is built into the model by allowing the elements in the strict lower triangle of B to be identically equal to zero. Although our method is most applicable to longitudinal data, because the Cholesky decomposition has a clear interpretation, the results of Wermuth (1980) and

Roverato (2000), as well as the results for the finance application in Section 5, suggest that our method may also be useful for some cross-sectional covariance matrices. The simulation results in Section 4 suggest that the Bayes estimators based on our model compare favorably with those produced by the method of Yang and Berger (1994). The real examples in Section 5 demonstrate that our method can be applied to high-dimensional covariance matrices, whereas it seems impractical to apply the method of Yang and Berger (1994) to these examples, because the high rejection rate Metropolis–Hastings step becomes computationally prohibitive as the size of the matrix increases.

APPENDIX: CALCULATING THE CONDITIONAL DENSITIES REQUIRED IN THE SAMPLING SCHEME IN SECTION 3.2

A.1 Generating From $B|\gamma, D, e$

Only the unrestricted elements B_γ in this conditional density require generation. From Sections 2.2 and 2.3, their density is

$$p(B_\gamma|D,\gamma,e) \propto p(e|B_\gamma,\gamma,D)p(B_\gamma|\gamma,D) \propto p(e|B_\gamma,\gamma,D)^{1+1/n}.$$

It follows from (3) that

$$\beta_k|D,\gamma,e \sim N\left(-A_k^{-1}a_k, \frac{n}{d_k(n+1)}A_k^{-1}\right).$$

A.2 Generating From $D|\gamma, B, e$

The conditional density for the elements of the diagonal matrix D is given by

$$p(D|B,\gamma,e) \propto p(e|D,B,\gamma)p(B_\gamma|\gamma,D)p(D|\gamma) \propto p(e|D,B,\gamma)\prod_{k=1}^{m-1} p(\beta_k|\gamma,D)\prod_{k=1}^{m} p(d_k).$$

Substituting in the priors and the likelihood yields

$$p(D|B,\gamma,e) \propto \prod_{k=1}^{m} \exp\left(-d_k\left(\frac{h_k}{2}+\frac{1}{\kappa}\right)\right) d_k^{\frac{n+q_k}{2}+\frac{\xi}{\kappa}-1}, \qquad (A.1)$$

where

$$h_k = \begin{cases} S_k(\gamma)+\left(1+\frac{1}{n}\right)(\beta_k-m_k)'A_k(\beta_k-m_k) & \text{if } k<m \text{ and } q_k>0, \\ a_{k,k} & \text{otherwise.} \end{cases}$$

Thus the d_k are conditionally independent with gamma distributions.

A.3 Generating $\gamma_{i,j}$

To generate $\gamma_{i,j}$ in step 3 of the sampling scheme in Section 3.2, it is necessary to calculate the likelihood term $l(\gamma_{i,j})$ in (8), up to a constant of proportionality. From (3) and (4),

$$l(\gamma_{i,j}) = p(e|\gamma,D) \propto \int p(e|D,\gamma,B)p(B_\gamma|D,\gamma)\,dB_\gamma \propto (n+1)^{-q_j/2}\exp(T_j),$$

where $T_j = d_j a_j' A_j^{-1} a_j/2$.

A.4 Calculating the Prior $\pi(\gamma_{i,j}) = p(\gamma_{i,j}|\gamma_{\setminus i,j})$

The conditional prior can be calculated by integrating out the hyperprior ω, with $p(\gamma_{i,j}|\gamma_{\setminus i,j}) \propto \int_0^1 (\omega)^{q_\gamma}(1-\omega)^{(r-q_\gamma)}d\omega = B(q_\gamma+1, r-q_\gamma+1)$, so that $p(\gamma_{i,j}=1|\gamma_{\setminus i,j}) = 1/(1+h)$, where $h = (r-s)/(s+1)$ and $s = q_\gamma$ if $\gamma_{i,j}=0$ and $s = q_\gamma - 1$ if $\gamma_{i,j}=1$.

[Received May 1999. Revised January 2002.]

REFERENCES

Campbell, J., Lo, A., and MacKinlay, A. (1997), *The Econometrics of Financial Markets*, Princeton, NJ: Princeton University Press.

Chang, T., and Eaves, D. (1990), "Reference Priors for the Orbit in a Group Model," *The Annals of Statistics*, 18, 1595–1614.

Daniels, M., and Kass, R. (1999), "Nonconjugate Bayesian Estimation of Covariance Matrices," *Journal of the American Statistical Association*, 94, 1254–1263.

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

——— (1972), "Covariance Selection," *Biometrics*, 28, 157–175.

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

Engle, R., Granger, W., Rice, J., and Weiss, A. (1986), "Semiparametric Estimates of the Relationship Between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310–320.

Fama, E., and French, K. (1997), "Industry Cost of Equity," *Journal of Financial Economics*, 43, 153–194.

Fiebig, D., Bartels, R., and Aigner, D. (1991), "A Random Coefficient Approach to the Estimation of Residential End-Use Load Profiles," *Journal of Econometrics*, 50, 297–327.

Gabriel, F. (1962), "Ante-Dependence Analysis of an Ordered Set of Variables," *The Annals of Mathematical Statistics*, 33, 201–212.

George, E., and McCulloch, R. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Gibbons, M. (1982), "Multivariate Tests of Financial Models: A New Approach," *Journal of Financial Economics*, 14, 217–236.

Giudici, P., and Green, P. J. (1999), "Decomposable Graphical Gaussian Model Determination," *Biometrika*, 86, 785–801.

Golub, G., and van Loan, C. (1996), *Matrix Computations* (3rd ed.), Baltimore: Johns Hopkins University Press.

Greene, W. (1997), *Econometric Analysis* (3rd ed.), Englewood Cliffs, NJ: Prentice-Hall.

Hand, D., and Crowder, M. (1990), *Practical Longitudinal Data Analysis*, London: Chapman and Hall.

Harvey, A., and Koopmans, S. (1993), "Forecasting Hourly Electricity Demand Using Time Varying Splines," *Journal of the American Statistical Association*, 88, 1223–1253.

Kohn, R., Smith, M., and Chan, D. (2001), "Nonparametric Regression Using Linear Combinations of Basis Functions," *Statistics and Computing*, 11, 313–322.

Macchiavelli, R., and Arnold, S. (1994), "Variable Order Ante-Dependence Models," *Communications in Statistics*, 23, 2683–2699.

Pinheiro, J., and Bates, D. (1996), "Unconstrained Parameterizations for Variance-Covariance Matrices," *Statistics and Computing*, 6, 289–296.

Pourahmadi, M. (1999), "Maximum Likelihood Estimation for Generalised Linear Models for Multivariate Normal Covariance Matrices," *Biometrika*, 86, 677–690.

——— (2000), "Joint Mean-Covariance Models With Application to Longitudinal Data: Unconstrained Parameterization," *Biometrika*, 87, 425–435.

Roverato, A. (2000), "Cholesky Decomposition of a Hyper Inverse Wishart Matrix," *Biometrika*, 87, 99–112.

Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Sharpe, W. (1964), "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *Journal of Finance*, 19, 425–442.

Smith, M. (2000), "Modeling and Short Term Forecasting of New South Wales Electricity Load," *Journal of Business and Economic Statistics*, 18, 465–478.

Smith, M., and Kohn, R. (1996), "Nonparametric Regression via Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–344.

Stein, C. (1956), "Some Problems in Multivariate Analysis, Part I," Technical Report 6, Stanford University, Dept. of Statistics.

Wermuth, N. (1980), "Linear Recursive Equations, Covariance Selection and Path Analysis," *Journal of the American Statistical Association*, 75, 963–972.

Whittaker, J. (1990), *Graphical Models in Applied Mathematical Analysis*, Chichester, UK: Wiley.

Yang, R., and Berger, J. (1994), "Estimation of a Covariance Matrix Using the Reference Prior," *The Annals of Statistics*, 22, 1195–1211.