Biometrika (2004), **91**, 1, *pp.* 1–14
© 2004 Biometrika Trust
Printed in Great Britain

Bayesian correlation estimation

By JOHN C. LIECHTY

*Departments of Marketing and Statistics, Pennsylvania State University,
University Park, Pennsylvania 16802-3007, U.S.A.*

jcl12@psu.edu

MERRILL W. LIECHTY

*LeBow College of Business, Drexel University, Philadelphia, Pennsylvania
19104–2875, U.S.A.*

merrill@drexel.edu

and PETER MÜLLER

*Department of Biostatistics, University of Texas M. D. Anderson Cancer Center,
Houston, Texas 77030-4009, U.S.A.*

pm@mdacc.tmc.edu

Summary

We propose prior probability models for variance-covariance matrices in order to address two important issues. First, the models allow a researcher to represent substantive prior information about the strength of correlations among a set of variables. Secondly, even in the absence of such information, the increased flexibility of the models mitigates dependence on strict parametric assumptions in standard prior models. For example, the model allows a posteriori different levels of uncertainty about correlations among different subsets of variables. We achieve this by including a clustering mechanism in the prior probability model. Clustering is with respect to variables and pairs of variables. Our approach leads to shrinkage towards a mixture structure implied by the clustering. We discuss appropriate posterior simulation schemes to implement posterior inference in the proposed models, including the evaluation of normalising constants that are functions of parameters of interest. The normalising constants result from the restriction that the correlation matrix be positive definite. We discuss examples based on simulated data, a stock return dataset and a population genetics dataset.

Some key words: Covariance matrix; Mixture prior; Separation strategy.

1. Introduction

Few existing probability models and parameterisations for covariance matrices allow for easy interpretation and prior elicitation. We propose a collection of models in which correlations are grouped based on similarities among the correlations or based on groups of variables, and which may thereby incorporate substantive prior information. For example, for financial time series it is often known that some returns are more closely related than others. Even in the absence of substantive prior information, the additional flexibility of the models mitigates dependence on strict parametric assumptions in standard models. Our main goal is to model correlation matrices, with the resulting grouping of correlations or variables as an insightful by-product. Alternative approaches to modelling correlation structure build on factor analysis; see for example West (2003) and Aguilar & West (2000). While factor models effectively reduce the dimensionality of the covariance matrix, it is difficult to interpret the factors and loadings, which restricts the ability for researchers to suggest informative prior distributions. These models may also overlook natural groupings or clusters of the underlying variables. Another alternative approach is explored by Karolyi (1992, 1993), who uses Bayesian methods to estimate the variance of individual stock returns based on stocks grouped a priori according to size, financial leverage and trading volume.

The constraint of positive definiteness and the typically high-dimensional nature of the parameter vector for the covariance matrix influence the choice of prior probability models for covariance matrices. Lack of conjugacy becomes a problem with departure from the inverse-Wishart parameterisation (Chib & Greenberg, 1998). Other important considerations are the need to incorporate substantive prior information into the probability model and the desire that posterior simulation should be efficient and straightforward.

We assume throughout a multivariate normal model $y_i \sim N(0, \Sigma)$, for J-dimensional data y_i $(i = 1, \ldots, n)$. Extensions to normal regression models and hierarchical models with multivariate normal random effects distributions are straightforward (Daniels & Kass, 1999).

The most commonly used prior model is the conjugate inverse-Wishart (Bernardo & Smith, 1994, Ch. 3). However, in this model the degree of freedom parameter v is the only 'tuning parameter' available to express uncertainty.

Several noninformative default priors have been proposed for covariance matrices. Jeffreys' prior is $p_J(\Sigma) = 1/|\Sigma|^{(J+1)/2}$. Alternatively, Yang & Berger (1994) propose a reference prior, $p_R(\Sigma) \propto 1/\{|\Sigma| \prod_{i<j}(d_i - d_j)\}$, where d_i are the eigenvalues of Σ. As with other similarly parameterised models, the lack of intuition of the relationship between eigenvalues and correlations makes it difficult to interpret this model. Daniels (1999) proposes a uniform shrinkage prior, based on considering the posterior mean as a linear combination of the prior mean and the sample average and assuming a uniform prior on the coefficient for the sample average; see Christiansen & Morris (1997), Everson & Morris (2000) and Daniels & Kass (2001) for more discussion of shrinkage priors. The log matrix prior introduced by Leonard & Hsu (1992) uses a logarithmic transformation of the eigenvalue/eigenvector decomposition of Σ and allows for hierarchical shrinkage to be done with the eigenvalues. The dimension of the problem is reduced, but it is difficult to interpret the relationship of the log of the eigenvalues to the correlations and standard deviations.

Barnard et al. (2000) propose a separation strategy for modelling $\Sigma = SRS$ by assuming independent priors for the standard deviations S and the correlation matrix R. They propose two alternative prior models for R. One is the marginally uniform prior, in which the marginal prior for each r_{ij} in R is a modified beta distribution over $[-1, 1]$; with an appropriate choice of the beta parameters, this becomes a uniform marginal prior distribution. The other model for R is called the jointly uniform prior. Here the matrix R is assumed to be a priori uniformly distributed over all possible correlation matrices.

Daniels & Kass (1999) discuss three alternative hierarchical priors. The first is a hierarchical extension of the inverse-Wishart prior, which assumes priors on the degrees of freedom parameter and on the unknown elements of a diagonal scale matrix. Alternatively they consider a separation strategy as in Barnard et al. (2000) and assume a normal prior for a transformation of the correlation coefficients. The constraint of positive definiteness amounts to appropriate truncations of the normal prior. A third model uses an eigenvalue/eigenvector parameterisation, with the orthogonal eigenvector matrix parameterised in terms of the Givens angles.

Wong et al. (2003) propose a prior probability model on the precision matrix, $P = \Sigma^{-1}$, that is similar to our approach. Their application is geared towards graphical models and partial correlations, focusing on the sparseness of the precision matrix.

In this paper we introduce additional hierarchical structure by allowing for correlations to be grouped in natural ways. We consider three models. Throughout we assume a separation strategy, modelling standard deviations S and the correlation matrix R separately. We focus on modelling R, as including S in the proposed posterior simulation schemes is straightforward. Our 'common correlation' model assumes a common normal prior for all correlations, with the additional restriction that the correlation matrix is positive definite. This follows the frequentist work of Lin & Perlman (1985) who use a version of the James–Stein estimator to model the off-diagonal elements of the correlation matrix; see Daniels & Kass (2001) for additional discussion of covariance shrinkage models. The second model, the 'grouped correlations' model, generalises the common correlation model by allowing correlations to cluster into different groups, where each group has a different mean and variance. The third model, the 'grouped variables' model, allows the observed variables y_i to cluster into different groups. Correlations between variables in the same group have a common mean and variance and the correlations between variables in different groups have a mean and variance that depend on the group assignment for each variable.

Posterior inference will rely on Markov chain Monte Carlo methods (Tierney, 1994). However, posterior simulation for these models is computationally challenging, mainly because we need to sample from truncated distributions that result from the positive definiteness constraint. The truncated distributions involve analytically intractable normalising constants that are functions of the parameters; see Chen et al. (2000, Ch. 6), for a general discussion of related problems. We investigate the following three strategies for evaluating these normalising constants: sidestepping the problem by assuming that ratios of these normalising constants are approximately constant; using importance sampling strategies; and introducing an additional set of latent variables, called 'shadow' priors, in the hierarchical structure.

Section 2 introduces the three proposed models. Implementation and posterior simulation are discussed in § 3. Section 4 reviews possible areas of application for each model, with a discussion of feasible extensions. In § 5 we give examples and we conclude in § 6.

2. Models

2·1. *A motivating example*

Throughout, we assume a multivariate normal likelihood function and follow the separation strategy of Barnard et al. (2000), writing $\Sigma = SRS$, as discussed in § 1. Without loss of generality we assume that $S = I$. We write $\mathscr{R}^{\mathscr{J}}$ for the space of all correlation matrices of dimension J.

Example 1. In an effort to simplify the task of diversification, the finance community is interested in classifying companies into industries based on the types of products and services provided by a company. Typically this classification is done by individuals who have industrial expertise. While many of these classifications may be straightforward, a number of companies engage in strategies where they expand and/or change the products and services that they offer, creating hybrid companies that may not fit into a specific industrial classification. As an example, before its demise, Enron transformed their basic business from being an energy company to being a finance company. Since it is debatable whether or not an industry expert may be able to classify Enron correctly, it would be of interest to determine whether Enron's stock behaviour is correlated with energy companies or with finance companies. To illustrate, consider the monthly stock returns from April 1996 to May 2002 for nine companies which are either energy or finance companies. The energy stocks include Reliant, Chevron, British Petroleum and Exxon, and the financial stocks include Citi-Bank, Lehman Brothers, Merrill Lynch and Bank of America. The stock in the middle is Enron; see Table 1 for their empirical correlations.

The grouped variable model introduced in § 2·4 and which classifies stocks into groups based on the correlations within and between each group offers a natural method for determining whether or not Enron successfully made the transition from being an energy company to a financial company before they encountered recent troubles; see § 5.

Table 1. *Empirical correlation matrix for monthly stock returns for nine equity securities from April* 1996 *to May* 2002. *Reliant, Chevron, British Petroleum and Exxon are energy companies and Citi-Bank, Lehman Brothers, Merrill Lynch and Bank of America are financial services companies. Enron could potentially be in either group based on different criteria*

Variable		1	2	3	4	5	6	7	8	9
Reliant	1	1	0·30	0·14	0·31	0·47	0·10	0·15	0·13	0·09
Chev.	2	0·30	1	0·75	0·60	0·54	0·34	0·22	0·13	0·41
BP	3	0·14	0·75	1	0·59	0·44	0·22	0·23	0·11	0·25
Exxon	4	0·31	0·60	0·59	1	0·32	0·38	0·15	0·21	0·26
Enron	5	0·47	0·54	0·44	0·32	1	0·08	0·11	−0·02	0·10
C.-B.	6	0·10	0·34	0·22	0·38	0·08	1	0·69	0·71	0·62
L.Bros.	7	0·15	0·22	0·23	0·15	0·11	0·69	1	0·77	0·59
ML	8	0·13	0·13	0·11	0·21	−0·02	0·71	0·77	1	0·50
BofA	9	0·09	0·41	0·25	0·26	0·10	0·62	0·59	0·50	1

2·2. *Common correlation model*

In the common correlation model we assume a priori that all correlations r_{ij} are sampled from a common normal distribution subject to $R \in \mathscr{R}^{\mathscr{I}}$:

$$f(R \mid \mu, \sigma^2) = C(\mu, \sigma^2) \prod_{i<j} \exp\{-(r_{ij} - \mu)^2/(2\sigma^2)\} I\{R \in \mathscr{R}^{\mathscr{I}}\}, \tag{1}$$

where

$$C^{-1}(\mu, \sigma^2) = \int_{R \in \mathscr{R}^{\mathscr{I}}} \prod_{i<j} \exp\{-(r_{ij} - \mu)^2/(2\sigma^2)\}\, dr_{ij}, \tag{2}$$

and where $I\{.\}$ represents an indicator function. We assume hyperpriors $\mu \sim N(0, \tau^2)$ and $\sigma^2 \sim \text{IG}(\alpha, \beta)$, where τ^2, α, the shape parameter, and β, the scale parameter, are treated as known. The indicator function in (1) ensures that the correlation matrix is positive definite and introduces dependence among the r_{ij}'s. The implication of the constraint on the conditional prior is not the same for each coefficient. The full conditional posterior density

$f(r_{ij} \mid -)$, which will play a prominent role in the posterior simulation discussed later, is

$$f(r_{ij} \mid -) \propto |R|^{-\frac{1}{2}n} \exp\{-\mathrm{tr}(R^{-1}B)/2\} \exp\{-(r_{ij}-\mu)^2/(2\sigma^2)\} I\{R \in \mathscr{R}^{\mathscr{I}}\}, \tag{3}$$

where B is the empirical variance-covariance matrix. The full conditional densities for μ and σ^2 are similar to the conjugate densities with an additional factor due to the positive definiteness constraint on R:

$$f(\mu \mid -) \propto C(\mu, \sigma^2) \prod_{i<j} \exp\{-(r_{ij}-\mu)^2/(2\sigma^2)\} \exp\{-\mu^2/(2\tau^2)\}, \tag{4}$$

$$f(\sigma \mid -) \propto C(\mu, \sigma^2) \prod_{i<j} \exp\{-(r_{ij}-\mu)^2/(2\sigma^2)\} (1/\sigma^2)^{\alpha-1} \exp(-\beta/\sigma^2). \tag{5}$$

By symmetry the prior is centred at $R=I$ and thus implements shrinkage towards a diagonal correlation matrix. Alternative shrinkage to positive and negative correlations is possible if we choose different prior means for μ.

2·3. *Grouped correlations model*

In many applications the common correlation model is too restrictive. For example, one might have substantive prior information that correlations cluster into groups of positive correlations and negative correlations.

To allow a priori for groups of correlations we generalise the common correlation model to a mixture prior:

$$f(R \mid \mu, \sigma^2, \vartheta) = C(\mu, \sigma^2, \vartheta) \prod_{i<j} \left[\sum_{k=1}^{K} I\{\vartheta_{ij}=k\} \exp\{-(r_{ij}-\mu_k)^2/(2\sigma_k^2)\} \right] I\{R \in \mathscr{R}^{\mathscr{I}}\}, \tag{6}$$

with $\vartheta_{ij} \sim \textsc{mn}(p)$ and $C(\mu, \sigma^2, \vartheta)$ analogous to (2). The indicator $I\{\vartheta_{ij}=k\}$ selects one of the K clusters. To avoid trivial identifiability problems associated with arbitrary permutation of indices, post-processing may be necessary; see for example Celeux et al. (2000) for a discussion of issues related to parameterising mixture models. The full conditional density (3) remains almost unchanged:

$$f(r_{ij} \mid -) \propto |R|^{-\frac{1}{2}n} \exp\{-\mathrm{tr}(R^{-1}B)/2\} \exp\{-(r_{ij}-\mu_{\vartheta_{ij}})^2/(2\sigma^2_{\vartheta_{ij}})\} I\{R \in \mathscr{R}^{\mathscr{I}}\}. \tag{7}$$

As with the common correlation model, the full conditional densities for the μ's and σ^2's are not conjugate and are similar to (4) and (5). The full conditionals for the ϑ_{ij}'s are multinomial distributions which will be dealt with in § 3.

Besides accommodating substantive prior information about clustering of correlations, the mixture prior (6) is also motivated by concerns about the strict normality assumption in (1). In (1), outliers with high correlations could unduly influence final inference. Also, bimodality arising from uncertainty about the direction of a correlation cannot be represented by the single normal prior in (1). For sufficiently large K the mixture model in the grouped correlation prior allows the model to approximate any random effects distribution, subject only to some technical constraints (Dalal & Hall, 1983).

Model (6) implements shrinkage of R towards a structure determined by clustering pairs (i, j) of variables. In many problems this is more appropriate than shrinkage towards a diagonal matrix. Additionally, the introduction of the mixture indicators ϑ_{ij} in the model allows a researcher to represent substantive prior information by choosing unequal prior probabilities $\mathrm{pr}(\vartheta_{ij}=k)$. The posterior distribution under (6) includes inference about grouping the r_{ij} into high and low correlations.

Model (6) includes as a special case model selection for different dependence structures: by including as one term in the mixture a point mass δ_0 at zero, the model allows inference about the presence of marginal dependence of any two variables. Similar mixture models are commonly used for variable selection in regression models (Clyde & George, 2000):

$$f(R \mid \mu, \sigma^2, \vartheta) = C(\mu, \sigma^2, \vartheta) \prod_{i<j} \left[\sum_{k=1}^{K-1} I\{\vartheta_{ij} = k\} \exp\{-(r_{ij} - \mu_k)^2/(2\sigma_k^2)\} \right.$$
$$\left. + I\{\vartheta_{ij} = K\} \delta_0(r_{ij}) \right] I\{R \in \mathscr{R}^{\mathscr{I}}\}.$$

Alternatively, a point mass at zero could be replaced by a small-variance normal distribution as in George & McCulloch (1993).

2·4. *Grouped variables model*

In many applications it is more natural to group the variables, rather than the correlations. For example, as discussed in the motivating example, one might expect a common correlation of returns between bank stocks and a different, common but smaller, correlation between the returns of bank stocks and energy company stocks.

If we group the variables instead of the correlations, the prior (6) changes to

$$f(R \mid \mu, \sigma^2, \vartheta) = C(\mu, \sigma^2, \vartheta) \prod_{i<j} \left[\sum_{k,h} I\{\vartheta_i = k\} I\{\vartheta_j = h\} \right.$$
$$\left. \times \exp\{-(r_{ij} - \mu_{kh})^2/(2\sigma_{kh}^2)\} \right] I\{R \in \mathscr{R}^{\mathscr{I}}\}, \qquad (8)$$

where again $C(\mu, \sigma^2, \vartheta)$ is analogous to (2) and $\vartheta_i \sim \textsc{mn}(p)$. The full conditional posterior density for r_{ij} is as in (7), with μ_k replaced by μ_{kh}:

$$f(r_{ij} \mid \vartheta_i = k, \vartheta_j = h, \ldots) \propto |R|^{-\frac{1}{2}n} \exp\{-\mathrm{tr}(R^{-1}B)/2\} \exp\{-(r_{ij} - \mu_{kh})^2/(2\sigma_{kh}^2)\} I\{R \in \mathscr{R}^{\mathscr{I}}\}.$$

Like (6), model (8) implements shrinkage of R towards a cluster structure, which now is determined by clustering variables; that is clustering is defined on indices i only. For each correlation r_{ij} the pair of indicators $(\vartheta_i, \vartheta_j)$ chooses the term in the prior mixture model. As with model (6), we can explore different dependence structures as a special case of model (8) by including a point mass at zero as a term in the model. This type of model could result in a block diagonal correlation matrix, potentially revealing independence between different groups of variables.

3. Implementation and posterior simulation

3·1. *Sampling the full conditional of* r_{ij}

Without loss of generality we consider only the full conditional (3) for r_{ij} in the common correlation model. The awkward manner in which r_{ij} is embedded in the likelihood complicates posterior simulation, leading us to use a Metropolis–Hastings algorithm to update one coefficient r_{ij} at a time (Barnard et al., 2000); see for example Chib & Greenberg (1995) for a review of the Metropolis–Hastings algorithm. The positive definiteness of R constrains $f(r_{ij} \mid -)$ to an interval (l_{ij}, u_{ij}). Once this interval is found, there are several different proposal densities that could be used, such as the uniform density on that interval, or more generally a Beta density that has been modified to fit the interval, with a mean equal to the current realisation of r_{ij} and a variance that is a fraction of the interval length.

3·2. *Sampling the full conditionals of μ and σ^2*

We focus our discussion on sampling from the full conditional density for μ based on the common correlation model, see (4). Extensions of these strategies to σ^2 for the common correlation model and to μ and σ^2 for the grouped models can be derived in a natural way. Since μ is hopelessly entangled in the normalising constant C, we again use a Metropolis–Hastings step to update μ. The proposal density is the normal distribution that results if $I\{R \in \mathscr{R}^{\mathscr{I}}\}$ is removed from (1). If μ^* denotes the generated proposal, the appropriate acceptance probability is

$$\alpha = \min\{1, C(\mu^*, \sigma^2)/C(\mu, \sigma^2)\}, \tag{9}$$

where C is given by (2). We propose several alternative strategies for evaluating α. The simplest approach is to assume $\alpha = 1$. This is the strategy that Daniels & Kass (1999) use when analysing their model which uses a Fisher z transformation of the correlations. While this may be reasonable when μ is close to μ^* or σ^2 is small, these conditions may not hold in practice.

Since C is not data-dependent, one may evaluate the value of $C(\mu, \sigma^2)$ at the outset, for a range of values for μ and σ^2 and then use an interpolation strategy to evaluate α. While this approach has the advantage of only being dependent on the dimension of the problem, we chose to focus on strategies which estimate C as needed.

The normalising constant C is proportional to the integral of a product of univariate normal densities restricted to a constrained space and can be estimated using an importance sampling scheme; see Chen et al. (2000, Ch. 5), for a general discussion of using importance sampling to estimate the ratio of multivariate integrals, as in (9). One strategy is to sample from unconstrained normal densities, $r_{ij}^m \sim N(\mu, \sigma^2)$, for $i < j$ and $m = 1, \ldots, M$, define $R^m = (r_{ij}^m)$ and use

$$\hat{C}(\mu, \sigma^2) = 1/M \sum_{m=1}^{M} I\{R^m \in \mathscr{R}^{\mathscr{I}}\}.$$

Another strategy is to modify the original model by inserting an additional layer of priors. We introduce latent variables δ_{ij} into the model hierarchy between r_{ij} and the prior moments (μ, σ^2) by assuming that

$$\delta_{ij} \sim N(\mu, \sigma^2), \tag{10}$$

and we replace the prior (1) by

$$f(R \mid \delta) = C(\delta, v^2) \prod_{i<j} \exp\{-(r_{ij} - \delta_{ij})^2/(2v^2)\} I\{R \in \mathscr{R}^{\mathscr{I}}\},$$

where C^{-1} is similar to (2) with μ replaced by δ_{ij} and σ^2 replaced by v^2. We refer to (10) as a 'shadow prior'. The resulting full conditional density for r_{ij} changes only slightly, with (δ_{ij}, v^2) replacing (μ, σ^2), but the full conditional densities for μ and σ^2 are now conjugate. The nature of the full conditional density for δ_{ij} is similar to the full conditional (3) in the original model, requiring a Metropolis–Hastings step to update the δ_{ij}. However, there are important advantages to using the additional parameters and model structure. First, as the researcher has complete control over the value of v^2, it can be set to an arbitrary value. As v^2 approaches zero, the ratio in (9) approaches one. In practice, it is often reasonable to set v^2 to a small number and assume that $C(\delta^*, v^2)/C(\delta, v^2) = 1$. Intuitively, if we set v^2 small enough, the full conditional density of R essentially lies inside the constrained space $\mathscr{R}^{\mathscr{I}}$, which allows the unconstrained normalising constant to be a reasonably good approximation for the constrained normalising constant $C(\delta, v^2)$. A second important advantage of introducing the shadow prior is the simplification of the computational burden associated with sampling the indicator variables in the two grouped models; see § 3·3. Although the shadow prior offers a general way of dealing with constraints, the full extent of its effectiveness and limitations are not explored in this paper.

One concern with setting v^2 to a small number is that it may affect the mixing properties of the Markov chain Monte Carlo algorithm with respect to μ and σ^2. In practice, we have found that setting v^2 to a small value does not have significant impact on the mixing properties of these variables in terms of the autocorrelation and the marginal posterior density of the parameters. For example, a small simulation study, based on a common correlation model and not described here, showed that for small values of v^2 it is reasonable to assume that $C(\delta^*, v^2)/C(\delta, v^2)=1$.

With regards to the performance of the Markov chain Monte Carlo analysis for different values of v^2, we found that the mixing properties, as summarised by the autocorrelation of μ and σ^2, and the posterior inference, as summarised by the marginal posterior density estimates of μ and σ^2, were almost identical across the range of v^2 considered. In practice this type of analysis could be used as a guide for calibrating v^2 such that the ratio of the normalising constants is approximately equal to one, and performance of the Markov chain Monte Carlo algorithm is not significantly affected.

Perhaps the most important benefit from using the shadow prior is a critical simplification in the full conditional for the indicators ϑ_i and ϑ_{ij} in the grouped variable and grouped correlation models, respectively. We discuss details in the next section.

3·3. *Sampling the full conditional of* ϑ

Without the shadow priors, the full conditional densities for ϑ for the grouped correlations model and the grouped variables model are as follows:

$$f(\vartheta_{ij}=k\,|\,-)=C(\mu, \sigma^2, \vartheta_{ij}=k, \vartheta_{-ij}) \exp\{-(r_{ij}-\mu_k)^2/(2\sigma_k^2)\}, \tag{11}$$

$$f(\vartheta_i=k\,|\,-)=C(\mu, \sigma^2, \vartheta_i=k, \vartheta_{-i}) \prod_{j\neq i} \exp\{-(r_{ij}-\mu_{k,\vartheta_j})^2/(2\sigma^2_{k,\vartheta_j})\}. \tag{12}$$

Evaluating these full conditional densities requires us to calculate the normalising constant C, which can be done using importance sampling strategies, as discussed previously. As the dimensionality of the correlation matrix increases, the computational task of evaluating (11) and (12) can make it difficult if not impossible to analyse these models in practice.

By introducing the shadow prior we have the good fortune that evaluation of (11) and (12) simplifies significantly. With the shadow prior included in each model, the multivariate integral C no longer involves the variables ϑ, and as a result the full conditional densities (11) and (12) become

$$f(\vartheta_{ij}=k\,|\,-) \propto \exp\{-(\delta_{ij}-\mu_k)^2/(2\sigma_k^2)\},$$

$$f(\vartheta_i=k\,|\,-) \propto \prod_{j\neq i} \exp\{-(\delta_{ij}-\mu_{k,\vartheta_j})^2/(2\sigma^2_{k,\vartheta_j})\}.$$

The full conditional densities for ϑ, using the shadow prior, are therefore easily evaluated, even for high-dimensional problems.

4. Extensions and applications

4·1. *Random effects distributions in hierarchical models*

The discussion and the proposed implementation of posterior simulation remain valid for more complex probability models in which the multivariate normal distribution defines only one component or level of the probability model, as in hierarchical models with multivariate normal random effects, $\theta_i \sim N(\mu, \Sigma)$ (Daniels & Kass, 1999). Here θ_i is a random effects vector specific to the ith experimental unit in the hierarchical model, for an observable y_i, and the proposed prior models would be used to define the hyperprior for the covariance matrix Σ of the random effects.

Example 2. Müller & Rosner (1997) describe a haematological study. The data record white blood cell counts over time for each of n chemotherapy patients. Denote by y_{it} the measured response for patient i on day t. The profiles of white blood cell counts over time can be reasonably well approximated by a piecewise linear-linear-logistic regression model, involving a patient-specific random effects vector θ_i and represented by $y_{it} = g_{\theta_i}(t) + \varepsilon_{it}$. The model is completed with a random effects model,

$$\theta_i \sim N(m, \Sigma), \tag{13}$$

and a hyperprior $h(m, \Sigma)$. Posterior predictive inference for future patients depends on the observed historical data only indirectly through learning about the random effects distribution (13). Thus a flexible hyperprior for Σ is essential. The grouped correlations model provides a possible hyperprior on Σ, as an alternative to Müller & Rosner's (1997) use of a flexible nonparametric model in place of (13).

4·2. *The* ARCH *models and time-varying correlations*

An autoregressive conditional heteroscedasticity, ARCH, model (Engle, 1982) is a discrete stochastic process for which the variance at time t is related to previous squared values of the process in an autoregressive scheme:

$$\varepsilon_t \sim N(0, h_t), \quad h_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2. \tag{14}$$

The GARCH(p, q), generalised ARCH, model (Bollerslev, 1986) includes lagged values of the variance itself in the variance equation, so that

$$h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \gamma_j h_{t-j};$$

see for example Bollerslev et al. (1992) for a survey of related models.

A multivariate version of the GARCH(p, q) model was first studied by Bollerslev et al. (1988). Let H_t denote the covariance matrix in a multivariate version of (14), such that $\varepsilon_t \sim N(0, H_t)$. The high-dimensional nature of H_t complicates modelling, and in practice the dimensionality is greatly reduced by imposing additional structural assumptions. For example, Bollerslev (1990) uses a time-varying conditional covariance matrix, H_t, but assumes time-invariant conditional correlations. Our framework can model multivariate time-series data by assuming a univariate (G)ARCH model for the marginal variances $h_{it} = (H_t)_{ii}$ and completing the model with a structured prior for the correlation matrix, using one of the models proposed in § 2. Let S_t denote a diagonal matrix with the standard deviations on the diagonal, and assume $H_t = S_t R S_t$ with equations (1), (6) or (8) as prior models for a time invariant correlation matrix R.

Not only can this framework model groupings of correlations among multivariate time-series data, but it can also jointly model time-varying variances and time-varying correlation structures. By shrinking correlations towards a set of common means, standard time-varying probabilistic models can be used to model the dynamic nature of these means, thereby offering a parsimonious way of modelling the time-varying correlation structure.

4·3. *Probit models*

The multivariate probit model implements regression of a set of binary response variables $y=(y_1, \ldots, y_p)$ on covariates $x=(x_1, \ldots, x_p)$. We introduce a p-dimensional normal latent variable vector y^* and assume that

$$y_i = I\{y_i^* > 0\}, \quad y^* \sim N(\beta' x, \Sigma). \tag{15}$$

Albert & Chib (1995) discuss the corresponding univariate probit model. In the multivariate model (15), the covariance matrix Σ defines the covariate effects. For identifiability Σ needs to be suitably constrained, for example as a correlation matrix. Models (1), (6) and (8) provide flexible prior models. Alternative Bayesian models for the identified parameters of a multivariate probit are discussed by Chib & Greenberg (1998) and McCulloch et al. (2000). Müller et al. (1999) discuss a parameterisation which allows conjugate Gibbs sampling.

Example 3. Liechty et al. (2001) use the multivariate probit model to estimate an empirical demand function for an information service. Potential customers were shown a collection of scenarios whereby they could purchase enhancements/services to a basic free listing. Each enhancement had a monthly fee and some had a one-time set-up fee. The levels of the fees along with several group discounting schemes were varied over different scenarios. Potential customers indicated whether or not they would choose each of the possible enhancements, resulting in a set of vectors of binary responses. Prices and discounting schemes were used as covariates, and the estimated correlation matrix revealed that the products had relationships net of price and group discount effects. One interpretation of this was that products with positive correlations were complements and products with negative correlations were substitutes. A two-group, grouped correlation model, would be a natural candidate for identifying complements and substitutes for this type of dataset.

The multivariate probit model is also often used to model the content of a household's shopping basket over time; see Manchanda et al. (1999). The resulting correlation matrix can become very large, but it would be natural to consider a correlation structure which groups correlations between products based on product categories using a grouped variables model. In this context, one can look for zeros in the correlation matrix, in much the same way as Wong et al. (2003) look for zeros in the precision matrix, so as to identify block diagonal matrices that define how consumers partition products into different types of market.

5. Examples

5·1. *Simulated data*

We generated an 8 × 8 correlation/covariance matrix with two different groups of correlations, using (6) with $\mu_1 = -0{\cdot}54$, $\sigma_1^2 = 0{\cdot}05$, $\mu_2 = 0{\cdot}37$ and $\sigma_2^2 = 0{\cdot}08$. This model has $(8 \times 7)/2 = 28$ correlations; 14 correlations are in each group, and can be thought of as being in a negative, group 1, or positive, group 2, group. Based on this matrix we generated

500 observations. The true group of each correlation and the posterior probabilities are shown in Table 2. The posterior estimates recover the true structure of the correlation model in the simulation model.

Additional studies with group means closer together and various values for group variances yield similar results. When the residual variances σ^2 are large, the posterior probabilities are more spread out across groups, as would be expected for any mixture with overlapping distributions. This suggests that inference about K should be interpreted with caution, unless the clusters are well separated. We found similar results when simulating from the grouped variables model. To investigate the performance of the approach in the absence of true cluster structure in the data, we considered the special case with $K = 1$ in the simulation model. Posterior inference with $K > 1$ finds means close enough for us to conclude that there is only one group.

Table 2. *Simulation study with* 8×8 *correlation matrix consisting of two groups of correlations and* 500 *observations. The* $(8 \times 7)/2 = 28$ *correlations are evenly divided between two groups that have parameters* $\mu_1 = -0{\cdot}54$, $\sigma_1^2 = 0{\cdot}05$, $\mu_2 = 0{\cdot}37$ *and* $\sigma_2^2 = 0{\cdot}08$. *The upper triangle shows the group* ϑ_0 *in which each of the correlations lies, and the lower triangle shows the posterior probability* $\hat{\vartheta}$ *that the correlations are from Group* 1

				ϑ_0				
$\hat{\vartheta}$	1	2	3	4	5	6	7	8
1	0	1	1	1	2	2	2	2
2	0·99	0	2	1	2	1	2	1
3	0·99	0·00	0	2	2	2	1	2
4	0·01	0·99	0·00	0	1	2	1	2
5	0·01	0·00	0·01	0·99	0	1	1	1
6	0·01	0·98	0·00	0·00	0·98	0	1	2
7	0·00	0·00	0·99	0·98	0·02	0·97	0	1
8	0·01	0·99	0·00	0·00	0·99	0·00	0·99	0

5·2. *Structure in the stock market*

As discussed in our motivating example, Enron was attempting to change from being an energy company to a finance company. As it may be difficult to classify Enron based on the range of products and services that they offer, it is possible to use the grouped variables model to determine whether to group Enron with energy stocks or finance stocks, based on the correlations between each group of stocks. Using the stocks introduced in § 2·1 we consider grouped variable models (8) with $K = 1$, which corresponds to the common correlation model (1), for $K = 2$ and $K = 3$, and use a reversible jump Markov chain Monte Carlo algorithm to infer the value of K. We also include a uniform prior as in Barnard et al. (2000) for comparison. With a uniform prior on the model space, the reversible jump Markov chain Monte Carlo method reduces to comparing the likelihoods of the current model and the proposed model at each step. This can be justified as follows. Let θ_k denote the parameter vector under $K = k$, including $k = 0$ for the model with the uniform prior. Let $K \in \{1, 2, 3, 0\}$ denote the index of the currently chosen model. We build a superparameter vector $\theta = (\theta_0, \ldots, \theta_3)$, and define a Markov chain Monte Carlo scheme that updates θ_k only when $K=k$. Proposing a move from K to $\tilde{K}$ leads to a Metropolis–Hastings acceptance probability that includes the ratio of the likelihood evaluated under the two competing models K and $\tilde{K}$, using the currently imputed parameter vectors θ_K and $\theta_{\tilde{K}}$. As a result of concerns with convergence properties we consider the results as approximate posterior inference only and advise against over-interpretation. We find approximate posterior probabilities of 0·06 for $K=1$, 0·5 for $K=2$, 0·25 for $K=3$, and 0·19 for the uniform prior. For the model with $K=2$, the posterior parameter estimates of (μ_{12}, μ_1, μ_2) are $(0{\cdot}15, 0{\cdot}41, 0{\cdot}61)$ and for $(\sigma_{12}^2, \sigma_1^2, \sigma_2^2)$ are $(0{\cdot}02, 0{\cdot}03, 0{\cdot}02)$. We see a distinct separation between the three group means.

Table 3 shows that Enron is clearly grouped with the energy companies, and was unsuccessful, in terms of stock performance, in making the transition from being an energy company to a finance company. The posterior estimate of the correlation matrix for these variables is also shown in Table 3.

It is interesting to note that classical factor models based on either the maximum likelihood method or the principal components method indicated five factors for these data. Loadings for both methods are very similar, but they offer no concise 'grouping' of variables as is found by the grouped variables model.

Table 3. *Posterior estimates of* $\mathrm{pr}(\vartheta_i \mid y)$, *the probability that variable i is in a particular group, and of R, the correlation matrix, in grouped variables model with two groups for monthly stock returns for nine equity securities from April* 1996 *to May* 2002. *Reliant, Chevron, British Petroleum and Exxon are the energy companies, and Citi-Bank, Lehman Brothers, Merrill Lynch and Bank of America are the financial services companies. According to this model, Enron is classified as an energy stock*

Variable		$\mathrm{pr}(1\mid y)$	$\mathrm{pr}(2\mid y)$	1	2	3	4	5	6	7	8	9
Reliant	1	0·99	0·01	1	0·32	0·18	0·33	0·43	0·13	0·16	0·16	0·11
Chev.	2	1	0	0·32	1	0·66	0·5	0·49	0·26	0·17	0·1	0·32
BP	3	1	0	0·18	0·66	1	0·51	0·4	0·13	0·2	0·08	0·16
Exxon	4	1	0	0·33	0·5	0·51	1	0·29	0·3	0·09	0·17	0·16
Enron	5	0·99	0·01	0·43	0·49	0·4	0·29	1	0·11	0·15	0·04	0·11
C.-B.	6	0·01	0·99	0·13	0·26	0·13	0·3	0·11	1	0·63	0·66	0·62
L.Bros.	7	0	1	0·16	0·17	0·2	0·09	0·15	0·63	1	0·72	0·56
ML	8	0	1	0·16	0·1	0·08	0·17	0·04	0·66	0·72	1	0·46
BofA	9	0·02	0·98	0·11	0·32	0·16	0·16	0·11	0·62	0·56	0·46	1

5·3. *Population genetics*

Murren et al. (2002) consider a dataset recording measurement of $J=11$ traits over a population of $n=40$ brassica plants, i.e. broccoli. One question of interest is how the different traits can be clustered into groups on the basis of correlation structure. Variables are a priori expected to cluster into groups of traits related to some common underlying themes. For example, possible groups might be variables related to 'life history', 'plant size' and so on.

We modelled the correlation matrix using a two-group, grouped variables model. The posterior probability that each variable is assigned to Group 1 is summarised in Table 4. The three variables associated with the size of the leaves are classified into Group 2; the remaining variables, placed in Group 1, have to do with the size of the other parts of the plant and with the number, but not the size, of leaves. The average correlation for the leaf-size group, Group 2, is 0·19, while the average correlation for the plant-size group,

Table 4. *Posterior estimates of* $\mathrm{pr}(\vartheta_i \mid y)$, *the probability that variable i is in a particular group, in a grouped variables model with two groups of* 11 *variables on* 40 *broccoli plants*

Variable	$\mathrm{pr}(1 \mid y)$	$\mathrm{pr}(2 \mid y)$	Variable	$\mathrm{pr}(1 \mid y)$	$\mathrm{pr}(2 \mid y)$
Days to leaf	1	0	Petiole length	0	1
Days to flower	1	0	Leaf number	1	0
Days to harvest	1	0	Height	1	0
Leaf length	0	1	Leaf biomass	1	0
Fruit number	1	0	Stem biomass	1	0
Leaf width	0	1			

Group 1, is 0·25. Interestingly, the average correlation between the two groups is negative, −0·31, which seems to indicate that the plant either emphasises leaf growth or plant growth, but not both.

6. Discussion

Our models provide a framework for representing and learning about dependence structure. An alternative approach is traditional factor analysis. Factor analysis explains dependence among a set of variables as arising from a few latent factors and the relative sizes of the factor loadings determine the strengths of the correlations. In contrast, the grouped variables and grouped correlations models proceed by assuming a partition of the set of correlations and variables, respectively. The extent of the relationship between these two modelling approaches is an interesting topic for future research.

The discussion has focused on inference on covariance and correlation matrices. However, since inference is based on posterior simulation, inference about any other function of the model parameters is possible. For example, in graphical models, inference about the precision matrix $P = R^{-1}$ is of interest. Small off-diagonal entries in the precision matrix correspond to small conditional correlation of the respective variables. This allows inference about the presence and absence of connecting edges in a graphical representation of a multivariate distribution.

We have introduced the models in the context of independent multivariate normal sampling, but the models lead to interesting generalisations of standard models in any modelling context which involves (hyper)prior probability models on variance-covariance or precision matrices. Examples are repeated measurement models, hierarchical models, multivariate probit models, graphical models and multivariate stochastic volatility models.

Acknowledgement

The authors are grateful to Robert McCulloch and Chris Carter for insightful discussion and comments.

References

Aguilar, O. & West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *J. Bus. Econ. Statist.* **18**, 338–57.

Albert, J. & Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika* **82**, 747–59.

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

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory*. New York: John Wiley & Sons, Inc.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Economet.* **31**, 307–27.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH approach. *Rev. Econ. Statist.* **72**, 498–505.

Bollerslev, T., Engle, R. F. & Wooldridge, M. (1988). A capital asset pricing model with time varying covariances. *Polit. Econ.* **96**, 116–31.

Bollerslev, T., Chou, R. & Kroner, K. F. (1992). ARCH modeling in finance. *J. Economet.* **52**, 5–59.

Celeux, G., Hurn, M. & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Am. Statist. Assoc.* **95**, 957–70.

Chen, M., Ibrahim, J. G. & Shao, Q. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Chib, S. & Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *Am. Statistician* **49**, 327–35.

Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–61.

Christiansen, C. & Morris, C. (1997). Hierarchical Poisson regression modeling. *J. Am. Statist. Assoc.* **92**, 618–32.

Clyde, M. & George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc.* B **62**, 681–98.

Dalal, S. & Hall, W. (1983). Approximating priors by mixtures of natural conjugate priors. *J. R. Statist. Soc.* B **45**, 278–86.

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *Can. J. Statist.* **27**, 567–78.

Daniels, M. J. & Kass, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J. Am. Statist. Assoc.* **94**, 1254–63.

Daniels, M. J. & Kass, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57**, 1174–84.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**, 987–1008.

Everson, P. J. & Morris, C. N. (2000). Inference for multivariate normal hierarchical models. *J. R. Statist. Soc.* B **62**, 399–412.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Statist. Assoc.* **88**, 881–9.

Karolyi, G. A. (1992). Predicting risk: some new generalizations. *Manag. Sci.* **38**, 57–74.

Karolyi, G. A. (1993). A Bayesian approach to modeling stock return volatility for option valuation. *J. Finan. Quantitat. Anal.* **28**, 579–94.

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

Liechty, J. C., Ramaswamy, V. & Cohen, S. H. (2001). Choice menus for mass customization: An experimental approach for analyzing customer demand with an application to a web-based information service. *J. Market. Res.* **38**, 183–96.

Lin, S. P. & Perlman, M. D. (1985). An improved procedure for the estimation of a correlation matrix. In *Statistical Theory and Data Analysis*, Ed. K. Matusita, pp. 369–79. North Holland: Elsevier.

Manchanda, P., Ansari, A. & Gupta, S. (1999). The 'shopping basket': A model for multi-category purchase incidence decisions. *Market. Sci.* **18**, 95–114.

McCulloch, R. E., Polson, N. G. & Rossi, P. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *J. Economet.* **99**, 173–93.

Müller, P. & Rosner, G. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *J. Am. Statist. Assoc.* **92**, 1279–92.

Müller, P., Parmigiani, G., Schildkraut, J. & Tardella, L. (1999). A Bayesian hierarchical approach for combining case-control and prospective studies. *Biometrics* **55**, 258–66.

Murren, C. J., Pendelton, N. & Pigliucci, M. (2002). Evolution of phenotypic integration in Brassica (Brassicaceae). *Am. J. Botany* **89**, 655–63.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with Discussion). *Ann. Statist.* **22**, 1701–62.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics* 7, Ed. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, pp. 723–32. Oxford University Press.

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

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

[*Received August* 2002. *Revised March* 2003]