Statistica Sinica **10**(2000), 1281-1311

MODELING COVARIANCE MATRICES IN TERMS OF STANDARD DEVIATIONS AND CORRELATIONS, WITH APPLICATION TO SHRINKAGE

John Barnard, Robert McCulloch* and Xiao-Li Meng*

*Harvard University and *The University of Chicago*

Abstract: The covariance matrix plays an important role in statistical inference, yet modeling a covariance matrix is often a difficult task in practice due to its dimensionality and the non-negative definite constraint. In order to model a covariance matrix effectively, it is typically broken down into components based on modeling considerations or mathematical convenience. Decompositions that have received recent research attention include variance components, spectral decomposition, Cholesky decomposition, and matrix logarithm. In this paper we study a statistically motivated decomposition which appears to be relatively unexplored for the purpose of modeling. We model a covariance matrix in terms of its corresponding standard deviations and correlation matrix. We discuss two general modeling situations where this approach is useful: shrinkage estimation of regression coefficients, and a general location-scale model for both categorical and continuous variables. We present some simple choices for priors in terms of standard deviations and the correlation matrix, and describe a straightforward computational strategy for obtaining the posterior of the covariance matrix. We apply our method to real and simulated data sets in the context of shrinkage estimation.

Key words and phrases: General location model, general location-scale model, Gibbs sampler, hierarchical models, Markov chain Monte Carlo, Wishart distribution.

1. A Separation Strategy for Modeling Covariance Matrices

Modeling a variance-covariance structure is one of the most common and important tasks in statistical analysis. It is also one of the most difficult. A covariance matrix may have many parameters, and these parameters are constrained by the complex requirement that the matrix be non-negative definite. In this paper we investigate a simple strategy that attempts to deal with these problems in some applications. Although our focus is on Bayesian analysis, our strategy is equally applicable for non-Bayesian modeling; in Section 4, we give one such application, which involves extensions to the general location model. Our strategy includes a simple method for computing the posterior of a covariance matrix using the Gibbs sampler.

Because a covariance matrix is complicated, it is helpful to start by breaking it down into components. For example, a dependence structure may be represented in terms of variance components. There are also several methods based on well-known matrix decompositions. For instance, Banfield and Raftery (1993), Yang and Berger (1994), Celeux and Govaert (1995), and Bensmail, Celeux, Raftery and Robert (1997) work with the spectral decomposition of the matrix. In transforming to the matrix logarithm, Leonard and Hsu (1992) and Chiu, Leonard and Tsui (1996) essentially start from the spectral decomposition. Another approach is to use the Cholesky decomposition of the inverse of the covariance matrix (e.g., Pourahmadi (1999, 2000)), which has a nice regression interpretation. Liu (1993) uses the same type of Cholesky decomposition to obtain a Bartlett-type decomposition of the posterior distribution of a covariance matrix with monotone missing data. There is also a literature on using the Cholesky decomposition directly for the covariance matrix (e.g., Pinheiro and Bates (1996)), though the resulting parameterizations do not have simple statistical interpretation.

For some applications we have encountered (e.g., shrinkage modeling), we found it is desirable to directly work with standard deviations and correlation matrix, which do not correspond to any parameterization from the aforementioned decompositions. The purpose of this paper is thus to study this direct decomposition. Specifically, we write

$$\Sigma = \operatorname{diag}(S) R \operatorname{diag}(S), \tag{1}$$

where S is the $k \times 1$ vector of standard deviations, $\operatorname{diag}(S)$ is the diagonal matrix with diagonal elements S, and R is the $k \times k$ correlation matrix. We refer to this decomposition as a *separation strategy*, as we separate out the standard deviations and correlations. Clearly, this separation has a strong practical motivation – most practitioners are trained to think in terms of standard deviations and correlations; the standard deviations are on the original scale, and the correlations are scale free.

Consider the problem of prior specification. Directly specifying a reasonable prior for a covariance matrix is a difficult task. The usual inverse-Wishart prior is often inadequate because of its restrictive form (e.g., common degrees of freedom for all components of S). It is sometimes the case that we are willing to express prior beliefs about S, the vector of standard deviations, but less willing about R, the correlation matrix. We argue in Section 3.1 that this is often the case in the important application of shrinkage estimation of regression coefficients. In such cases we can write our prior on Σ in terms of (S, R) in the form

$$p(S, R) = p(S)p(R \mid S). \tag{2}$$

We want to incorporate our prior information in $p(S)$ but wish to choose $p(R \mid S)$ in a manner that is convenient and "diffuse". This is in the same spirit as in Sun and Berger (1998), who consider the more general situation where one writes $p(\theta_1, \theta_2) = p(\theta_1)p(\theta_2 \mid \theta_1)$. Prior information is then used to specify $p(\theta_1)$ and a reference prior is chosen for θ_2 conditional on θ_1. In Section 2, we discuss some simple choices for $p(S)$ and for $p(R \mid S)$, which have nice properties from both a practical and theoretical point of view; empirical evidence is provided in Section 3 in the context of shrinkage estimation. Obviously, decomposition (2) also allows the use of "non-informative" priors for S, if such a choice is desired (e.g., a constant prior for $\log S$).

In some cases, we may have a set of covariance matrices to deal with rather than just one, such as with the general location model (e.g., Olkin and Tate (1961)). Rather than just letting the matrices vary freely, we may want to model them somehow (e.g., in order to reduce the number of parameters). The separation strategy suggests a modeling approach. For $\Sigma_i = \text{diag}(S_i)R_i\text{diag}(S_i)$, we assume that R_i's do not vary with i, and model the vector S_i's as depending on explanatory variables. In Section 4 we use this approach to propose a generalization of the general location model (GLOM), which relaxes the restrictive common-covariance assumption of GLOM without bringing an unmanageable number of parameters into the model.

In Section 5 we show that the separation strategy also leads to a simple computational strategy for obtaining the posterior distribution of S and R. We simply draw each component of both S and R one at time in a Gibbs sampler. The key here is that it is straightforward to find the set of values for one correlation given the others that preserves the positive-definiteness of the correlation matrix. Because we know the Jacobian of the transformation from Σ to (S, R), this strategy can also (though not always) be useful for computing the posterior of Σ when the prior is specified directly on Σ, albeit it may not be the most efficient method for particular problems. Our limited experience suggests that this simple strategy is promising, even in high dimensional situations, for problems likely to occur in practice (e.g., when the posterior for Σ is reasonably smooth).

2. Some Prior Models for Covariance Matrices

2.1. Basic assumptions

Following the separation strategy, we specify a prior on Σ by choosing $p(S)$ and $p(R \mid S)$. Since S is simply a k-dimensional vector with component-wise non-negativity as the only constraint, there are many multivariate distributions that can be potentially used for $p(S)$. In the applications in this paper we use

$$\log(S) \sim N(\xi, \Lambda), \tag{3}$$

where $\log(S) \equiv (\log(s_1), \ldots, \log(s_k))^\top$. Typically, the matrix Λ will be chosen to be diagonal, that is, we are choosing independent log normal distributions for each of the standard deviations. An obvious alternative would be to choose independent scaled inverted chi-squared distributions for each of the variances, as this is the commonly used conjugate prior for a variance. In the real application discussed in Section 3.3 we found the log normal prior more appealing, as it was more difficult to deal with the tail behavior of the inverted chi-squared with low degrees of freedom. Note that if we use the usual prior for covariance matrices, the inverse-Wishart, all the diagonal elements share the same degrees of freedom parameter, making it impossible to separately assess marginals for the diagonal elements. The flexibility in dealing with tails of individual components is a key practical advantage of the separation strategy.

The choice of prior for R given S is more complicated, due to the complexity of the space of correlation matrices. We also know that, particularly in high dimensional problems, priors are never really "non-informative", so some care is needed. Before we discuss the specific priors used in this paper, we want to lay out some basic assumptions that underline our choices.

First, we are willing to assume that S and R are independent. In our shrinkage application, it is difficult, if not impossible, to solicit reliable prior information on how correlations depend on the variances. Our experience and beliefs are that, in the absence of reliable knowledge of the dependence structure, it is less harmful to adopt the assumption of independence, and, thus, gain flexibility in dealing with S and R separately, than to blindly use common models such as the inverse-Wishart distribution. Of course, if an application calls for dependence of R on S, then it should be modeled – the separation strategy allows for explicit modeling of such dependence through the specification of $p(R \mid S)$.

Second, we assume we are in a situation where we do not have much *a priori* information to distinguish among $\{r_{ij},\ i \neq j\}$, thus *a priori* the prior distribution for $\{r_{ij},\ i \neq j\}$ is invariant to permutations of indexes, that is, $\{r_{ij},\ i \neq j\}$ are *a priori* exchangeable. Note that inherent in this assumption is that k is not too large with respect to the amount of information we have; if k is (relatively) large, then it is often the case that *a priori* we can group the underlying variables into several groups and then assume exchangeability within each group and independence between groups. What we discuss in this paper can be straightforwardly extended to handle multiple groups, but for simplicity of presentation we will focus on the single group case.

Finally, we also intend to choose priors that are "diffuse" in some sense to reflect our weak knowledge about R. We do not attempt to specify a realistic prior distribution for the the correlation matrix, but rather, give a few default choices for R. In the shrinkage application of Section 3, a diffuse prior for R also

seems desirable because it reduces possible conflicts between *a priori* restrictions on R (e.g., high correlations) and our informative prior for S, since the amount of shrinkage is not determined by S alone. The next two subsections investigate two choices for $p(R)$, which possibly represent two extreme modeling strategies that a "diffuse modeler" is likely to adopt, and thus they are also useful, as a pair, for sensitivity studies to the specification of "default" priors for R. An alternative to our two strategies is to use uniform prior distributions on the partial correlations, which has been shown to be useful for informative modeling, e.g., see Ramsey (1974) and Le, Martin and Raftery (1996). We have not explored this option, but suspect that it falls between our two modeling extremes.

2.2. A marginally uniform prior

Given each r_{ij} is between $[-1, 1]$ and our desire to use "diffuse" priors, it is natural to seek a joint distribution on the correlation matrix space $\mathcal{R}^k$, which consists of all $k \times k$ correlation matrices, such that all the implied marginal densities for $r_{ij} (i \neq j)$ are uniform. Such a distribution can be obtained from the commonly used inverse-Wishart distribution for Σ. To see this, we first derive the marginal distribution of R when Σ has a standard inverse-Wishart distribution, $W_k^{-1}(I, \nu)$, $\nu \geq k$, that is, when Σ has a density function

$$f_k(\Sigma \mid \nu) \propto |\Sigma|^{-\frac{1}{2}(\nu+k+1)} \exp(-\frac{1}{2} \operatorname{tr}\{\Sigma^{-1}\}). \tag{4}$$

Under the transformation $\Sigma \to (S, R)$, the Jacobian is given by $2^k(\prod s_i)^k$, thus,

$$f_k(R \mid \nu) \propto |R|^{-\frac{1}{2}(\nu+k+1)} \prod_i \int_0^\infty s_i^{-(\nu+1)} \exp(-\frac{r^{ii}}{2s_i^2}) ds_i, \tag{5}$$

where r^{ii} is the ith diagonal element of R^{-1}. Now let $\xi_i = r^{ii}/(2s_i^2)$, then

$$f_k(R \mid \nu) \propto |R|^{-\frac{1}{2}(\nu+k+1)} (\prod_i r^{ii})^{-\frac{\nu}{2}} \prod_i \int_0^\infty \xi^{(\nu-2)/2} \exp(-\xi) d\xi.$$

Thus,

$$f_k(R \mid \nu) \propto |R|^{-\frac{1}{2}(\nu+k+1)} (\prod_i r^{ii})^{-\frac{\nu}{2}} = |R|^{\frac{1}{2}(\nu-1)(k-1)-1} (\prod_i |R_{ii}|)^{-\frac{\nu}{2}}, \tag{6}$$

where the equality is due to the fact $r^{ii} = |R_{ii}|/|R|$, with R_{ii} being the ith principal sub-matrix of R.

Thanks to the marginalization property of the inverse-Wishart (i.e., a principal sub-matrix of an inverse-Wishart is still an inverse-Wishart), we can easily derive the marginal distributions of (6). For any $k_1 \times k_1$ sub-covariance matrix

Σ_1 of Σ, its distribution is $W_{k_1}^{-1}(I_1, \nu-(k-k_1))$, where I_1 is the $k_1 \times k_1$ identity matrix. Thus, the marginal density of the corresponding sub-correlation matrix R_1 is obtained when we replace k in (6) by k_1 and *at the same time* ν by $\nu_1 = \nu - (k - k_1)$. In particular, taking $k_1 = 2$, we obtain that the marginal distribution of $r_{ij}(i \neq j)$ as

$$f_2(r_{ij}|\nu) \propto (1-r_{ij}^2)^{\frac{\nu-k-1}{2}}, \qquad |r_{ij}| \leq 1, \tag{7}$$

which is the same result as given in Box and Tiao (1973, Section 8.5.4) but with different notation because their ν is our $\nu - k + 1$; also see Jeffreys (1983) and Tan (1969). Density (7) can be viewed as a $Beta(\frac{\nu-k+1}{2}, \frac{\nu-k+1}{2})$ on [-1, 1], and is uniform when $\nu = k + 1$. In other words, if we take

$$f_k(R \mid \nu = k+1) \propto |R|^{\frac{k(k-1)}{2}-1}(\prod_i R_{ii})^{-\frac{(k+1)}{2}}, \tag{8}$$

then the marginal distributions for all the individual correlations are uniform. In addition, by choosing $k \leq \nu < k+1$ or $\nu > k+1$, we can control the tail behavior of $f_2(r_{ij} \mid \nu)$, that is, whether we want it be heavier or lighter than the uniform. Thus, we have a family of priors for R indexed by a single "tuning" parameter ν. Since we will use an independent prior for S, as detailed in Section 2.1, ν only controls the tails of the distributions of the $r'_{ij}s$, in contrast to the use of the inverse-Wishart for Σ, where it controls the tails of all components of S as well. It is also possible, as with hierarchical modeling, to estimate ν from the data, or to use posterior predictive checks (Rubin (1984), Gelman and Meng (1996), Gelman, Meng and Stern (1996)) to see if the model resulting from a specific ν contradicts the data in important ways.

By taking $k_1 > 2$, we can also study the properties of the higher dimensional marginal distributions, though we typically have much less, if any, prior knowledge to tell whether such model specification is reasonable. For example, with $k_1 = 3$, (6) implies

$$f_3(r_{12}, r_{13}, r_{23} \mid \nu) \propto \frac{(1-r_{12}^2-r_{13}^2-r_{23}^2+2r_{12}r_{13}r_{23})^{\nu-k+1}}{[(1-r_{12}^2)(1-r_{13}^2)(1-r_{23}^2)]^{\frac{\nu-k+3}{2}}},$$

where $1-r_{12}^2-r_{13}^2-r_{23}^2+2r_{12}r_{13}r_{23} \geq 0$.

2.3. The jointly uniform prior

Given that the correlation matrix space $\mathcal{R}^k$ is a compact subspace of the $k(k-1)/2$ dimensional cubic $[-1,1]^{k(k-1)/2}$, one may also consider the (proper) uniform prior on $\mathcal{R}^k$:

$$p(R) \propto 1, \quad R \in \mathcal{R}^k. \tag{9}$$

However, due to the shape of $\mathcal{R}^k$, one must be aware of the fact that the joint uniformity on $\mathcal{R}^k$ results in marginal priors for individual correlations, r_{ij}, which are not uniform. Figure 1 shows the marginal distribution of one of the $(k-1)k/2$ individual correlations when $k = 3$ and when $k = 10$. Clearly, these priors are not uniform. They favor values close to zero over values close to ± 1. Intuitively, this is because the positive definite constraint is more restrictive as the correlations move away from zero in $\mathcal{R}^k$. This can be visualized in Figure 2, which compares the jointly uniform prior (9) with the marginally uniform prior (8) when $k = 3$. It is seen that in order to maintain the marginal uniformity, density (8) has to place more mass at the corners of $\mathcal{R}^3$, as made clear by the first two sets of panels on the left in Figure 2 (i.e., when r_{23} is close to -1). For readers who are interested in more discussion of the shape of correlation matrices, see Rousseeuw and Molenberghs (1994) for an intriguing exploration. Note that in some applications, it is desirable to have a prior that favors value of r_{ij} close to zero.

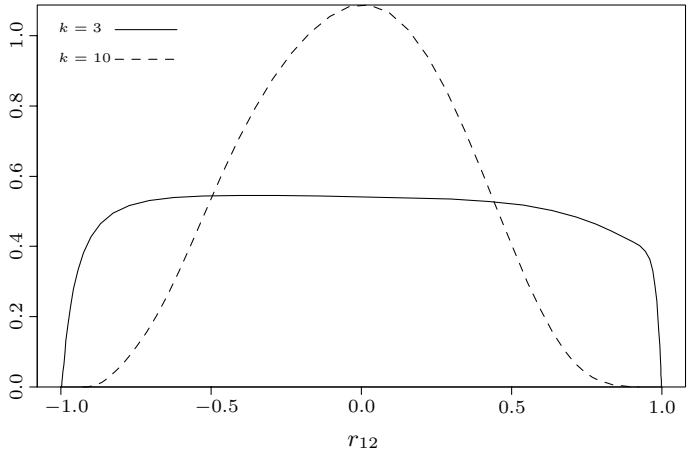


Figure 1. Marginal prior density for r_{12} when $k = 3$ and $k = 10$ under the prior $p(R) \propto 1$. Estimated densities based on 2000 draws.

We saw in Figure 1 that as k increases the marginal prior on individual correlations under the jointly uniform prior (9) tightens up around zero. (Notice the estimated densities are slightly asymmetric when the real ones should be symmetric about zero – we purposely did not use any estimate throughout the simulation that is numerically constrained to be symmetric in order to use the symmetry as a diagnostic tool for the performance of the simulation). This suggests the possibility that this prior is highly informative in that data cannot drive the marginal posterior of a correlation toward ± 1. To investigate this

possibility we simulated i.i.d. data from the multivariate normal distribution with known zero mean and

$$\Sigma = \frac{1}{11}(I + a\iota\iota'), \tag{10}$$

where ι is a column vector of k ones, I is the identity matrix, and a is either 1 or 10. These two choices of Σ result in common correlation $r_{ij} \equiv r = 0.50$ and 0.91, respectively. The large common correlation (i.e., 0.91) is chosen simply as an extreme illustrative example, because in real applications we rarely have correlations that are all very large – if that happens then we would use a model with a random common mean (e.g., a random-effect model) to deal with the high correlations. Also see Gelman, Bois and Jiang (1996) for the use of reparameterization to reduce substantial population correlation. The method we discuss in this paper is most beneficial when some correlations are large and some are small (but we do not know which is which *a priori* and would like the data to "speak").

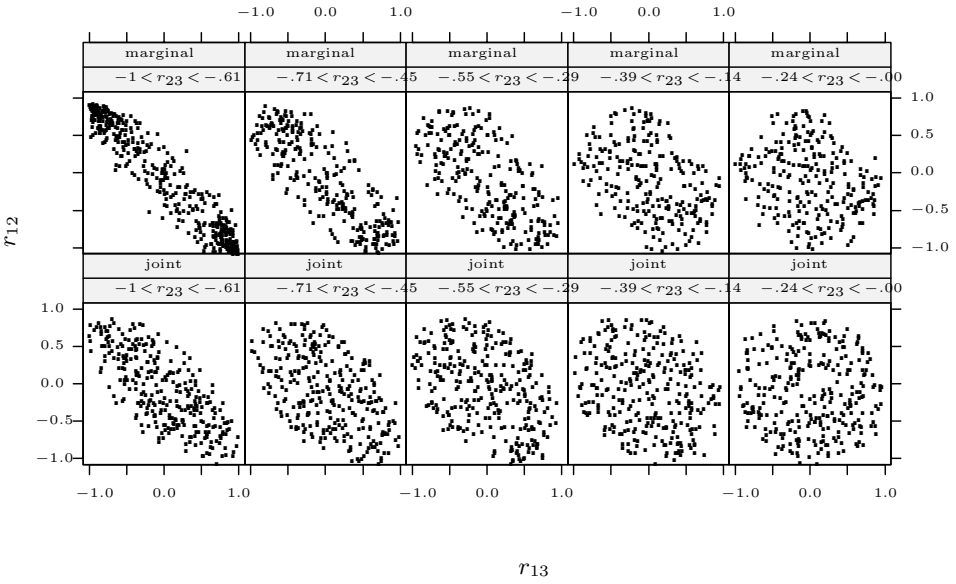
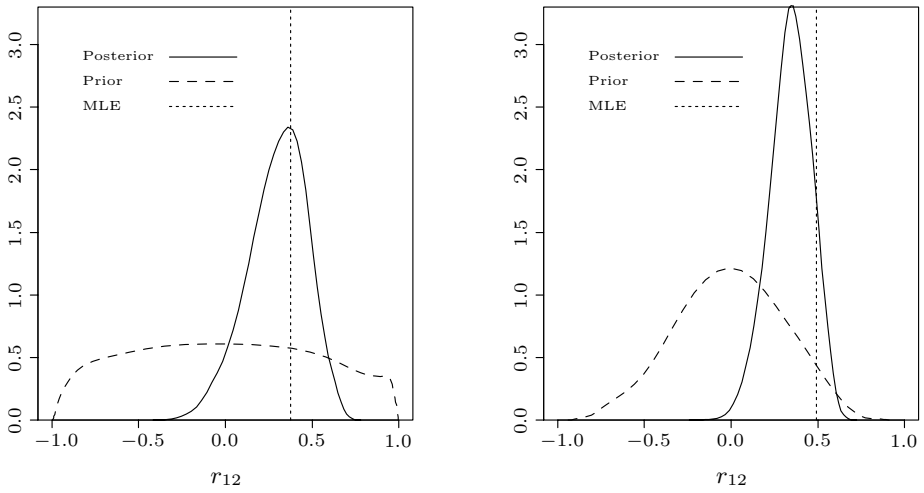


Figure 2. Multi-panel scatter plots comparing draws from the marginally uniform prior distribution (8) (in row 1 of the figure) for R and the jointly uniform prior distribution (9) (in row 2 of the figure) for R when $k = 3$. Each panel gives a scatter plot of draws of r_{12} versus r_{13} when the value of r_{23} is in the range given above the respective plot. Only negative values of r_{23} are given because the plots for the corresponding positive values of r_{23} are essentially mirror images. Each row in the multi-panel figure is based on 1000 random draws from the respective prior distribution.

We considered $k = 3$ with 20 observations and $k = 10$ with 50 observations; the sample sizes were chosen to be relatively small (as otherwise the impact of prior is not too much of a concern) and to roughly equate the sample sizes per dimension (this choice is by no means "fair", which is difficult to achieve when comparisons are made across different dimensions). In both cases, we used independent log normal priors with normal mean 0 and standard deviation .1 for the elements of S. For $r = 0.5$, Figure 3(a) displays the marginal prior and posterior densities for $k = 3$, and Figure 3(b) displays the densities for $k = 10$. For $r = 0.91$, Figure 4(a) displays the marginal prior and posterior densities for $k = 3$, and Figure (4b) displays the densities for $k = 10$.



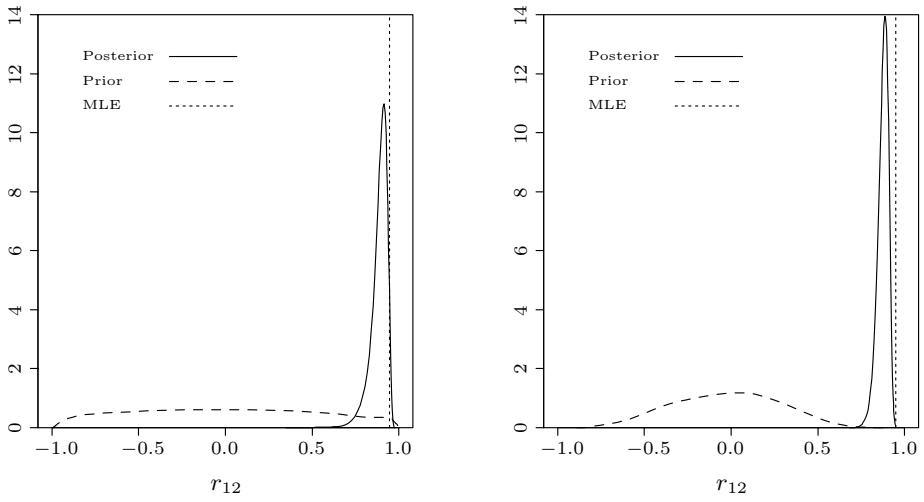
(a) $k = 3$ with 20 observations. Posterior mode is 0.36, posterior median is 0.32, and MLE is 0.37.

(b) $k = 10$ with 50 observations. Posterior mode is 0.35, posterior median is 0.35, and MLE 0.49.

Figure 3. Marginal prior and posterior densities for r_{12} with $k = 3$ and $k = 10$ when the true Σ is given by (10) with $a = 1$, thus, the true value of r_{12} is 0.50. Estimated densities based on 2000 draws.

In assessing the figures we looked at two criteria: (1) the position of the MLE of r_{12} (i.e., the sample correlation with divisor n) with respect to its marginal posterior; (2) the position of the true value of r_{12} with respect to its marginal posterior. Regarding the first criterion, the figures show that the posterior mode and median are slightly closer to the MLE when k is 3, indicating that when k is 10, the more informative marginal prior on r_{12} pulls the posterior slightly towards zero. In other words, when k is 10 compared to k being 3, the uniform prior on R has more of an impact on the posterior estimates, as expected from Figure 1. Regarding the second criterion, the figures show that the true value of r_{12} is well within the posterior mass for both cases. Therefore, the posterior has

not been misled by the somewhat informative priors on r_{12}; this is partially due to the fact that the marginal posterior depends on the joint prior for (S, R), not just on the marginal prior for r_{12}.



(a) $k = 3$ with 20 observations. Posterior mode is 0.92, posterior median is 0.90, and MLE is 0.94.

(b) $k = 10$ with 50 observations. Posterior mode is 0.89, posterior median is 0.88, and MLE is 0.95.

Figure 4. Marginal prior and posterior densities for r_{12} with $k = 3$ and $k = 10$ when the true Σ is given by (10) with $a = 10$, and thus the true value of r_{12} is 0.91. Estimated densities based on 2000 draws.

For either value of r and for both cases of k the posterior tightens up around large correlation values relative to the priors. The posteriors when $r = 0.91$ are much tighter about their true values than the posteriors when $r = 0.5$, since the likelihood is much sharper for large correlations. Although our marginal priors are tight, particularly when k is 10, because of the shape of the space of correlation matrices, this marginal tightness does not stop the likelihood from pushing the posterior toward one of the corners of the space. Thus, while examination of the marginal prior distributions of the correlations might suggest that the jointly uniform prior is highly informative when k is large, our limited empirical investigations suggest that such informativeness may have tolerable impact on the marginal posteriors, as long as k is not too large relative to the amount of data. (As we mentioned in Section 2.1, if k is too large, there are often other considerations that take place, such as grouping the underlying variables.) Note also that centering the marginal priors of correlations around zero is much less extreme than setting $R = I$, which has been a common strategy when flexibility in dealing with individual components is desired. In the absence of any realistic

prior knowledge of r_{ij}, it is difficult to argue a prior distribution centered at a point other than zero.

2.4. Invariance and coherence considerations

The primary motivation for the separation strategy is its flexibility and directness in soliciting modeling information, as most practitioners are more comfortable thinking in terms of standard deviations and correlations rather than in terms of the spectral decomposition or matrix logarithm of Σ. Nevertheless, caution is needed when adopting this strategy. For example, the standard inverse-Wishart model is invariant, in terms of the distributional form, under rotation of variables underlying a covariance matrix, but priors resulting from the separation strategy are generally not. Whether one should dismiss such priors on this ground depends on the application. If rotation invariance is important, such as in some shape analyses (e.g., Dryden and Mardia (1997)), then any distributional family that lacks rotation invariance is clearly inadequate as a candidate for modeling; a model can be thrown out just for not having desirable invariance or other theoretical properties (e.g., Gelman (1996)).

For the shrinkage application discussed in Section 3.1, the underlying variables are regression coefficients, and their rotations are of interest only if we are (simultaneously) also interested in regressing the same dependent variable on rotations of the regressors, a situation that we have not encountered. For the general location model application in Section 4, it is often the case that a linear combination of the components of the dependent variable has no substantive meaning (e.g., a linear combination of log body weight and log blood pressure, as in the imputation application described in Barnard (1995)), though it makes perfect sense to model the correlations between the components (e.g., between log body weight and log blood pressure). In such cases, it is not fruitful to insist on rotation invariance, especially if such an invariance is achieved at the expense of model restrictions that have adverse effects (e.g., as with the standard inverse-Wishart). We do, however, want scale invariance for obvious reasons – the priors from the separation strategy are scale invariant, with respect to a distribution *family*, as long as $p(S)$ is chosen to be so (e.g., with the log-normal prior, $\log S$ and $\log(cS)$ belong to the same normal distribution family). In general, while a joint distributional specification of a set of variables determines the joint distribution of any transformation of the variables, whether these two joint distributions should be required to belong to the same distributional family should depend on whether the two sets of variables have the same substantive meaning. See Zellner (1991) for more discussions about the dependence of invariance requirements on the underlying problems.

Coherence is another theoretical consideration for (Bayesian) modeling. The lower-dimensional joint distributions of the correlation coefficients from both the jointly uniform distribution on $\mathcal{R}^k$ and the marginal inverse-Wishart distributions depend on the total number of underlying variables (i.e., k). While it may be useful to have equally flexible distributions without such dependence, we point out that the commonly employed inverse-Wishart distribution has the same dependence (e.g., as can be seen from (7)). In other words, the models (priors) we use for Σ in this paper are no worse than what is being commonly used with respect to the issue of coherence. In fact, it is not always insensible to allow such dependence, especially in the context of (imperfect) prior elicitation. As an extreme example, knowing r_{12} is a common correlation among X_1, X_2, and X_3 (i.e., $r_{12} = r_{13} = r_{23}$) changes the support of r_{12} from $[-1, 1]$ to $[-0.5, 1]$ regardless of whether we actually have data for X_3.

3. Applications to Shrinkage Estimation

3.1. A shrinkage model for normal regressions

As in Lindley and Smith (1972), we have m normal regressions:

$$Y_j \mid X_j, \beta_j, \tau_j \sim N(X_j\beta_j, \tau_j^2 I_{n_j}), \quad j = 1, \ldots, m. \tag{11}$$

As usual, Y_j is a vector in R^{n_j} and X_j is an $n_j \times k$ matrix of explanatory variables, and given X_j and the parameters β_j and τ_j, $j = 1, \ldots, m$, the observations Y_j are independent. Our basic modeling intuition is that each regression is a particular instance of the same type of relationship. Consequently, while the parameters may vary from regression to regression, we may have prior beliefs that they are similar. For instance, in Section 3.3 we present an example where each regression corresponds to a different firm. From each firm, the same type of Y and X are measured. We have beliefs that firms in the same industry will have similar parameter values.

To model our prior beliefs about the *degree of similarity*, we start by letting $\beta_j \sim$ i.i.d. $N(\bar{\beta}, \Sigma)$. The i.i.d. assumption represents the fact that the β's are *a priori* exchangeable because we do not have prior information to distinguish them. The assumed normality is a part of the prior belief about the the degree of similarity. With $\bar{\beta}$ and Σ considered fixed and known, the inferences in the m regressions are still independent, but the posterior of each β_j will be shrunk towards the common mean $\bar{\beta}$. Typically, however, we are uncertain about $\bar{\beta}$ and Σ. We must specify (hyper-)prior distributions for $\bar{\beta}$ and Σ to capture this uncertainty.

The model discussed above describes what the different regressions have in common. The model has a hierarchical form as follows:

$$\begin{aligned} &Y_j \mid X_j, \beta_j, \tau_j^2 \sim N(X_j\beta_j, \tau_j^2 I_{n_j}), \quad j = 1, \ldots, m, \\ &\beta_j \mid \bar{\beta}, \Sigma \overset{\text{i.i.d.}}{\sim} N(\bar{\beta}, \Sigma), \quad j = 1, \ldots, m, \end{aligned} \tag{12}$$

where priors must be chosen for the τ_j^2's, $\bar{\beta}$, and Σ. Convenient and usually adequate choices are available for $\bar{\beta}$ and the τ_j^2's, e.g., the (conditionally) conjugate normal prior for $\bar{\beta}$ and inverse-gamma priors for the τ_j^2's. In many cases the assumption that $\bar{\beta}$, τ_j^2's, and Σ are *a priori* independent is also reasonable.

Our focus is on the choice of prior for Σ. The choice is crucial, because it determines the nature of the shrinkage of the posterior of the individual β_j towards a common target. Prior beliefs that Σ is small result in more shrinkage, and prior beliefs that Σ is large result in less shrinkage. The tightness of the prior on Σ determines the degree to which the shrinkage adapts to the data. As an example, consider a situation where our prior on Σ is quite vague. Suppose all but one of the β_j's are clearly shown by the data to be very similar, and there is weak evidence that the remaining β vector is far from the rest. The well estimated β_j's will convince us that Σ is small, resulting in shrinkage of the remaining β vector towards the rest.

In the shrinkage context our prior information is primarily about the level of variation of the individual coefficients across the m regressions. Let $\beta_j = (\beta_{1j}, \beta_{2j}, \ldots, \beta_{kj})^\top$. We are interested in the variation in the set of values $\{\beta_{i1}, \beta_{i2}, \ldots, \beta_{im}\}$. Marginally, we have

$$\beta_{ij} \overset{\text{i.i.d.}}{\sim} N(\bar{\beta}_i, \sigma_i^2), \quad j = 1, \ldots, m, \tag{13}$$

where $\sigma_i^2 = \Sigma_{ii}$ is the ith diagonal element of Σ. Let $s_i = \sqrt{\sigma_i^2}$. The prior belief that s_i is small says that the coefficient for the ith explanatory variable is similar in the different regressions. We want our prior to be able to express beliefs about each s_i in a simple way. Since the explanatory variables have different units in general, so do the β_{ij} and s_i for different i. Thus, we also need to be able to assign a simple prior to each s_i which may support quite different values.

Thus, a key prior belief in our shrinkage model is about the standard deviations corresponding to Σ. If we make the further (often realistic) assumption that there is little prior information about the correlations, and, thus, we do not want to use a strong prior such as $R = I$, then we are exactly in the situation discussed in Sections 1 and 2. We choose an independent log normal prior for each of the standard deviations. For R we can use a variety of prior distributions, for example the joint uniform prior and the marginal prior from the inverse-Wishart with various choices of the degrees of freedom, including the marginal uniform prior, as discussed in Section 2.

It is worthwhile to contrast the separation approach with other approaches in the literature in the context of constructing prior distributions for Σ. Besides not shrinking or complete shrinking (i.e., $\Sigma = \infty$ or $\Sigma = 0$), using an inverse-Wishart prior for Σ is by far the most common approach in practice, mainly because it is the (conditionally) conjugate prior and, thus, renders some

mathematical simplicity. The restriction of common degrees of freedom for all components of S, however, makes it virtually useless when we have different *a priori* assessments for different components of S, as we discussed earlier. The matrix logarithm approach of Leonard and Hsu (1992) is much more flexible but suffers from a different kind problem. The matrix logarithm of Σ has no direct statistical meaning, and S is not a simple function of the matrix-logarithm parameters. Consequently, we found it difficult to use the matrix logarithm approach to produce a prior that captures our prior knowledge about S. The same thing can be said about the spectral decomposition approach, another common method in practice. For example, Banfield and Raftery (1993) and Bensmail and Celeux (1996) have successfully utilized the spectral decomposition of a covariance matrix in clustering and discrimination modeling problems by viewing the decomposition in terms of the shape, volume, and orientation of a covariance matrix. We have found it difficult, however, to use these ideas to capture our prior beliefs in the shrinkage context.

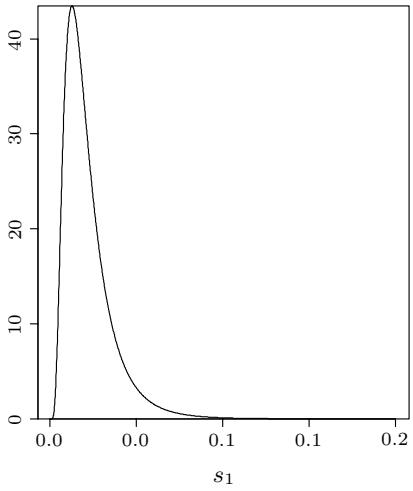
The method of Pourahmadi (1999, 2000) is useful for modeling the covariance matrix for a response variable whose components have a natural ordering (e.g., from a longitudinal study), but it loses its appealing properties when the goal is to model the prior of the covariance matrix of regression coefficients (e.g., it is necessary to specify priors on the parameters of the conditional regressions of elements of regression coefficient β on the other elements of β assuming an ordering of the elements of β). Similarly, the reference prior approach of Yang and Berger (1994) is very useful for constructing posterior estimators with good frequentist properties, considerably better than estimators derived using the Jeffreys prior. But reference priors are designed for cases where one is seeking a non-informative prior, and thus they are not suitable when the goal is to provide a simple means of capturing informative prior beliefs in Σ, especially in S. However, the reference prior approach might be useful for constructing a noninformative prior for R when our modeling focus is on S.

In the next two sections we provide two examples to show the flexibility and simplicity of the separation strategy for modeling a covariance matrix in the shrinkage context. The first example is simulated with three regression coefficients and ten regressions, and is designed to illustrate some of the shrinkage properties associated with the priors in Section 2. The second example uses a real data set involving six regressions with two regression coefficients. The number of regression coefficients (i.e., k) is kept low for ease of illustration.

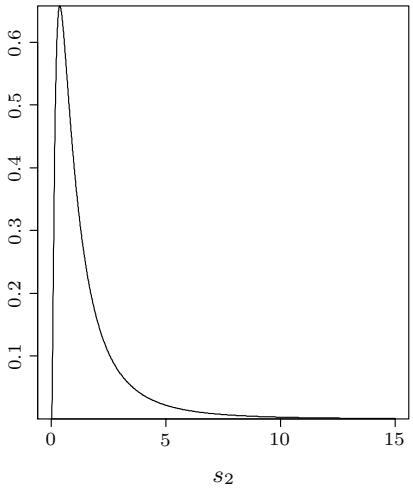
3.2. A simulation study

In this simulation study, the true values of the first coefficient in the ten regressions were evenly spaced from 1 to 2.8. (We have not let our prior normal

assumption on the β's influence our simulation configuration in order to reflect a realistic situation.) The sample standard deviation of the ten values is 0.6. To illustrate how our choice of prior affects the posterior through shrinkage, we choose a prior for s_1, the standard deviation of the first coefficient, that concentrates on small values. We choose our prior for $\log(s_1)$ to be $N(-4, .6^2)$. Figure 5(a) shows the corresponding prior density for s_1. The 10%, 50%, and 90% quantiles are 0.01, 0.02, and 0.04, respectively. Thus, we are inputing the prior belief that values of the first coefficient in the ten regressions are much closer together than they actually are. The long right tail of the prior is intended to reflect our prior believe that, although s_1 is likely to be small, there is a chance that it could be much bigger.



(a) Prior density for s_1.



(b) Prior density for s_2. Prior density for s_3 is the same as that for s_2.

Figure 5. Prior density for s_1, s_2, and s_3. *A priori*, $\log s_1 \sim N(-4, 0.6^2)$ and $\log s_2 \sim N(0, 1)$.

The true values for the second coefficient are the same as those of the first. In contrast to the prior chosen for s_1, our prior for s_2 is spread out. We choose our prior for $\log(s_2)$ to be $N(0, 1)$. Figure 5(b) shows the corresponding prior density for s_2. The 10%, 50%, and 90% quantiles are 0.3, 1.0, and 3.7, respectively.

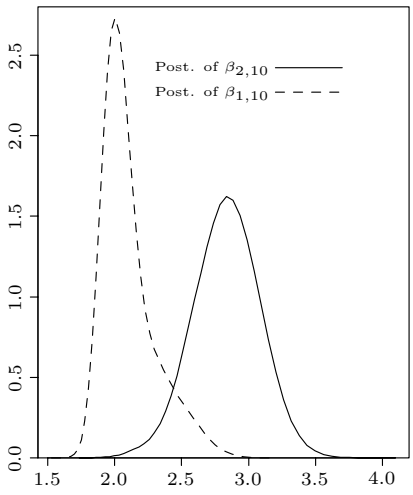
The true values for the first two coefficients are chosen to illustrate the effect of our priors on the degree of shrinkage. The prior for s_1, the standard deviation of the first coefficient, is chosen so that the posteriors of the first coefficient will be shrunken together across the ten regressions. The prior for s_2, in contrast, is chosen to express the possibility of substantial variation in the coefficients, that is, we expect little shrinkage for the posteriors of the second coefficient. The

values of the third coefficient and the prior for s_3 are chosen to illustrate the situation where the goal is to learn about the variation of the coefficient and there is little prior information. In each regression, the true values of the third coefficient are set to one. Our prior for s_3 is the same as that for s_2, i.e., the prior for s_3 is spread out. In this case we hope to learn something about the degree of similarity of the third coefficient across regressions from the posterior of s_3.

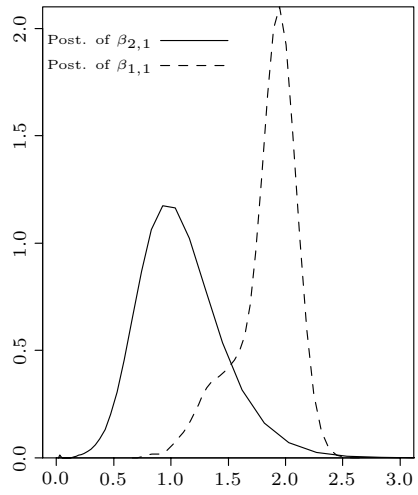
This example illustrates the flexibility of our prior modeling approach. By allowing separate priors on each of the standard deviations, it is easy to express our prior beliefs for each variable about the similarity of the individual regression coefficients, and at the same time we can use almost any prior for R. In our example we believe *a priori* that the regression coefficients for the first variable are quite similar, while our prior beliefs about the degree of similarity of the regression coefficients for each of the other two variables are much less certain. In contrast, it is essentially impossible to express these beliefs for the standard deviations using the commonly used inverse-Wishart prior for Σ.

In each regression, all the values for the explanatory variables are drawn from the standard normal distribution, there are 20 observations, and $\tau_i = 1$. The prior for each τ_i is $\log(\tau_i) \sim N(0,1)$. For $\bar{\beta}$ we have $p(\bar{\beta}) \sim N(0, 1000I)$, which will be extremely flat relative to the likelihood. As discussed, our prior for Σ has $\log(S) \sim N((-4,0,0)^\top, \mathrm{diag}(.36,1,1))$ and $p(R) \propto 1$. Note that we also performed the simulations using the marginally uniform prior (8) instead of the jointly uniform prior $p(R) \propto 1$. The results were very similar under the two priors, which is not unexpected given $k = 3$.

Figure 6(a) displays an estimate of the posterior density of $\beta_{2,10}$, the second coefficient in the tenth regression, and an estimate of the posterior density of $\beta_{1,10}$, the first coefficient in the tenth regression. The true value of both coefficients is 2.8. We see that the posterior of $\beta_{2,10}$ is centered close to 2.8, while the posterior of $\beta_{1,10}$ is strikingly different from that of $\beta_{2,10}$. The mode of the posterior of $\beta_{1,10}$ is close to 2, which is the average value of the first coefficient across the ten regressions. This reflects the extreme shrinkage caused by the prior of s_1. The posterior has a long right tail. This tail is trying to stretch out to cover the true value suggested by the data. We find this posterior very appealing in the manner in which it reflects prior and sample information. The shrinkage demanded by the prior is reflected in the mass around 2. The right tail correctly indicates that there is substantial uncertainty about the value, and that it may actually be much larger than 2. Figure 6(b) shows a similar story for the first and second coefficients from the first regression, $\beta_{1,1}$ and $\beta_{2,1}$. For the first regression, the true value of both coefficients is 1.0.

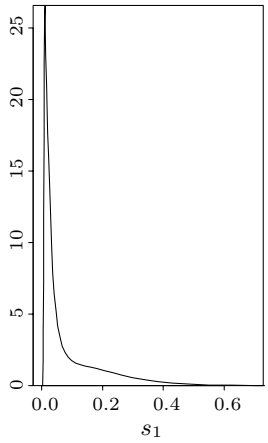


(a) Estimated posterior density for $\beta_{1,10}$ and $\beta_{2,10}$. The true value of both β's is 2.8.

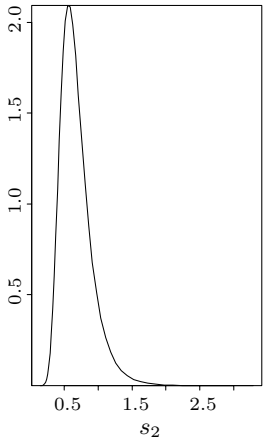


(b) Estimated posterior density for $\beta_{1,1}$ and $\beta_{2,1}$. The true value of both β's is 1.0.

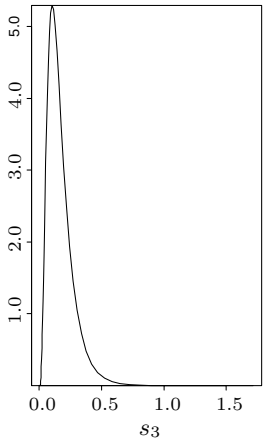
Figure 6. Comparisons of estimated posterior densities of β_1 and β_2. All estimates based on 5000 posterior draws, which were obtained by subsampling every 100th draw after a burn-in period of 100 iterations.



(a) Estimated posterior density for s_1.



(b) Estimated posterior density for s_2.



(c) Estimated posterior density for s_3.

Figure 7. Estimated posterior densities of s_1, s_2, and s_3. All estimates based on 5000 posterior draws, which were obtained by subsampling every 100^{th} draw after a burn-in period of 100 iterations.

Figures 7(a), 7(b), and 7(c) display estimates of the posterior distributions of s_1, s_2, and s_3, respectively. Notice the long right tail of the posterior of s_1 (in

comparison with its prior in Figure 1(a)) trying to stretch out to cover the kind of values suggested by the data in defiance of the strong prior on smaller values. The tail corresponds to the tail on the posterior of $\beta_{1,10}$ and on the posterior of $\beta_{1,1}$. All three of these long tails are made possible by the long right tail in the prior for s_1. We consider the ability to conveniently capture such a tail a strong point in favor of the choice of the log-normal prior for the elements of S. The posterior of s_2 correctly indicates a value a bit bigger than 0.5. The posterior of s_3 is concentrated on small values relative to the prior, reflecting the sample information that there is little variation in the third coefficient.

3.3. A simple real-data example

This example is a simplified version of an approach which has been used recently in the finance literature (e.g., Stevens (1996)). Much empirical work in finance relates the returns of small assets (e.g., individual firms) to returns on portfolios, which are made up of many assets. These portfolio returns are designed to capture broad market factors.

In our simple example we will consider only one factor which we call "the market". It is meant to capture overall market activity. The market is represented by returns on the value weighted market portfolio, which is a portfolio of many stocks, where the weight of each stock in the portfolio is proportional to value of the firm (stock price multiplied by number of outstanding shares).

We regress returns of an individual stock on market returns. Our data consist of monthly returns on each stock of interest and monthly returns on the market portfolio. The regression slope is of particular interest in certain financial calculations (Fama (1976), Brealy and Myers (1984)). In practice there may be difficulty estimating this coefficient. While a great deal of data is available, firms change considerably over time so that only fairly recent data may be deemed relevant.

Our use of the hierarchical model is motivated by the need to improve estimates of these regression coefficients and the belief that regression coefficients for firms in the same industry may be similar. The estimates are improved because if the coefficients are similar, then the s_i's from Σ will be small and, thus, the shrinkage will give more precise estimates of the individual coefficients. We consider six airlines: American, Continental, Delta, KLM, Southwest, and United. Our data consists of the returns for the six airlines and the value weighted market in each of the 36 months from 92/01/31 to 94/12/30. Our prior is used to specify how similar we think the regression slope is in the six regressions of airline returns on the market returns. Since none of us is an expert on the airline industry, nor do we have reliable information for realistic prior specification, we will use three

prior specifications to conduct a sensitivity study to illustrate the impact of the choice of the prior on the posterior.

Figure 8 shows the scatter plots of each airline's returns versus the market returns (returns have been multiplied by 100). To give the reader unfamiliar with this data some idea of the magnitudes we are dealing with, the least squares estimates (standard errors) of the six regression slopes are 0.57 (.53) [United], 1.21 (.63) [Continental], 1.27 (.50) [Delta], 1.47 (.70) [KLM], 1.53 (.49) [American], and 1.55 (.65) [Southwest]. Note that the first slope (United) is substantially less than the rest, but all of the SEs are quite large. It is of interest to see the effect of our shrinkage model with its associated prior specifications on the estimation of the United slope. To what extent will it be shrunk towards the others? In practice there is a big difference between a slope of 1.0 and a slope of 0.5.

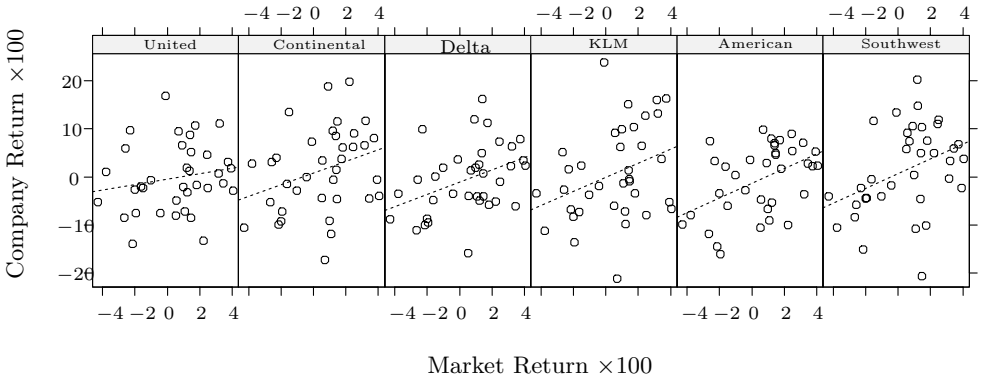
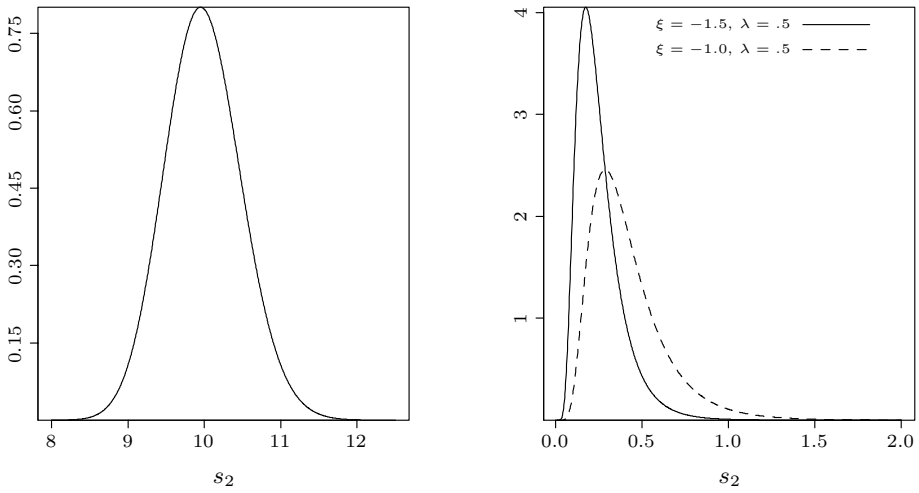


Figure 8. Monthly airline return times 100 versus monthly market return times 100 for each of the six airlines for the period 92/01/31 to 94/12/30. The dashed line in each panel is the least-squares regression line from the regression of the corresponding airline returns on the market returns. The panels are ordered by the slope of the least-squares line in each panel.

Although several choices must be made in the prior specification, we concentrate on our choice of prior for s_2, which denotes the standard deviation of the regression slopes across the regressions (we let s_1 be the standard deviation of the intercept). This is the prior choice that is closely related to the shrinkage of the posteriors of the regression slopes corresponding to the six airlines. Because $k = 2$ for this problem, the jointly uniform prior of R is the same as the marginally uniform prior. All other priors were carefully chosen to be plausible but not too informative. We attempted to make them spread out without giving support to values which are clearly *a priori* improbable. In particular, the prior for s_1, the standard deviation of the regression intercepts (i.e., the standard

deviation of β_{1j}), is log-normal with normal mean 1 and standard deviation 1. This prior is quite disperse, indicating an uncertain prior belief about the degree of similarity of the regression intercepts. A key feature of our prior specification strategy is the ability to allow strong prior shrinkage for some variables, and weak or no prior shrinkage for other variables. In this example, we are uncertain about shrinking the intercepts, but will (and easily can with the separation strategy) entertain a variety of prior shrinkage beliefs about the regression slopes.

Figure 9(a) shows our first choice of prior for s_2. It is log-normal with normal mean $\xi = 2.3$ and standard deviation $\lambda = 0.05$. Most of the mass is concentrated on values between 9 and 11. This highly informative and clearly ridiculous prior (as can be seen from a posterior predictive check) keeps the amount of shrinkage small; this extreme prior is used purely for illustration because in practice such an extreme prior is unlikely to be used even as part of a sensitivity study. Figure 10(a) displays boxplots of draws from the corresponding posteriors of each of the six regression slopes. The medians are 0.58, 1.32, 1.23, 1.50, 1.47, and 1.63 (corresponding to the order of the airlines in Figure 10). Thus, the medians are quite close to the least squares estimates, reflecting the absence of shrinkage. Figure 11 shows an estimate under this prior of the posterior density of the slope for United. We see that much of the mass is on values less than one.



(a) First prior density of s_2. Prior parameters are $\xi = 2.3$ and $\lambda = 0.05$.

(b) Second and third prior densities of s_2. Prior parameters specified in the key.

Figure 9. Prior densities of s_2, the standard deviation of the regression slope, for three choices. For all priors, $\log(s_2) \sim N(\xi, \lambda^2)$.

Figure 9(b) shows our second and third choices of prior for s_2. The second choice is log-normal with normal mean $\xi = -1.0$ and standard deviation $\lambda = 0.5$,

which is chosen to represent a plausible ball-park prior (i.e., it is reasonable but not too informative). For example, the right tail extends well past values of 0.5. A standard deviation of 0.5 (i.e., $s_2 = 0.5$) suggests that most slopes are within 1 of the mean, which *a priori* seems like a large interval based on the authors' experience with similar studies. (Note that the sample standard deviation of the six least-squares estimates is 0.37.) Figure 10(b) displays boxplots of draws from the posteriors of each of the six regression slopes. Although the prior is not meant to be informative, the results are quite different than those obtained from the first prior. The posteriors have been noticeably shrunk together. Figure 11 shows that the marginal posterior of the slope for United has shifted to the right compared to the corresponding posterior for the first prior on s_2.

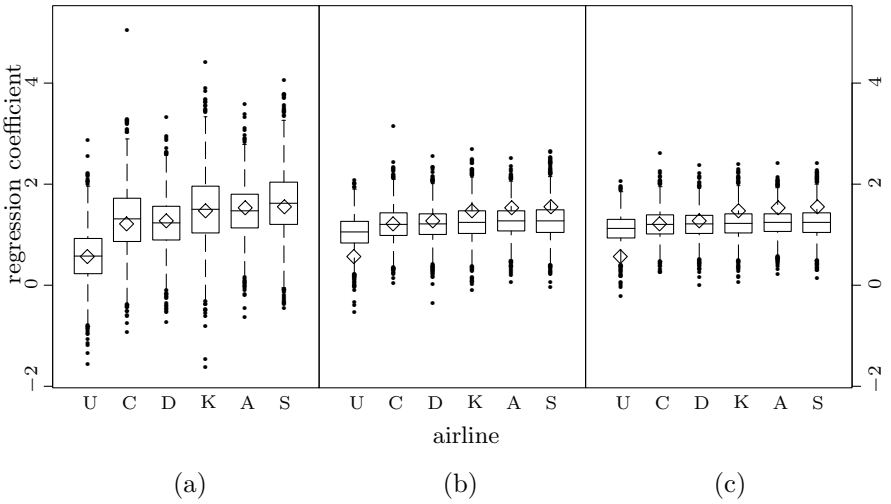


Figure 10. Boxplots of 3000 posterior draws of the regression slopes of the 6 airlines: United (U), Continental (C), Delta (D), KLM (K), American (A), and Southwest (S). Draws were obtained by subsampling every 20th draw after a burn-in period of 100 iterations. The $\diamond$ for each airline is the least-square estimate based on the data from the corresponding airline. Figures (a) ($\xi = 2.3$, $\lambda = 0.05$), (b) ($\xi = -1.0$, $\lambda = 0.5$), and (c) ($\xi = -1.5$, $\lambda = 0.5$) are the posterior draws under the three priors for s_2 given in Figure (a) (corresponding prior parameters for s_2 are given in parentheses).

The third prior is log-normal with normal mean $\xi = -1.5$ and standard deviation $\lambda = 0.5$, which is intended to represent a plausible actual prior. This prior was the result of one of the authors honest attempt at eliciting a prior specification. Figures 10 and 11 show that this prior results in substantial shrinkage relative to the first prior. The posterior medians are 1.12, 1.20, 1.21, 1.22, 1.25,

and 1.24. Even relative to the second prior, we see that in Figure 11, the posterior of the United slope is shifted a little to the right and is tighter.

Both the second and the third prior result in substantial shrinkage. The median of the United slope moves from 0.58 to a value larger than one. However, the basic conclusion has to be that even with the shrinkage model, the posteriors are quite spread out, so that with 36 months of data we have not obtained very precise inferences about the quantities of interest. Should we really shrink United's slope? Our hierarchical setup is a model, and, like any model, it should be evaluated in the light of the data and substantive knowledge. For example, we could conduct a posterior predictive check (Rubin (1984), Gelman and Meng (1996), Gelman, Meng and Stern (1996)) of our hierarchical model to access its goodness-of-fit (e.g., the normality assumption on the β's) with the observed data. The fact that the United slope is so unlike the rest may suggest that we learn more about the company, for example, by analyzing its historical data or by trying to find out any unusual event during that 36 months that seriously invalidates the normality assumption underlying our hierarchical model.

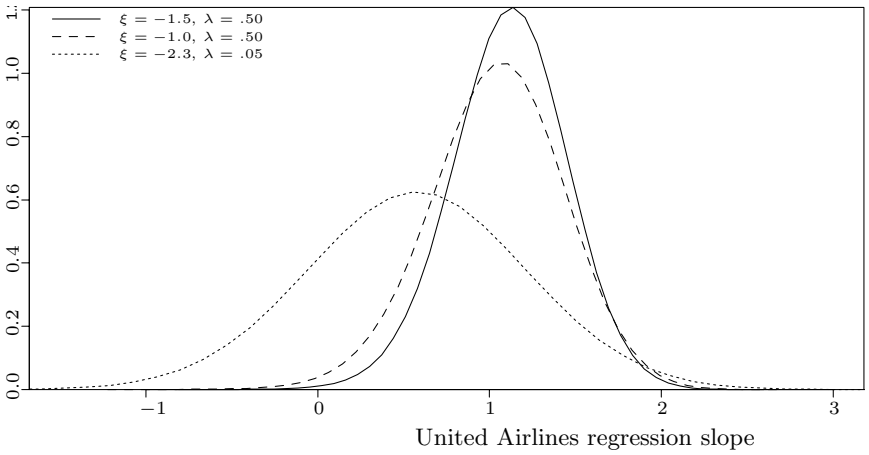


Figure 11. Estimated posterior densities of the regression slope for United Airlines. The three densities correspond to the three priors on s_2 shown in Figure 9. The corresponding prior parameters for each estimated posterior density are given in the key. All estimates are based on 3000 posterior draws.

4. General Location-Scale Model

4.1. General location model

Many applications, particularly in social science, require models that can simultaneously deal with categorical and continuous variables. The general location model (GLOM) (Olkin and Tate (1961), Krzanowski (1980, 1982), Little and

Schluchter (1985), Little and Rubin (1987), Schafer (1997)) has been one of the most frequently used models of this kind. One important application of GLOM is for producing multiply-imputed public-use data files (e.g., Schafer (1997), Rubin (1987, 1996), Meng (1994)). Here we first review the GLOM and its recent extensions and then we propose a generalization of the GLOM, based on our strategy of decomposing a covariance matrix into its correlations and standard deviation. Our generalization allows for much greater flexibility in specifying the conditional covariance structure of the continuous variables, and allows for shrinkage estimation of covariances under a Bayesian hierarchical model.

Suppose we have a k-dimensional continuous variable $Y = (Y_1, \ldots, Y_k)^\top$ and a q-dimensional categorical variable $Z = (Z_1, \ldots, Z_q)^\top$. Each element of Z, Z_j, has c_j levels for $j = 1, \ldots, q$. The categorical variables form a contingency table with $c = \prod_j c_j$ cells, one cell for each possible value z of Z. It is convenient to index the cells either according to the values of Z, or by assigning each cell a number from 1 to c, where the number is a function of the cell value z. We will use both notations interchangeably.

Under the GLOM, the marginal distribution of the categorical variable Z is multinomial and the conditional distribution of Y given Z (i.e., given a particular cell) is multivariate normal with different means across cells defined by the categorical variables, but a common covariance matrix across cells. More precisely, under the GLOM the joint distribution of (Y, Z) is as follows:

1. $Z \mid \pi = (\pi_1, \ldots, \pi_c)^\top \sim \text{Multinomial}(\pi)$, $\sum_{i=1}^{c} \pi_i = 1$, where π_i is the probability that a realization falls in cell i, $i = 1, \ldots, c$, where (as noted above) $i(z)$ is a function of z.
2. $Y \mid Z = z, \mu_z, \Sigma \sim N(\mu_z, \Sigma)$, where μ_z is the mean of Y in the cell specified by z, and Σ is the common (across cells of the contingency table) conditional variance of Y.

Structure is often imposed on the cell probabilities π and the cell means μ_z, for example, by specifying a log-linear model for π and a linear (hierarchical) regression model for the μ_z (see Schafer (1991), Schafer, Khare and Ezzati-Rice (1993), Raghunathan and Grizzle (1995), Schafer (1997)).

A key aspect of the GLOM is the assumption of a common covariance for Y across cells defined by the categorical variable Z. While it is clearly inadequate in many applications to assume common covariance across all cells, especially when there are many cells, it is also clear that allowing each cell to have its own covariance matrix is impractical. For one thing, such a model often has many more parameters than data points. An example of the inadequacy of the GLOM is mentioned in Barnard (1995), where restricting all of the Σ_i to be

the same results in drastic over-estimation of the covariance matrices for cells with much less variable Y's than average. This over-estimation can, for example, lead to substantial over-coverage of multiple-imputation confidence intervals (see Barnard (1995, Section 6.6.2)). We would like to have some flexibility in modeling the relationship among the cell covariances Σ_i, while at the same time keeping the number of unknowns reasonably small.

4.2. Extensions to GLOM

To relax the homogeneous covariance restriction of the GLOM model, Liu and Rubin (1998) present extensions to the GLOM model that allow a different but proportional covariance matrix across cells, i.e., $\Sigma_i = \lambda_i \Sigma$, where the geometric mean of the λ's is set to 1. This model assumes that the multivariate normals within each cell have the same ellipsoidal shape, but possibly different sizes. They also allow additional log-linear constraints on the proportionality coefficients λ_i, making it possible to put further structure in terms of covariates on the cell covariances. Liu and Rubin give details on how to get maximum likelihood parameter estimates using the EM algorithm, and how to conduct Bayesian inference using Markov chain Monte Carlo methods for their GLOM extensions.

Recently, Chiu, Leonard and Tsui (1996) presented a modeling approach for covariance matrices based on the matrix-logarithm transformation. They put linear models on the unique elements of the matrix-logarithm of a covariance matrix. This approach, while quite flexible, suffers from the same complication as the use of the matrix-logarithm transformation in prior specification (Leonard and Hsu (1992)) – the interpretability of the parameters resulting from the transformation. Placing sensible models on these parameters can be a demanding task for general users, particularly because one has to examine the hidden implications of the model choice on the parameters of interest.

Our extension to GLOM, which we call a general location-scale model, is more in the spirit of Liu and Rubin (1998). We allow cell-dependent vectors of standard deviations, but a common correlation matrix. Writing $\Sigma_i = \mathrm{diag}(S_i)\, R_i\, \mathrm{diag}(S_i)$, $i = 1, \ldots, c$, where R_i is the correlation matrix of Y and S_i is the vector of standard deviations of Y in cell i, our model assumes $R_i = R$ for all i, but leaves the S_i unrestricted. This extension of GLOM gives greater flexibility than that of Liu and Rubin (1998), but keeps the number of unknown parameters to a manageable size. If the assumption of a common correlation structure is too restrictive, the cells could be partitioned into a small number of groups (e.g., two or four based on some of the categorical variables), with a separate correlation matrix allowed for each group.

Structure could be imposed on the S_i's by putting linear regression models (with possibly common regression coefficients) on the $\log(S_i)$:

$$\log(S) = X\beta, \tag{14}$$

where $\log(S) \equiv (\log(S_1), \ldots, \log(S_c))^\top$ is a $(c \times k)$ matrix, X is a $c \times p$ design matrix, and β is the corresponding $p \times k$ coefficient matrix. Obtaining maximum likelihood estimates under this model, with either complete or incomplete data, will require iterative methods, such as the EM algorithm.

By combining equation (14) with distributional assumptions on the $\log(S_i)$, it is also possible to do (Bayesian) hierarchical modeling of the covariance structure through the vector of standard deviations, in addition to hierarchical modeling of the mean structures. In other words, we can improve not only the estimation of the mean-regression parameters but also the estimation of the standard deviation (or variance) parameters through shrinkage, an improvement that is particularly important when the data are sparse relatively to the number of cells. For example, with a common correlation matrix R across cells, we can use any prior discussed in Section 2.2 as $p(R)$ together with the following to specify a hierarchical model for the Σ_i: $\log(S_i) \mid X_{[i]}, \beta, \Lambda \sim N([X_{[i]}\beta]^\top, \Lambda)$, $p(\beta) \propto 1$, and $p(\Lambda) \propto |\Lambda|^{-(k+1)/2}$, where $X_{[i]}$ is the ith row of X. The computations for this model can be performed using the approach in Section 5.

5. Computational Details

There are many potential ways to do the computation. What we describe below is a relatively simple approach that has worked well for our examples. Since all of the models discussed in Section 3 and Section 4 can be expressed in a hierarchical structure, it is convenient to compute the posterior using the Gibbs sampler. We assume this is the case, so that we wish to draw from the distribution of (S, R) given the other model parameters and the data (i.e., what we describe below is to be incorporated into a larger computer program, which includes the simulation for, say, β.) In addition, we assume that we can compute this conditional posterior up to a proportionality constant.

To draw (S, R) we use the Gibbs sampler and draw each of the components of S and R one at a time. The drawing of a particular r_{ij} $(i > j)$ given the other correlations and S (as well as whatever other parameters are in the model) is complicated by the requirement that R be positive definite. We need to know what values of r_{ij} keep R positive definite given that the other correlations are held fixed.

Two observations enable us to solve this problem easily. Let us start with a correlation matrix R, which we know to be positive definite (as will be the case in

a Gibbs iteration). Let $R(r)$ be the matrix obtained from R by changing the i, j correlation to r and let $f(r) = |R(r)|$. Our first observation is that $f(r) > 0$ is a sufficient (and necessary) condition for $R(r)$ to be positive definite. To see this we first let $i = k$; recall R is $k \times k$. Let $R_m(r)$ be the submatrix of $R(r)$ obtained by selecting all correlations r_{uv} with $u \leq m$ and $v \leq m$, where m is an integer between 1 and k. Then $R(r)$ is positive definite if and only if $|R_m(r)| > 0$ for all m. Since R is positive definite, this will be true for $m = 1, \ldots, k-1$. Hence, we need only check that $|R_k(r)| = f(r) > 0$. By permuting rows and columns of $R(r)$ as necessary the result follows when $1 \leq i \leq k-1$.

Our second observation is that $f(r)$ is a quadratic function in r. Hence, the set of values r which correspond to a positive definite correlation matrix are those in the interval determined by the roots of the quadratic $f(r) = ar^2 + br + c$. The coefficients in $f(r)$ can be found using $a = [f(1) + f(-1) - 2f(0)]/2$, $b = [f(1) - f(-1)]/2$, and $c = f(0)$. Thus, for each draw of r_{ij} given the other correlations and S, we can easily determine the largest possible support of the distribution, which takes the form of an interval. The computational resources involved in obtaining the required determinants of R are minor in current computer environments, as long as k is not too large.

To actually make the draws we have used the griddy Gibbs sampler (Ritter and Tanner (1992)). We adopted this strategy because it is easy to program, not because it is efficient in terms of computing time. Each component of S has only the constraint that it be positive given the other components and R, so that implementation of the griddy Gibbs strategy is straightforward, although some experimentation may be needed to choose a grid. Choosing a grid from which to draw r_{ij} given the other parameters is relatively simple. We first determined the interval of values which preserve the positive definiteness of R as discussed above. We then put down a grid of equally spaced values. Usually in application we are only interested in the first two digits of a correlation so that choosing a grid with 100 points is more than adequate and results in fast computation.

In general, this griddy Gibbs strategy can be used to draw from the posterior distribution of Σ given prior distributions other than those discussed in Section 2, by first transforming from Σ to (S, R) and then using the above griddy Gibbs strategy, assuming, of course, the resulting Markov chain satisfies the standard regularity conditions (see, e.g., Roberts (1996), Tierney (1996)). We only assume that we can (conditionally) evaluate the posterior distribution up to a proportionality constant. The use of the Gibbs sampler also allows easy adaptation of our computational strategy for missing data problems simply by adding another Gibbs step, which draws from the conditional density of the missing data given the observed data and the model, including its parameters.

Our one-at-a-time Gibbs strategy is also useful for drawing from the joint uniform prior $p(R) \propto 1$. We do this by drawing each correlation given the others under the prior distribution. Each draw is uniform on the interval determined by the roots of the quadratic as discussed above. Clearly, many other prior distributions on Σ could be drawn from using a similar strategy, assuming the prior distributions can be evaluated up to a proportionality constant. For some prior distributions, however, transformations of R may be needed before implementing the Gibbs strategy. For example, we have found that directly implementing the one-at-a-time Gibbs sampler for the marginally uniform prior is not effective, as the chain can stay in the corners of the correlation space $\mathcal{R}^k$ space (see Figure 2) for a long time.

The use of the griddy Gibbs method makes implementation (coding) of our approach relatively straightforward (our code was written in C++). There is the danger, however, that drawing all of the elements of S and R one at a time will make the resulting Markov chain slow to converge and highly dependent. Our limited experience suggests that this may not be as serious as one might expect. We have found that the resulting Markov chains are able to move reasonably quickly and produce draws with low to moderate autocorrelations. For example, to assess convergence for the simulated example in Section 3.2, we ran five independent Markov chains, using "disperse" starting values, and calculated Gelman and Rubin's (1992) potential scale reduction, $\sqrt{\hat{R}}$, for consecutive iterations beginning at iteration 10 and ending at iteration 200 for a variety of parameters. Figure 12 gives plots of the reductions versus iteration number. For all parameters (including those not graphed), by iteration 150 (which includes a burn-in of 75 iterations) $\sqrt{\hat{R}}$ is close to 1.0, i.e., below 1.1, producing evidence of acceptable convergence behavior of these chains.

Although we found the one-at-a-time Gibbs strategy coupled with griddy Gibbs to be a useful computational approach for obtaining posterior quantities, we also explored a Metropolis-Hastings strategy with an independence jumping distribution in which we used the posterior of Σ under the conjugate inverse-Wishart prior as the jumping distribution for generating trial draws of Σ (giving draws of S and R). This alternative strategy, which seems quite natural, turned out to be not useful in our applications – the resulting Markov chain often got stuck for lengthy periods, giving a strong indication of the large differences between our priors and the standard inverse-Wishart prior. More flexible choices of jumping distribution, such as the posteriors of Σ under the priors of Leonard and Hsu (1992), may provide attractive alternatives to our Gibbs strategy. However, we have not explored such options.

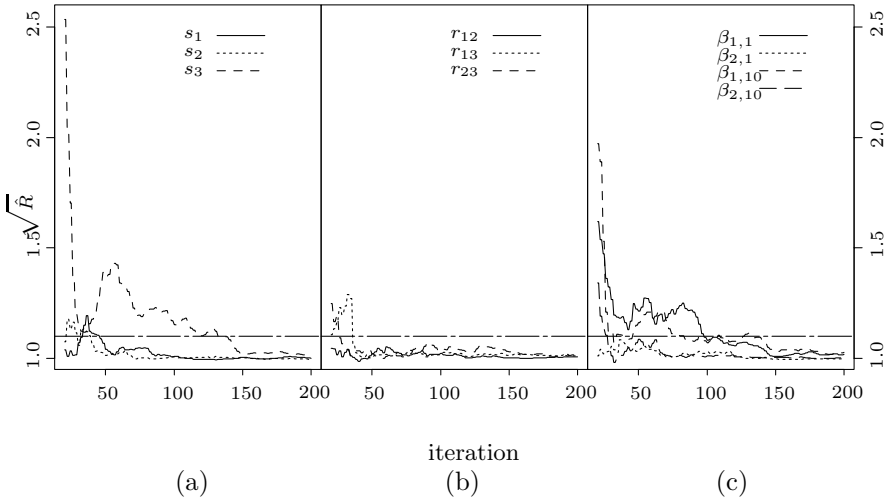


Figure 12. Gelman and Rubin's (1992) potential scale reduction, $\sqrt{\hat{R}}$, for a variety of parameters ((a) standard deviations, (b) correlations, and (c) selected regression coefficients) from the simulation example in Section 3.2 plotted against iteration number. Reductions are based on consecutive draws from five independent chains (first reductions calculated at iteration 10) and are calculated from the second half of the corresponding chains.

6. Summary

Modeling a covariance matrix in terms of the standard deviations and correlation matrix is a common strategy when the positive definite constraint for the correlation matrix is easy to deal with, such as with bivariate variables (e.g., Jeffreys (1983), Zellner (1979), Gelman, Carlin, Stern and Rubin (1995)). In this paper we show that the strategy can be applied in a much more general setting. We provide empirical evidences to demonstrate that this strategy has much to offer in the context of prior elicitation. The strategy is applicable in general because the positive definite constraint can be easily handled via a Gibbs-sampler formulation, that is, one correlation at a time. We also give an important application in a non-Bayesian setting, namely, we propose the general location-scale model to relax the restrictive common-covariance assumption of the popular general location model for simultaneously modeling continuous and categorical variables. There are many other applications of this approach, such as prior specification of a covariance matrix in multinomial probit model (e.g., McCulloch and Rossi (1998)) and Bayesian analysis of seemingly unrelated regression (Zellner (1962)).

We do not claim that the computational method we used in this paper is the best possible one, albeit we did find it easy to use and reliable compared to other methods we have tried. We also do not claim that the separation strategy

is universally applicable (e.g., when rotation invariance is desirable). We do feel, however, that the separation strategy, given it is directly motivated by statistical considerations, is likely to be a more effective tool in certain contexts for applied statisticians than current common methods such as inverse-Wishart or those based on various mathematical decompositions. In fact our approach has already been applied by practioners in hierarchical Bayesian modeling (e.g. Brav (2000)). We thus hope our work will stimulate more applications of and research on the separation strategy, including findings on its limitations.

Acknowledgement

The authors thank Jim Berger, Andrew Gelman, Adrian Raftery, Peter Rossi and reviewers for very helpful comments. We also thank Moshen Pourahmadi for preprints and reprints of his papers. Meng's research was supported in part by NSF grants DMS-9505043 and DMS-9626691, and in part by NSA grant MDA 904-96-1-0007. Barnard's research was supported in part by NSF grant DMS-9705158.

References

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gussian and non-Gaussian clustering. *Biometrics* **49**, 803-821.

Barnard, J. (1995). Cross-Match Procedures for Multiple Imputation Inference: Bayesian Theory and Frequentist Evaluation. Ph.D. thesis, Dept. of Statistics, University of Chicago.

Bensmail, H. and Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *J. Amer. Statist. Assoc.* **91**, 1743-1748.

Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997). Inference in model-based cluster analysis. *Statist. Comput.* **7**, 1-10.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. John Wiley, New York.

Brav, A. (2000). Inference in long-horizon event studies: A Bayesian approach with application to initial public offerings. *J. Finance*. To appear.

Brealy, R. and Myers, S. (1984). *Principle of Corporate Finance*. McGraw-Hill, New York.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *J. Pattern Recognition Soc.* **28**, 781-793.

Chiu, T. Y., Leonard, T. and Tsui, K.-W. (1996). The matrix-logarthimic covariance model. *J. Amer. Statist. Assoc.* **91**, 198-210.

Dryden, I. and Mardia, K. V. (1997). *Statistical Shape Analysis*. John Wiley, New York.

Fama, E. (1976). *Foundations of Finance*. Basic Books, New York.

Gelman, A. (1996). Bayesian model-building by pure thought: Some principles and examples. *Statist. Sinica* **6**, 215-232.

Gelman, A., Bois, F. and Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Amer. Statist. Assoc.* **91**, 1400-1412.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

Gelman, A. and Meng, X.-L. (1996). Model checking and model improvement. In *Practical Markov Chain Monte Carlo* (Edited by W. Gilks, S. Richardson and D. Spiegelhalter), 189-201. Chapman and Hall, London.

Gelman, A., Meng, X.-L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica* **6**, 733-807.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7**, 457-511.

Jeffreys, H. (1983). *Theory of Probability.* Oxford University Press, New York.

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics* **36**, 493-499.

Krzanowski, W. J. (1982). Mixtures of continuous and categorical variables in discriminant analysis: A hypothesis-testing approach. *Biometrics* **38**, 991-1002.

Le, N. D., Martin, R. D. and Raftery, A. E. (1996). Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models. *J. Amer. Statist. Assoc.* **91**, 1504-1515.

Leonard, T. and Hsu, J. S. J. (1992). Bayesian inference for a covariance matrix. *Ann. Statist.* **20**, 1669-1696.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 1-41.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* John Wiley, New York.

Little, R. J. A. and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**, 497-512.

Liu, C. (1993). Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data. *J. Multivariate Anal.* **46**, 198-206.

Liu, C. and Rubin, D. B. (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika* **85**, 673-688.

McCulloch, R. and Rossi, P. (1998). Bayesian analysis of the multinomial probit model. In *Simulation-Based Inference in Econometrics: Methods and Applications* (Edited by T. Schuermann, M. Weeks and R. Mariano). Cambridge University Press, Cambridge.

Meng, X. L. (1994). Multiple imputation with uncongenial sources of input (with discussion). *Statist. Sci.* **9**, 538-573.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables (corr: V.36 p.343). *Ann. Math. Statist.* **32**, 448-465.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statist. Comput.* **6**, 289-296.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterization. *Biometrika* **86**, 677-690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika* **87**. To appear.

Raghunathan, T. E. and Grizzle, J. E. (1995). A split questionnaire survey design. *J. Amer. Statist. Assoc.* **90**, 54-63.

Ramsey, F. L. (1974). Characterization of the partial autocorrelation function. *Ann. Statist.* **2**, 1296-1301.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *J. Amer. Statist. Assoc.* **87**, 861-868.

Roberts, G. (1996). Markov chain concepts related to sampling algorithms. In *Practical Markov Chain Monte Carlo* (Edited by W. Gilks, S. Richardson and D. Spiegelhalter), 45-57. Chapman and Hall, London.

Rousseeuw, P. J. and Molenberghs, G. (1994). The shape of correlation matrices. *Amer. Statist.* **48**, 276-279.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12**, 1151-1172.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley, New York.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *J. Amer. Statist. Assoc.* **91**, 473-489.

Schafer, J. L. (1991). Algorithms for Multiple Imputation and Posterior Simulation from Incomplete Multivariate Data with Ignorable Nonresponse. Ph.D. thesis, Department of Statistics, Harvard University.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data by Simulation*. Chapman and Hall, New York.

Schafer, J. L., Khare, M. and Ezzati-Rice, T. M. (1993). Multiple imputation of missing data in NHANES III (disc: P.502-510). In *Proceedings of the Bureau of the Census Annual Research Conference*, 459-487.

Stevens, R. (1996). Three Essays in the Selection and Estimation of Factor Models for Returns Data. Ph.D. thesis, Graduate School of Business, University of Chicago.

Sun, D. and Berger, J. (1998). Reference priors with partial information. *Biometrika* **85**, 55-72.

Tan, W.-Y. (1969). Bayesian estimation of a multivariate covariance model when the covariables are uncontrollable. Technical Report No. 198, Department of Statistics, University of Wisconsin.

Tierney, L. (1996). Introduction to general state-space Markov chain theory. In *Practical Markov Chain Monte Carlo* (Edited by W. Gilks, S. Richardson and D. Spiegelhalter), 59-74. Chapman and Hall, London.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22**, 1195-1211.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57**, 348-368.

Zellner, A. (1979). An error-components procedure (ECP) for introducing prior information about covariance matrices and analysis of multivariate regression models. *Internat. Econom. Rev.* **20**, 679-692.

Zellner, A. (1991). Bayesian methods and entropy in economics and econometrics. In *Maximum Entropy and Bayesian Methods* (Edited by W. Grandy and L. Schick), 17-31. Kluwer Academic, The Netherlands.

Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

E-mail: barnard@stat.harvard.edu

Graduate School of Business, The University of Chicago, Chicago, IL 60637, U.S.A.

E-mail: robert.mcculloch@gsb.uchicago.edu

Department of Statistics, The Unversity of Chicago, Chicago, IL 60637, U.S.A.

E-mail: meng@galton.uchicago.edu

(Received October 1999; accepted March 2000)