MCMC Estimation of Restricted Covariance Matrices

Joshua Chi-Chun Chan*
University of Queensland

Ivan Jeliazkov†
University of California, Irvine

November 19, 2009

Abstract

This article is motivated by the difficulty of applying standard simulation techniques when identification constraints or theoretical considerations induce covariance restrictions in multivariate models. To deal with this difficulty, we build upon a decomposition of positive definite matrices and show that it leads to straightforward Markov chain Monte Carlo samplers for restricted covariance matrices. We introduce the approach by reviewing results for multivariate Gaussian models without restrictions, where standard conjugate priors on the elements of the decomposition induce the usual Wishart distribution on the precision matrix and vice versa. The unrestricted case provides guidance for constructing efficient Metropolis-Hastings and accept-reject Metropolis-Hastings samplers in more complex settings, and we describe in detail how simulation can be performed under several important constraints. The proposed approach is illustrated in a simulation study and two applications in economics. Supplemental materials for this article (appendices, data, and computer code) are available online.

Keywords: Accept-reject Metropolis-Hastings algorithm; Bayesian estimation; Cholesky decomposition; Correlation matrix; Markov chain Monte Carlo; Metropolis-Hastings algorithm; Multinomial probit; Multivariate probit; Unconstrained parameterization; Wishart distribution.

1 Introduction

Theoretical and practical considerations frequently motivate the imposition of restrictions on covariance structures in a variety of multivariate statistical models, such as models for binary and ordinal outcomes, simultaneous equation systems, Gaussian copulas, or models with incidental truncation or endogenous treatment indicators. For example, multinomial probit (MNP) models are usually estimated subject to the identification constraint that a diagonal element of the covariance matrix (usually the first) is restricted to one, while multivariate probit (MVP), multivariate ordinal probit (MOP), and Gaussian copula models are identified by requiring that the covariance matrix is in correlation form, *i.e.*, all diagonal elements are equal to one. Off-diagonal restrictions appear frequently in empirical work as well, such as in systems of simultaneous equations, graphical models, structural vector autoregressions, or in circumstances where parsimony may be desirable – for instance, when the dimension of the covariance matrix is large relative to the sample size.

*Department of Mathematics, University of Queensland, Brisbane, QLD 4072, Australia. E-mail: chancc@maths.uq.edu.au. This author's research was supported by the Australian Research Council (Discovery Grant DP0558957).

†Department of Economics, University of California, Irvine, 3151 Social Science Plaza, Irvine, CA 92697-5100. E-mail: ivan@uci.edu.

Constructing Markov chain Monte Carlo (MCMC) samplers for models with covariance restrictions, however, is non-trivial owing to the non-standard form of the resulting conditional densities and the positive definiteness requirement. Advances have been made in imposing a constraint on a single diagonal element of the covariance matrix as in MNP models (McCulloch et al., 2000; Nobile, 2000; Imai and van Dyk, 2005), treatment models (Munkin and Trivedi, 2003; Chib, 2007), and incidental truncation models (Chib et al., 2009). Unfortunately, those techniques are not readily extendable to more general cases where additional diagonal or off-diagonal elements are constrained. When such complications are present, samplers generally require a Metropolis-Hastings (MH) step in which the specification of a suitable proposal density plays a crucial role. A number of MH simulation approaches have been suggested in this context. To draw the elements of a correlation matrix in MVP models, Chib and Greenberg (1998) use independence or random walk MH chains, while Liu and Daniels (2006) consider a reparameterization in which a covariance matrix, drawn from an inverse Wishart distribution, is subsequently translated to a correlation matrix that is passed to an MH step. One-at-a-time MH sampling of the components of a correlation matrix has been implemented in Gaussian copula models by Pitt et al. (2006) using the parameterization of Wong et al. (2003), and in hierarchical shrinkage models by Barnard et al. (2000) using the griddy Gibbs sampler (Ritter and Tanner, 1992). In the context of Gaussian graphical models, Atay-Kayis and Massam (2005) and Carvalho et al. (2007) discuss estimation of precision matrices with off-diagonal zero constraints implied by the graph of the model. Everson and Morris (2000) use accept-reject sampling to simulate draws from Wishart distributions with eigenvalue restrictions.

In this article we study the applicability of a particular decomposition of the covariance matrix that can accommodate both diagonal and off-diagonal restrictions. Specifically, since the covariance matrix $\boldsymbol{\Sigma}$ is positive definite, one can uniquely factor it by a modified Cholesky decomposition $\boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}' = \boldsymbol{D}$, or equivalently $\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}'\boldsymbol{D}^{-1}\boldsymbol{L}$, where $\boldsymbol{L}$ is a lower triangular matrix with ones on the main diagonal and $\boldsymbol{D}$ a diagonal matrix with positive diagonal elements. This parameterization is appealing because the free elements in $\boldsymbol{L}$ are unrestricted, while the positivity of the diagonal elements of $\boldsymbol{D}$ is easy to check and impose. Because of these features, the decomposition has recently been employed in maximum likelihood estimation (Pourahmadi, 1999, 2000, 2007) and covariance matrix modeling through partial autocorrelations (Daniels and Pourahmadi, 2008). The decomposition is useful in Bayesian estimation since it produces the usual Wishart conjugate sampler in the unrestricted case – in Section 2 we show that simple conjugate priors on $\boldsymbol{D}$ and $\boldsymbol{L}$ induce the usual Wishart prior on the precision matrix $\boldsymbol{\Sigma}^{-1}$ (and vice versa, since the decomposition is one-to-one). These results offer theoretical continuity as well as means for incorporating additional flexibility in prior modeling by assuming other prior distributions for the elements of the decomposition so as to yield classes of priors beyond the Wishart family. The desirability of such extensions was advocated by Leonard and Hsu (1992) who modeled the matrix logarithm of the covariance matrix. More importantly, however, the parameterization allows for straightforward Bayesian estimation in a number of important cases involving restrictions on $\boldsymbol{\Sigma}$. Specifically, the method provides a natural proposal density for MH or accept-reject Metropolis-Hastings (ARMH) sampling (Tierney, 1994; Chib and Greenberg, 1995; Chib and Jeliazkov, 2005) that in many instances may minimize, or entirely eliminate, the costs of tailoring by constrained optimization. (Appendix B in the online supplemental materials offers details on the ARMH algorithm.) To illustrate the proposed approach, we construct Markov chain samplers for models involving the following restrictions: (i) the first diagonal element of $\boldsymbol{\Sigma}$ is one, (ii) all diagonal elements are ones, and (iii) any off-diagonal elements are zeros.

The rest of this article is organized as follows. In Section 2, we present the conjugate priors on the decomposition matrices and show that they induce a Wishart prior on the precision matrix. The details of the proposed sampler are discussed and it is shown that the unrestricted case requires

only direct sampling from known densities. The focus of Section 3 is on the handling of various covariance matrix restrictions, using the unrestricted case as a guide in constructing efficient MH or ARMH Markov chains. Section 4 illustrates the proposed method with simulated data experiments, and Section 5 provides two real data applications concerning women's labor force participation and commuters' scheduling of work trips. Section 6 offers brief concluding remarks.

2 An Alternative Parameterization of the Covariance Matrix

Suppose that we have N independent and identically distributed observations from the p-dimensional normal distribution $\boldsymbol{u}_i \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, $i = 1, \ldots, N$, and the goal is to estimate the covariance matrix $\boldsymbol{\Sigma}$, which might involve certain diagonal or off-diagonal restrictions. Since $\boldsymbol{\Sigma}$ is positive definite, there exist unique matrices $\boldsymbol{L}$ and $\boldsymbol{D}$ such that $\boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}' = \boldsymbol{D}$ or equivalently $\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}'\boldsymbol{D}^{-1}\boldsymbol{L}$, where $\boldsymbol{L}$ is a lower triangular matrix with ones on the diagonal and $\boldsymbol{D}$ a diagonal matrix with positive diagonal elements (for a proof from first principles, see Golub and van Loan, 1983). A number of other Cholesky-type decompositions have been used (see, for example, Pourahmadi, 2007, and the references therein), but the one considered here lends itself well to simulation-based estimation. To establish notation, let λ_k, $k = 1, \ldots, p$ denote the diagonal entries of $\boldsymbol{D}$ and let $a_{kj}, 1 \leqslant j < k \leqslant p$ denote the free elements of the lower unitriangular matrix $\boldsymbol{L}$, *i.e.*,

$$\boldsymbol{D} \equiv \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}, \; \boldsymbol{L} \equiv \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 1 & \cdots & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ a_{p1} & a_{p2} & & \cdots & 1 \end{pmatrix}.$$

Also define $\boldsymbol{\lambda} \equiv (\lambda_1, \ldots, \lambda_p)'$, $\boldsymbol{a}_k \equiv (a_{k1}, \ldots, a_{k,k-1})'$, $k = 2, \ldots, p$, and $\boldsymbol{a} \equiv (\boldsymbol{a}_2', \ldots, \boldsymbol{a}_p')'$. The notation $\boldsymbol{a}$ and $\boldsymbol{L}$ (similarly $\boldsymbol{\lambda}$ and $\boldsymbol{D}$) will be used interchangeably in the rest of the article. With this parameterization, consider the priors

$$\lambda_k \overset{ind}{\sim} \mathcal{IG}\left((\nu + k - p)/2, 1/2\right), \quad \nu > p, \quad k = 1, \ldots, p, \tag{1}$$

and

$$\boldsymbol{a}_k | \lambda_k \overset{ind}{\sim} \mathcal{N}(\boldsymbol{0}, \lambda_k \boldsymbol{I}_{k-1}), \quad k = 2, \ldots, p, \tag{2}$$

where $\mathcal{IG}(\cdot, \cdot)$ is the inverse gamma distribution. We then have the following result.

Theorem 1 *Under the priors in (1) and (2), $\boldsymbol{\Sigma}^{-1} \equiv \boldsymbol{L}'\boldsymbol{D}^{-1}\boldsymbol{L}$ has a Wishart distribution $\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\nu, \boldsymbol{I}_p)$. Moreover, the converse is also true. If $\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\nu, \boldsymbol{I}_p)$, then the induced priors on λ_k, $k = 1, \ldots, p$ and a_{kj}, $1 \leqslant j < k \leqslant p$ are (1) and (2), respectively.*

The theorem can be derived from the Bartlett decomposition, and a detailed proof is provided in Appendix A in the supplemental materials. This result provides an important equivalence relation for the case where $\boldsymbol{\Sigma}$ is unrestricted. It also offers several straightforward extensions. The first of these follows from the re-scaling property of the Wishart distribution under the more general priors

$$\lambda_k \overset{ind}{\sim} \mathcal{IG}\left(\nu_{k0}/2, \delta_{k0}/2\right), \quad k = 1, \ldots, p, \tag{3}$$

and

$$\boldsymbol{a} | \boldsymbol{\lambda} \sim \mathcal{N}\left(\boldsymbol{a}_0, \boldsymbol{A}_0\right), \tag{4}$$

where the elements of $\boldsymbol{a}_0$ need not necessarily be zero, and those of $\boldsymbol{A}_0$ are allowed to be freely correlated and can possibly depend on $\boldsymbol{\lambda}$. Then, we have the following result.

Corollary 1 *As a special case, the priors in* (3) *and* (4) *will induce a Wishart distribution for the matrix* $\boldsymbol{\Sigma}^{-1} \equiv \boldsymbol{L}'\boldsymbol{D}^{-1}\boldsymbol{L}$, $\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\nu, \boldsymbol{R})$ *with an arbitrary* $p \times p$ *symmetric positive definite scale matrix* $\boldsymbol{R}$. *More general hyperparameter settings in* (3) *and* (4) *will result in distributions for* $\boldsymbol{\Sigma}^{-1}$ *beyond the Wishart class.*

A detailed proof of the corollary is given in Appendix A. While those derivations show that the full spectrum of Wishart results can be recovered in this framework and that some additional generality is possible in the modeling of the precision matrix, an important point to note is that in principle one can place alternative, possibly non-conjugate, priors on $\boldsymbol{\lambda}$ and $\boldsymbol{a}$ to achieve yet more flexible distributions for the resultant covariance matrix. Extensions are possible by specifying other priors on the components of the decomposition, *e.g.* $\boldsymbol{\lambda}$ can be distributed as gamma, lognormal, or log-t (the last two of which also allow correlation among the elements of $\boldsymbol{\lambda}$); similarly, $\boldsymbol{a}$ can be modeled more flexibly by mixtures or scale-mixtures of normals, which include the Student-t and logistic distributions, among others.

In addition to the aforementioned results, parameterization in terms of $\boldsymbol{a}$ and $\boldsymbol{\lambda}$ can also be shown to be, in fact, conjugate with respect to a multivariate Gaussian likelihood. Upon defining $\boldsymbol{w}_i \equiv \boldsymbol{L}\boldsymbol{u}_i$ and $\boldsymbol{u} = (\boldsymbol{u}_1', \ldots, \boldsymbol{u}_N')'$, and recognizing that $|\boldsymbol{\Sigma}^{-1}| = |\boldsymbol{L}'||\boldsymbol{D}^{-1}||\boldsymbol{L}| = |\boldsymbol{D}|^{-1} = \prod_{k=1}^{p} \lambda_k^{-1}$, the likelihood can be written as

$$\begin{aligned}
\ell(\boldsymbol{u}|\boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{-N/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N} \boldsymbol{u}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{u}_i\right\} \\
&= \left(\prod_{k=1}^{p} \lambda_k^{-N/2}\right) \exp\left\{-\frac{1}{2}\sum_{i=1}^{N} \boldsymbol{w}_i'\boldsymbol{D}^{-1}\boldsymbol{w}_i\right\} \\
&= \left(\prod_{k=1}^{p} \lambda_k^{-N/2}\right) \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{D}^{-1}\sum_{i=1}^{N} \boldsymbol{w}_i\boldsymbol{w}_i')\right\} \\
&= \prod_{k=1}^{p} \lambda_k^{-N/2} \exp\left\{-\frac{s_k}{2\lambda_k}\right\},
\end{aligned}$$

where s_k is the (k,k)-element of $\sum_{i=1}^{N} \boldsymbol{w}_i\boldsymbol{w}_i'$. Thus, under the inverse gamma prior in (3), for $k = 1, \ldots, p$, the full conditional distributions $\lambda_k|\boldsymbol{u}, \boldsymbol{a}$ are also independent inverse gamma

$$\lambda_k|\boldsymbol{u}, \boldsymbol{a} \overset{ind}{\sim} \mathcal{IG}\left(\frac{\nu_{k0} + N}{2}, \frac{\delta_{k0} + s_k}{2}\right). \tag{5}$$

To obtain the full conditional density for $\boldsymbol{a}$, observe that the elements a_{ij} enter the likelihood as the coefficients in the regressions of u_{ij} on the negative values of $\{u_{ik}\}_{k<j}$ for $j = 2, \ldots, p$. To see this, note that since

$$\boldsymbol{L}\boldsymbol{u}_i = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 1 & \cdots & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ a_{p1} & a_{p2} & a_{p3} & \cdots & 1 \end{pmatrix} \begin{pmatrix} u_{i1} \\ u_{i2} \\ u_{i3} \\ \vdots \\ u_{ip} \end{pmatrix} = \begin{pmatrix} u_{i1} \\ u_{i2} + a_{21}u_{i1} \\ u_{i3} + a_{31}u_{i1} + a_{32}u_{i2} \\ \vdots \\ u_{ip} + \sum_{k=1}^{p-1} a_{pk}u_{ik} \end{pmatrix},$$

we can rewrite the likelihood as

$$\begin{aligned}\ell(\boldsymbol{u}|\boldsymbol{\Sigma}) &\propto \left(\prod_{k=1}^{p}\lambda_k^{-N/2}\right)\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{L}\boldsymbol{u}_i)'\boldsymbol{D}^{-1}(\boldsymbol{L}\boldsymbol{u}_i)\right\}\\ &= \left(\prod_{k=1}^{p}\lambda_k^{-N/2}\right)\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\frac{u_{i1}^2}{\lambda_1}\right\}\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\frac{(u_{i2}+a_{21}u_{i1})^2}{\lambda_2}\right\}\cdots\\ &\qquad\cdots\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\frac{(u_{ip}+\sum_{k=1}^{p-1}a_{pk}u_{ik})^2}{\lambda_p}\right\}, \quad (6)\end{aligned}$$

which can in turn be written in a more familiar form in which $\boldsymbol{a}$ enters as a vector of regression coefficients

$$\ell(\boldsymbol{u}|\boldsymbol{\Sigma}) \propto \left(\prod_{k=1}^{p}\lambda_k^{-N/2}\right)\exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{u}_i-\boldsymbol{U}_i\boldsymbol{a})'\boldsymbol{D}^{-1}(\boldsymbol{u}_i-\boldsymbol{U}_i\boldsymbol{a})\right\}, \quad (7)$$

with (note the negative sign)

$$\boldsymbol{U}_i = -\begin{pmatrix} 0 & \cdots & & & & & & \cdots & 0\\ u_{i1} & 0 & \cdots & & & & & & \vdots\\ 0 & u_{i1} & u_{i2} & 0 & \cdots & & & & 0\\ 0 & \cdots & 0 & u_{i1} & u_{i2} & u_{i3} & 0 & \cdots & \vdots\\ \vdots & \cdots & & \ddots & & & \ddots & & 0\\ 0 & \cdots & & 0 & & \cdots & 0 & u_{i1} & \cdots & u_{ip}\end{pmatrix}.$$

Therefore, given the prior in (4), the full-conditional distribution becomes

$$\boldsymbol{a}|\boldsymbol{u},\boldsymbol{\lambda} \sim \mathcal{N}(\widehat{\boldsymbol{a}},\widehat{\boldsymbol{A}}), \quad (8)$$

with $\widehat{\boldsymbol{A}} = \left(\boldsymbol{A}_0^{-1}+\sum_{i=1}^{N}\boldsymbol{U}_i\boldsymbol{D}^{-1}\boldsymbol{U}_i\right)^{-1}$ and $\widehat{\boldsymbol{a}} = \widehat{\boldsymbol{A}}\left(\boldsymbol{A}_0^{-1}\boldsymbol{a}_0+\sum_{i=1}^{N}\boldsymbol{U}_i\boldsymbol{D}^{-1}\boldsymbol{u}_i\right)$.

Because $\boldsymbol{D}$ is diagonal, if $\boldsymbol{A}_0$ happens to be diagonal or block-diagonal for the rows of $\boldsymbol{L}$, *i.e.*

$$\boldsymbol{a}_k \overset{ind}{\sim} \mathcal{N}(\boldsymbol{a}_{k0},\boldsymbol{A}_{k0}),\ k=2,\ldots,p, \quad (9)$$

the derivations are simplified and the elements of $\boldsymbol{a}$ can be updated in a series of independent steps

$$\boldsymbol{a}_k|\boldsymbol{u},\lambda_k \overset{ind}{\sim} \mathcal{N}_k(\widehat{\boldsymbol{a}}_k,\widehat{\boldsymbol{A}}_k),\ k=2,\ldots,p, \quad (10)$$

where, keeping in mind that $\boldsymbol{A}_{k0}$ can possibly depend on λ_k, we have

$$\begin{aligned}\widehat{\boldsymbol{A}}_k &= \left(\boldsymbol{A}_{k0}^{-1}+\lambda_k^{-1}\boldsymbol{X}_k'\boldsymbol{X}_k\right)^{-1},\ \widehat{\boldsymbol{a}}_k = \widehat{\boldsymbol{A}}_k\left(\boldsymbol{A}_{k0}^{-1}\boldsymbol{a}_{k0}+\lambda_k^{-1}\boldsymbol{X}_k'\boldsymbol{z}_k\right),\\ \boldsymbol{z}_k &= (u_{1k},\ldots,u_{Nk})',\ \boldsymbol{X}_k = [\boldsymbol{z}_1:\cdots:\boldsymbol{z}_{k-1}].\end{aligned}$$

Given posterior draws $\boldsymbol{a}$ and $\boldsymbol{\lambda}$ (equivalently $\boldsymbol{L}$ and $\boldsymbol{D}$) from (5) and (8) (or from (5) and (10) if $\boldsymbol{A}_0$ is diagonal or block diagonal), a posterior draw of $\boldsymbol{\Sigma}$ can be obtained simply by computing $\boldsymbol{\Sigma} = \boldsymbol{L}^{-1}\boldsymbol{D}(\boldsymbol{L}^{-1})'$.

In this section we have considered a parameterization of $\boldsymbol{\Sigma}^{-1}$ that can be used as an alternative to, or an extension upon, Wishart modeling. We have also presented full-conditional distributions for the elements of the decomposition that can be used to form a straightforward Gibbs sampler in this setting. In the next section, we turn to evaluating the usefulness of this representation for estimating restricted versions of $\boldsymbol{\Sigma}$. While direct sampling is still possible in some limited cases, the results presented thus far serve as a helpful guide in constructing straightforward and efficient Metropolis-Hastings or accept-reject Metropolis-Hastings algorithms for more general problems.

3 Restrictions on the Covariance Matrix

An important advantage of the proposed approach lies in its ability to accommodate various covariance restrictions, both on and off the main diagonal, that are often found in empirical studies. Here we discuss how this can be done within the framework discussed in Section 2. While a restricted precision matrix can no longer be modeled through a Wishart prior, we emphasize that prior information can still be incorporated into the analysis through priors on $\boldsymbol{\lambda}$ and $\boldsymbol{a}$, although such priors must depart from the unconstrained versions presented in Section 2 to reflect the desired restrictions on the matrix. However, we also emphasize that in each of the following cases the required modifications are conceptually straightforward.

In order to be able to deal with covariance restrictions, it is convenient to express each element of $\boldsymbol{\Sigma}$ in terms of elements of $\boldsymbol{D}$ and $\boldsymbol{L}$. By definition, $\boldsymbol{\Sigma} = \boldsymbol{L}^{-1}\boldsymbol{D}(\boldsymbol{L}^{-1})'$ and therefore

$$\sigma_{kk} = \lambda_k + \sum_{j=1}^{k-1} \lambda_j (a^{kj})^2, \quad k = 1, \ldots, p, \tag{11}$$

$$\sigma_{kj} = a^{kj}\lambda_j + \sum_{h=1}^{j-1} a^{kh} a^{jh} \lambda_h, \quad 1 \leqslant j < k \leqslant p, \tag{12}$$

where $\{a^{kj}\}$ are the lower diagonal elements of $\boldsymbol{L}^{-1}$ and σ_{kj} is the (k, j) element of $\boldsymbol{\Sigma}$.

3.1 The first diagonal element is restricted to one

This restriction is easy to impose: we only need to observe that by (11), $\sigma_{11} = \lambda_1$. Thus the condition that the first diagonal element equals one, *i.e.* $\sigma_{11} = 1$ is the same as restricting $\lambda_1 = 1$, or equivalently $\Pr(\lambda_1 = 1) = 1$. For other diagonal elements λ_k, $k = 2, \ldots, p$, the full-conditional densities are as in (5), whereas the full-conditional density for $\boldsymbol{a}$ is as in (8) or (10). Hence, this case only requires Gibbs sampling with direct simulation from known densities, similarly to approaches using a decomposition of the inverse Wishart distribution such as in Dreze and Richard (1983), McCulloch et al. (2000), Munkin and Trivedi (2003), and Chib et al. (2009).

3.2 All diagonal elements are restricted to ones

To impose the condition that all diagonal elements are ones, we first observe that from (11) it follows that $\sigma_{11} = \cdots = \sigma_{pp} = 1$ if and only if λ_i satisfies the recursive equations:

$$\lambda_1 = 1, \tag{13}$$

$$\lambda_k = 1 - \sum_{j=1}^{k-1} (a^{kj})^2 \lambda_j, \quad k = 2, \ldots, p. \tag{14}$$

These equations reveal that under the restriction $\sigma_{11} = \cdots = \sigma_{pp} = 1$, $\boldsymbol{\lambda}$ is a deterministic function of $\boldsymbol{a}$ and its role is purely notational. It is simply a working parameter that will be used to keep the notation concise and consistent throughout the paper and will also serve to simplify the derivation of a reasonable MH proposal density below. Formally, the joint prior for $\boldsymbol{\lambda}$ and $\boldsymbol{a}$ takes the form

$$p(\boldsymbol{\lambda}, \boldsymbol{a}) \propto p(\boldsymbol{a}) p(\boldsymbol{\lambda}|\boldsymbol{a}) I(\boldsymbol{\lambda} > 0), \tag{15}$$

where $p(\boldsymbol{a})$ is the density implied by either (4) or (9) depending on the prior specification, $p(\boldsymbol{\lambda}|\boldsymbol{a})$ takes the values defined recursively by (13) and (14) with probability 1, and $I(\cdot)$ is the indicator

function. This representation guarantees that all elements of $\boldsymbol{\lambda}$ are positive and satisfy (13) and (14), so that the resulting $\boldsymbol{\Sigma}$ is both positive definite and in correlation form.

With the restriction $\sigma_{11} = \cdots = \sigma_{pp} = 1$ imposed, the conditional density $\boldsymbol{a}|\boldsymbol{u}$ does not belong to a known family and a Metropolis-Hastings step is required. However, the results of Section 2 suggest a reasonable way to proceed. By analogy with (8) and (10), we consider MH proposal densities of the form

$$f(\boldsymbol{a}|\boldsymbol{u}) = f_T(\boldsymbol{a}|\boldsymbol{\mu}, \tau\boldsymbol{V}, \kappa), \tag{16}$$

and

$$f(\boldsymbol{a}_k|\boldsymbol{u}) = f_T(\boldsymbol{a}_k|\boldsymbol{\mu}_k, \tau\boldsymbol{V}_k, \kappa), \quad k = 2, \ldots, p, \tag{17}$$

where $f_T(\cdot)$ is the multivariate-t density with mean $\boldsymbol{\mu}$, scale matrix $\tau\boldsymbol{V}$ with tuning parameter τ, and κ degrees of freedom. We build on the results in Section 2 by combining the expression for $p(\boldsymbol{a})$ in (15) with an approximation to (6) or (7) that uses pre-specified values $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\boldsymbol{D}} = \text{diag}(\widehat{\boldsymbol{\lambda}})$ ($\boldsymbol{\lambda}$ and $\boldsymbol{D}$ are unavailable since they are a deterministic function of the MCMC draw $\boldsymbol{a}^{(t)}$ that has yet to be sampled). Then, by analogy with (8), we have $\boldsymbol{V} = (\boldsymbol{A}_0^{-1} + \sum_{i=1}^N \boldsymbol{U}_i\widehat{\boldsymbol{D}}^{-1}\boldsymbol{U}_i)^{-1}$ and $\boldsymbol{\mu} = \boldsymbol{V}(\boldsymbol{A}_0^{-1}\boldsymbol{a}_0 + \sum_{i=1}^N \boldsymbol{U}_i\widehat{\boldsymbol{D}}^{-1}\boldsymbol{u}_i)$, and by analogy with (10), we have $\boldsymbol{V}_k = (\boldsymbol{A}_{k0}^{-1} + \widehat{\lambda}_k^{-1}\boldsymbol{X}_k'\boldsymbol{X}_k)^{-1}$ and $\boldsymbol{\mu}_k = \boldsymbol{V}_k(\boldsymbol{A}_{k0}^{-1}\boldsymbol{a}_{k0} + \widehat{\lambda}_k^{-1}\boldsymbol{X}_k'\boldsymbol{z}_k)$. A simple choice of $\widehat{\boldsymbol{\lambda}}$ can be obtained by iterating a few times between the expression for $\boldsymbol{\mu}$ given here and the equations for $\boldsymbol{\lambda}$ in (13) and (14) (convergence is usually achieved in three to four iterations). This choice of $\widehat{\boldsymbol{\lambda}}$ has performed competitively in our examples and tends to be fast and undemanding. Moreover, the values for $\boldsymbol{\mu}$ and $\boldsymbol{V}$ discussed above can be useful starting points for optimization when one is interested in finding the mode and modal dispersion matrix of the posterior for $\boldsymbol{a}$. Random walk versions of the proposal densities can be obtained by centering the proposal at $\boldsymbol{\mu} = \boldsymbol{a}^{(t-1)}$, where $\boldsymbol{a}^{(t-1)}$ is the latest available draw in the Markov chain.

The proposal densities in (16) and (17) can be used in MH or ARMH sampling (see Appendix B for details on the latter algorithm). Once a candidate draw $\boldsymbol{a}^c$ is available from the proposal density $f(\boldsymbol{a}|\widehat{\boldsymbol{\lambda}}, \boldsymbol{u})$, all terms in (15) and the likelihood in (6) (or equivalently (7)) can be easily evaluated as a function of $\boldsymbol{a}^c$ and the implied $\boldsymbol{\Sigma}^c = (\boldsymbol{L}^c)^{-1}\boldsymbol{D}^c(\boldsymbol{L}'^c)^{-1}$. A posterior draw of $\boldsymbol{a}$ (and therefore $\boldsymbol{\Sigma}$) is obtained by proceeding with an MH step, setting $\boldsymbol{a}^{(t)} = \boldsymbol{a}^c$ with probability

$$\min\left\{1, \frac{\ell(\boldsymbol{u}|\boldsymbol{\Sigma}^c)p(\boldsymbol{a}^c)I(\boldsymbol{\lambda}^c > 0)f(\boldsymbol{a}^{(t-1)}|\widehat{\lambda}, \boldsymbol{u})}{\ell(\boldsymbol{u}|\boldsymbol{\Sigma}^{(t-1)})p(\boldsymbol{a}^{(t-1)})I(\boldsymbol{\lambda}^{(t-1)} > 0)f(\boldsymbol{a}^c|\widehat{\lambda}, \boldsymbol{u})}\right\}$$

and returning $\boldsymbol{a}^{(t)} = \boldsymbol{a}^{(t-1)}$ otherwise. Alternatively, if ARMH sampling is applied, the acceptance probabilities are spelled out in Appendix B. The ARMH algorithm has the advantage of nesting MH sampling and being more robust to irregular densities, which, at the cost of additional draws in the AR step, can mitigate some of the difficulties with MH sampling when good approximations to the target density are difficult to find, as is often the case in high-dimensional problems. Moreover, in the current context ARMH simulation is useful because only draws for which $I(\boldsymbol{\lambda}^c > 0)$ is satisfied pass through the accept-reject part of the algorithm, thus enabling better acceptance rates in the MH part. For these reasons, draws from the ARMH algorithm generally tend to exhibit better properties than those from similarly constructed MH samplers. We illustrate the application of the MH and ARMH algorithms suggested above in several examples in Sections 4 and 5.

3.3 Some off-diagonal elements are restricted to zero

We now discuss how to impose the restriction $\sigma_{kj} = 0$, $k > j$ (by symmetry, σ_{jk} is automatically zero). Multiple restrictions of this type can be imposed similarly, and the examples we consider

here and in Sections 4 and 5 provide illustration. By (12), it follows that $\sigma_{kj} = 0$ if and only if

$$a^{kj} = -\lambda_j^{-1} \sum_{h=1}^{j-1} a^{kh} a^{jh} \lambda_h, \tag{18}$$

where it is understood that the right hand side of the equation is zero when summing over an empty set (*i.e.* $a^{k1} = 0$). The goal is to derive an equivalent condition on a_{kj}. Then we can incorporate the restriction $\sigma_{kj} = 0$ into the prior of $\boldsymbol{a}$ in an analogous manner as in the previous case. To this end, we first express each element of the inverse matrix $\boldsymbol{L}^{-1}$ in terms of elements of $\boldsymbol{L}$. Denote the columns of $\boldsymbol{L}^{-1}$ as $\boldsymbol{L}^{-1} = [\boldsymbol{l}_1 : \cdots : \boldsymbol{l}_p]$. By virtue of being the inverse, we have $\boldsymbol{L}\boldsymbol{L}^{-1} = \boldsymbol{I}_p$, i.e., $\boldsymbol{L}\boldsymbol{l}_k = e_k,\ k = 1, \ldots, p$, where e_k is a column of zeros except the kth element which is equal to one. Since $\boldsymbol{L}$ is a lower triangular matrix, we can solve the above system of equations by back substitution. In fact, one can easily check that we have the explicit formula for a^{ij}:

$$a^{ij} = -a_{ij} + \sum_{j<k_1<i} a_{ik_1} a_{k_1 j} - \sum_{j<k_1<k_2<i} a_{ik_1} a_{k_1 k_2} a_{k_1 j} + \ldots + (-1)^{i+j} a_{i,i-1} a_{i-1,i-2} \ldots a_{21}. \tag{19}$$

For example, a^{52} is given by $a^{52} = -a_{52} + (a_{54}a_{42} + a_{53}a_{32}) - a_{54}a_{43}a_{32}$. By the above formula and (18), we have the desired condition: $\sigma_{ij} = 0$ if and only if

$$\begin{aligned} a_{ij} \;=\; & \sum_{j<k_1<i} a_{ik_1} a_{k_1 j} - \sum_{j<k_1<k_2<i} a_{ik_1} a_{k_1 k_2} a_{k_1 j} + \ldots \\ & + (-1)^{i+j} a_{i,i-1} a_{i-1,i-2} \cdots a_{21} + \lambda_j^{-1} \sum_{k=1}^{j-1} a^{ik} a^{jk} \lambda_k. \end{aligned}$$

Now, to impose the condition $\sigma_{ij} = 0$, a_{ij} must be set to be equal the solution of the above expression with probability 1. It is worth mentioning that the right hand side of the above expression might involve a_{ij} (see the examples below). In general, a Metropolis-Hastings step is required as the resulting full conditional densities are nonstandard, although, as in the previous case, a natural proposal density is available based on the results in Section 2. In addition, for certain special cases, only a Gibbs step is needed. For concreteness of discussion, we demonstrate the proposed algorithm in two examples.

Example 1. Suppose we have a 4×4 covariance matrix and want to impose the conditions $\sigma_{31} = \sigma_{32} = 0$. By (18), we have

$$\begin{aligned} \sigma_{31} = 0 \;\; &\Leftrightarrow \;\; a^{31} = 0, \\ \sigma_{32} = 0 \;\; &\Leftrightarrow \;\; a^{32} = -\lambda_2^{-1} a^{31} a^{21} \lambda_1 = 0, \end{aligned}$$

where the last equality holds because $a^{31} = 0$. Now by the formula for the inverse (19)

$$\begin{aligned} a^{31} &= -a_{31} + a_{32}a_{21}, \\ a^{32} &= -a_{32}. \end{aligned}$$

Combining all equations, we finally have $\sigma_{31} = \sigma_{32} = 0$ if and only if $a_{31} = a_{32} = 0$. Therefore, to produce the restrictions $\sigma_{31} = \sigma_{32} = 0$, we have $a_{31} = a_{32} = 0$ with probability one. For the purposes of illustration, suppose that the priors for the other parameters are $a_2|\lambda_2 \sim \mathcal{N}_1(0, \lambda_2)$, $\boldsymbol{a}_4|\lambda_4 \sim \mathcal{N}_3(\boldsymbol{0}, \lambda_4 \boldsymbol{I}_3)$, and $\lambda_k \overset{ind}{\sim} \mathcal{IG}\left((\nu + k - p)/2, 1/2\right)$, for $\nu > p$ and $k = 1, \ldots, p$. Then the full conditional densities for $\boldsymbol{a}_2$ and $\boldsymbol{a}_4$ are given in (10) while those of $\lambda_1, \ldots, \lambda_p$ are given in (5). Notice

that in this case only direct sampling is required, so that simulation-based estimation under these restrictions turns out to be very simple with the proposed approach; in contrast, many existing sampling approaches would be quite difficult to adapt to these restrictions.

Example 2. Suppose we have a 4×4 covariance matrix and want to impose the conditions $\sigma_{31} = \sigma_{42} = 0$. By (18), we have

$$\begin{aligned} \sigma_{31} = 0 &\Leftrightarrow a^{31} = 0, \\ \sigma_{42} = 0 &\Leftrightarrow a^{42} = -\lambda_2^{-1} a^{41} a^{21} \lambda_1. \end{aligned}$$

By the formula for the inverse (19)

$$\begin{aligned} a^{21} &= -a_{21}, \\ a^{31} &= -a_{31} + a_{32} a_{21}, \\ a^{41} &= -a_{41} + a_{42} a_{21} + a_{43} a_{31} - a_{43} a_{32} a_{21}, \\ a^{42} &= -a_{42} + a_{43} a_{32}. \end{aligned}$$

Combining these equations and solving for a_{31} and a_{42}, we have $\sigma_{31} = \sigma_{42} = 0$ if and only if

$$a_{31} = a_{32} a_{21}, \tag{20}$$

$$a_{42} = \frac{a_{43} a_{32} + \lambda_1 \lambda_2^{-1} a_{21} a_{41}}{1 + \lambda_1 \lambda_2^{-1} a_{21}^2}. \tag{21}$$

For convenience, partition $\boldsymbol{a}$ into 2 sets: $\boldsymbol{b} = (a_{21}, a_{32}, a_{41}, a_{43})$ and $\boldsymbol{c} = (a_{31}, a_{42})$. Notice that now $\boldsymbol{c}$ is a deterministic function of $\boldsymbol{b}$ and thus its role is purely notational. We incorporate the restrictions $\sigma_{31} = \sigma_{42} = 0$ into the prior as follows: we let $a_{31} = a_{32} a_{21}$ and $a_{42} = (a_{43} a_{32} + \lambda_1 \lambda_2^{-1} a_{21} a_{41})/(1 + \lambda_1 \lambda_2^{-1} a_{21}^2)$ with probability one while the priors for λ and other elements in $\boldsymbol{a}$ (*i.e.*, $\boldsymbol{b}$) can be taken as in (1) and (2), respectively. Observe that given $\boldsymbol{a}$, the full conditional densities of λ_3 and λ_4 are exactly the same as (5). For λ_1 and λ_2, a Metropolis-Hastings step is required and we consider a natural proposal density suggested by the results in Section 2

$$g(\lambda_1, \lambda_2 | \boldsymbol{a}, \boldsymbol{u}) = f_{IG}\left(\lambda_1 \,|\, \frac{N + \nu + 1 - p}{2}, \frac{s_1 + 1}{2}\right) f_{IG}\left(\lambda_2 \,|\, \frac{N + \nu + 2 - p}{2}, \frac{s_2 + 1}{2}\right), \tag{22}$$

where f_{IG} is the inverse gamma density, s_k the (k,k)-element of $\sum_{i=1}^N \boldsymbol{w}_i \boldsymbol{w}_i'$ and $\boldsymbol{w}_i \equiv L \boldsymbol{u}_i$. Given a candidate draw $(\lambda_1^c, \lambda_2^c)$, define $D^c = \mathrm{diag}(\lambda_1^c, \lambda_2^c, \lambda_3, \lambda_4)$ and let L^c denote the matrix L with a_{42} replaced by $a_{42}^c = (a_{43} a_{32} + \lambda_1^c (\lambda_2^c)^{-1} a_{21} a_{41})/(1 + \lambda_1^c (\lambda_2^c)^{-1} a_{21}^2)$. Then the candidate draw is accepted with probability

$$\min\left\{1, \frac{\ell(\boldsymbol{u}|\boldsymbol{\Sigma}^c) p(\lambda_1^c, \lambda_2^c) g(\lambda_1, \lambda_2 | \boldsymbol{a}, \boldsymbol{u})}{\ell(\boldsymbol{u}|\boldsymbol{\Sigma}) p(\lambda_1, \lambda_2) g(\lambda_1^c, \lambda_2^c | \boldsymbol{a}, \boldsymbol{u})}\right\}.$$

To obtain a draw from $\boldsymbol{a}|\lambda, \boldsymbol{u}$, we utilize the proposal density

$$f(\boldsymbol{b}|\boldsymbol{\lambda}, \boldsymbol{u}) = f_T(a_{21} | D_2 d_2, D_2, \kappa) f_T(a_{32} | D_3 d_3, D_3, \kappa) f_T(a_{41}, a_{43} | \boldsymbol{D}_4 \boldsymbol{d}_4, \boldsymbol{D}_4, \kappa), \tag{23}$$

where, given our assumed priors,

$$\begin{aligned} D_2 &= \lambda_2 (1 + \boldsymbol{U}_1' \boldsymbol{U}_1)^{-1}, \;\; d_2 = -\boldsymbol{U}_1' \boldsymbol{U}_2 / \lambda_2, \\ D_3 &= \lambda_3 (1 + \boldsymbol{U}_2' \boldsymbol{U}_2)^{-1}, \;\; d_3 = -\boldsymbol{U}_2' \boldsymbol{U}_3 / \lambda_3, \\ \boldsymbol{D}_4 &= \lambda_4 (\boldsymbol{I}_2 + [\boldsymbol{U}_1 : \boldsymbol{U}_3]' [\boldsymbol{U}_1 : \boldsymbol{U}_3])^{-1}, \;\; d_4 = -[\boldsymbol{U}_1 : \boldsymbol{U}_3]' \boldsymbol{U}_4 / \lambda_4, \\ \boldsymbol{U}_k &= (u_{1k}, \cdots, u_{Nk})', \;\; k = 1, \ldots, 4. \end{aligned}$$

Given a candidate draw $\boldsymbol{b}^c$, a_{43}^c and a_{32}^c are determined by equations (20) and (21). Then accept the draw $\boldsymbol{a}^c = \{\boldsymbol{b}^c, \boldsymbol{c}^c\}$ with probability

$$\min\left\{1, \frac{\ell(\boldsymbol{u}|\boldsymbol{\Sigma}^c)p(\boldsymbol{b}^c)f(\boldsymbol{b}|\boldsymbol{\lambda}, \boldsymbol{u})}{\ell(\boldsymbol{u}|\boldsymbol{\Sigma})p(\boldsymbol{b})f(\boldsymbol{b}^c|\boldsymbol{\lambda}, \boldsymbol{u})}\right\}.$$

Upon completion of these sampling steps, the resulting posterior draws for $\boldsymbol{\Sigma} = (\boldsymbol{L}^c)^{-1}\boldsymbol{D}^c(\boldsymbol{L}'^c)^{-1}$ satisfy the restriction $\sigma_{31} = \sigma_{42} = 0$. As suggested earlier, the MH proposal densities discussed above can also be used as pseudo-dominating proposal densities in ARMH simulation.

4 Examples and Extensions

This section begins with three examples based on the discussion in Section 3 that illustrate the performance of the proposed approach for dealing with covariance restrictions. We then suggest techniques for addressing a number of complications that may arise in practice and present evidence from a simulation study. The first three examples use a sample size of $N = 700$ observations and MCMC runs of 11000 iterations of which the first 1000 are discarded as burn-in. To gauge the performance of the MCMC algorithms, we also report the *inefficiency factors* for the sampled parameters, which give a useful metric of the performance of the Markov chain. For a given scalar parameter θ, the inefficiency factor approximates the ratio of the numerical variance of the posterior mean from the MCMC output relative to that from hypothetical iid draws. To obtain an estimate of the latter quantity, we use the method of batch means, where the m draws in the MCMC sample are batched into v equal non-overlapping groups such that the respective batch means $\bar{\theta}_j$, $j = 1, \ldots, v$, are approximately serially uncorrelated. Then the inefficiency factor is given by the ratio

$$\frac{var(\theta|y)/m}{var(\bar{\theta}|y)/v},$$

where the numerator is computed with draws from the main MCMC run and the denominator gives the variance of the overall mean as implied by the batch means. It can be easily seen that the ratio will approach 1 as the posterior draws from the Markov chain become less serially correlated.

4.1 Illustrations

Illustration 1: A 4×4 covariance matrix with $\sigma_{11} = \sigma_{22} = \sigma_{33} = \sigma_{44} = 1$. In the first example we generate data from $u_i \overset{iid}{\sim} \mathcal{N}_4(\boldsymbol{0}, \boldsymbol{\Sigma})$, $i = 1, \ldots, 700$, where

$$\boldsymbol{\Sigma} \equiv \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix} = \begin{pmatrix} 1 & 0.2 & 0.3 & -0.4 \\ 0.2 & 1 & 0.6 & 0.2 \\ 0.3 & 0.6 & 1 & -0.2 \\ -0.4 & 0.2 & -0.2 & 1 \end{pmatrix}.$$

We impose the condition $\sigma_{11} = \sigma_{22} = \sigma_{33} = \sigma_{44} = 1$ by the MH method discussed in Section 3.2. We assume the following prior for $\boldsymbol{a}_k$: $\boldsymbol{a}_k \overset{ind}{\sim} \mathcal{N}_{k-1}(\boldsymbol{0}, \boldsymbol{I}_{i-1})$, $k = 2, \ldots, p$, while the priors for λ_k, $k = 1, \ldots, p$ are determined by (13) and (14) together with the restriction that $\lambda_k > 0$. The one-block MH algorithm for $\boldsymbol{a}$ described in Section 3.2 produces MCMC draws for which all autocorrelations drop below 0.05 after a few lags (examples are given in Figure 1). We report the true values, posterior means and standard deviations, together with the corresponding inefficiency factors in Table 1.

Figure 1: Examples of parameter autocorrelations in the first illustration.

Table 1: Simulation results for a 4×4 correlation matrix where $\sigma_{11} = \sigma_{22} = \sigma_{33} = \sigma_{44} = 1$.

Parameter	True Value	Posterior Mean	Posterior SD	Inefficiency
σ_{21}	0.2	0.195	0.035	2.6
σ_{31}	0.3	0.298	0.032	2.6
σ_{41}	-0.4	-0.382	0.030	2.4
σ_{32}	0.6	0.589	0.021	2.1
σ_{42}	0.2	0.216	0.032	2.5
σ_{43}	-0.2	-0.205	0.032	2.6

Illustration 2: A 4×4 covariance matrix with $\sigma_{31} = \sigma_{32} = 0$. In this example we generate data from $u_i \overset{iid}{\sim} \mathcal{N}_4(\mathbf{0}, \mathbf{\Sigma})$, $i = 1, \ldots, 700$, where

$$\mathbf{\Sigma} \equiv \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix} = \begin{pmatrix} 1.2 & 0.9 & 0 & 0.5 \\ 0.9 & 1 & 0 & 0.3 \\ 0 & 0 & 0.9 & 0.2 \\ 0.5 & 0.3 & 0.2 & 1.1 \end{pmatrix}.$$

It is shown in Example 1 in Section 3.3 that the conditions $\sigma_{31} = \sigma_{32} = 0$ are equivalent to $a_{31} = a_{32} = 0$. The details of the prior densities with the restrictions imposed, together with the posterior densities are also given in that section. The simplicity of the resulting sampler in this example can be rather surprising – despite the restrictions, only trivial Gibbs sampling is required without any MH steps. In the simulation below all posterior autocorrelations drop below 0.05 after the first lag. We report the true values, posterior means and standard deviations in Table 2.

Table 2: Simulated results for a 4×4 covariance matrix with $\sigma_{31} = \sigma_{32} = 0$.

Parameter	True Value	Posterior Mean	Posterior SD	Inefficiency
σ_{11}	1.2	1.200	0.064	1.00
σ_{21}	0.9	0.897	0.054	1.00
σ_{41}	0.5	0.491	0.047	1.00
σ_{22}	1.0	1.001	0.053	1.00
σ_{42}	0.3	0.302	0.041	1.00
σ_{33}	0.9	0.891	0.048	1.00
σ_{43}	0.2	0.191	0.035	1.00
σ_{44}	1.1	1.121	0.060	1.00

Illustration 3: A 4×4 covariance matrix with $\sigma_{11} = 1$ and $\sigma_{31} = \sigma_{42} = 0$. In this example we generate data from $u_i \overset{iid}{\sim} \mathcal{N}_4(\mathbf{0}, \mathbf{\Sigma})$, $i = 1, \ldots, 700$, where

$$\mathbf{\Sigma} \equiv \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{pmatrix} = \begin{pmatrix} 1 & 0.5 & 0 & 0.4 \\ 0.5 & 0.9 & -0.2 & 0 \\ 0 & -0.2 & 1.1 & -0.3 \\ 0.4 & 0 & -0.3 & 0.8 \end{pmatrix}.$$

We impose the conditions $\sigma_{11} = 1$ and $\sigma_{31} = \sigma_{42} = 0$ by letting $\lambda_1 = 1$, $a_{31} = a_{32}a_{21}$ and $a_{42} = (a_{43}a_{32} + \lambda_2^{-1}a_{21}a_{41})/(1 + \lambda_2^{-1}a_{21}^2)$ with probability 1. A posterior simulator was discussed in detail in example 2 of Section 3.3 – the acceptance rates in both MH steps of that sampler were above 90%, and all posterior autocorrelations dropped below 0.05 after the first lag. We report the true values, posterior means and standard deviations in Table 3.

Table 3: Simulated results for a 4×4 covariance matrix with $\sigma_{11} = 1$ and $\sigma_{31} = \sigma_{42} = 0$.

Parameter	True Value	Posterior Mean	Posterior SD	Inefficiency
σ_{21}	0.5	0.500	0.029	1.00
σ_{41}	0.4	0.395	0.026	1.00
σ_{22}	0.9	0.890	0.044	1.00
σ_{32}	-0.2	-0.211	0.030	1.00
σ_{33}	1.1	1.096	0.057	1.00
σ_{43}	-0.3	-0.314	0.034	1.00
σ_{44}	0.8	0.823	0.040	1.00

4.2 Practical caveats, extensions, and suggestions

The above three illustrations show that a variety of covariance restrictions can be handled by adopting the proposed approach. However, we caution that practical applications often involve one or more unfavorable factors that could impede the performance of the Markov chain. For this reason, we now draw attention to such potential complications, point out their impact on the simulation of the covariance matrix, and discuss practical solutions for addressing them.

One complication, which is of particular practical relevance, occurs because restricted covariance matrices arise most commonly in discrete data models, where latent data augmentation is an essential part of the sampling algorithm. However, in multivariate models, the latent data vector $\boldsymbol{y}_i^*$ that underlies the observed data vector $\boldsymbol{y}_i$ (for example, in binary data models $y_{ij} = I(y_{ij}^* > 0)$ for $i = 1, \ldots, N$, $j = 1, \ldots, p$) is sampled one-element-at-a-time by drawing from $[y_{ij}^* | \boldsymbol{y}, \boldsymbol{y}_{i,-j}^*]$ for $j = 1, ..., p$ (see, for example, Geweke, 1991; Robert, 1995). When the dimension of $\boldsymbol{y}_i^*$ is large, or when the correlations between its elements are high, the latent data draws mix slowly and as a consequence also slow down the mixing of the entire chain, including the sampler for the covariance matrix. In this case, even a well performing sampler for $\boldsymbol{\Sigma}$ can suffer because at every MCMC iteration it may be conditioned on poorly mixing latent data. Such circumstances may require longer MCMC chains to reduce the simulation standard errors of the posterior estimates.

A second obvious complicating factor is dimensionality. It can present serious difficulties since the number of parameters in $\boldsymbol{\Sigma}$ grows quadratically with p. In cases that require MH sampling, the problem of dimensionality can manifest itself through difficulties in tuning the proposal density – approximations that work reasonably well for small p may deteriorate due to the compounding of small discrepancies as p is increased. To deal with this problem, the parameters of the covariance matrix can be sampled in a sequence of smaller, more manageable blocks as in Chib and Greenberg (1998). Appropriate blocking and sequential sampling can be applied rather naturally in the current context. For the setting in Section 3.2, blocking can take place according the rows of $\boldsymbol{L}$ as in (17), where, in case $\boldsymbol{A}_0$ is not block diagonal, $\boldsymbol{A}_{k0}$ and $\boldsymbol{a}_{k0}$ should be taken as the conditional, not marginal, moments of the prior distribution given the other rows of $\boldsymbol{L}$; in Section 3.3, it is more sensible to block the parameters along the lines of $\boldsymbol{a}$ and $\boldsymbol{\lambda}$. Such blocking offers a straightforward way of extending the techniques of Sections 3.2 and 3.3 to larger matrices when the limitations of

single-block MH sampling may become more pronounced as p is increased. These techniques are applied in the examples in Section 5.

While appropriate blocking of the elements of $\boldsymbol{\Sigma}$ may be a useful way of dealing with high-dimensional matrices, there are benefits to joint sampling of the entire covariance matrix in a single block, especially when the parameters are correlated. In such cases, we argue in favor of using tuned ARMH sampling, built upon the parameterization of Sections 2 and 3, as a way of mitigating some of the pitfalls of high-dimensional MH samplers. Improvements in sampling can be expected because, at the cost of additional tuning and simulation in the AR step, the MH draws from the ARMH algorithm tend to exhibit better properties than similarly constructed MH samplers. We study the performance of such an algorithm next.

To consider the impact of issues such as dimensionality, latent data augmentation, and magnitude of correlations in $\boldsymbol{\Sigma}$, we focus on the case where $\boldsymbol{\Sigma}$ is a $p \times p$ correlation matrix and illustrate the aforementioned issues with data from the model

$$\boldsymbol{y}_i^* = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad i = 1, \ldots, N, \tag{24}$$

where $\boldsymbol{X}_i$ is a $p \times k$ covariate matrix with entries sampled iid from a standard normal distribution, and $\boldsymbol{\beta}$ is a $k \times 1$ vector which we set to equal $0.3 * \boldsymbol{1}$. We let $N = 1500$ and $k = 4$, and study the effect of dimensionality by varying p over three possible settings, namely $p = 4$, $p = 6$, and $p = 8$. In order to elicit the effects of latent data augmentation, we fit the model (i) as if $\boldsymbol{y}_i^*$ were continuous observed data *i.e.* $\boldsymbol{y}_i = \boldsymbol{y}_i^*$ and, (ii) as if $\boldsymbol{y}_i^*$ were unobserved latent data underlying the observed binary outcomes $\boldsymbol{y}_i$, where $y_{ij} = 1\{y_{ij}^* > 0\}$, as in MVP models. We also consider a "high correlation" case where

$$\boldsymbol{\Sigma}[j,k] = \max\{0, (1 - 0.25|j-k|)\},$$

and a "low correlation" case where

$$\boldsymbol{\Sigma}[j,k] = \begin{cases} (1/2)^{|j-k|} & \text{if} \quad |j-k| \leq 2 \\ 0 & \text{otherwise} \end{cases}.$$

In each case we estimate the $p(p-1)/2$ free correlations in $\boldsymbol{\Sigma}$ through a single-block ARMH algorithm. Parameterization of the correlation matrix is in terms of $\boldsymbol{a}$ as in Section 3.2. The pseudo-dominating density in the algorithm is based on (16), with $\boldsymbol{\mu}$ and $\boldsymbol{V}$ being determined by additional quasi-Newton maximization starting with the approximations based on $\widehat{\boldsymbol{\lambda}}$ that were discussed in Section 3.2. The tuning parameters τ and κ in (16) are set at $\tau = 1.5$ and $\kappa = 10$ in order to provide a sufficiently heavy-tailed proposal since there is no guarantee that local features of the target, such as its mode and modal dispersion, will be useful for addressing the potential for multimodality, skewness, kurtosis, or other complications in the tails. In addition, we choose the constant c in the AR step of the ARMH algorithm so that the degree of domination $\frac{cf_T(\boldsymbol{a}|\boldsymbol{\mu},\tau\boldsymbol{V},\kappa)}{\ell(\boldsymbol{u}|\boldsymbol{\beta},\boldsymbol{\Sigma})f_{\mathcal{N}}(\boldsymbol{a}|\boldsymbol{a}_0,\boldsymbol{A}_0)} = 1.5$ evaluated at $\boldsymbol{a} = \boldsymbol{\mu}$. The sampling of $\boldsymbol{\beta}$ (and $\{\boldsymbol{y}_i^*\}$ in the discrete data case) is done as in Chib and Greenberg (1998).

Boxplots of the inefficiency factors for the elements of $\boldsymbol{\Sigma}$ under a number of different simulation settings are presented in Figure 2. The four panels in that figure depict continuous and discrete data cases under the "high correlation" and "low correlation" scenarios discussed above. From the figure we see that as the dimension of $\boldsymbol{\Sigma}$ is increased within each panel (implying a larger number of parameters to be estimated – 6 when $p = 4$, 15 when $p = 6$, and 28 when $p = 8$), the inefficiency factors increase in both the continuous and discrete data settings. As one might expect, the inefficiency factors in the discrete data panels of the figure are higher than those in the continuous data panels. This is due, on the one hand, to the fact that binary data are

less informative about covariate effects because the threshold-crossing nature of the discretization transformation $y_{ij} = 1\{y_{ij}^* > 0\}$ only contains information on signs, and not on magnitudes; on the other hand, the larger inefficiency factors in the binary case can be attributed in part to the fact that the elements of $\boldsymbol{y}_i^*$ are generally sampled one-at-a-time, conditionally on all other elements in $\boldsymbol{y}_i^*$ (Geweke, 1991; Robert, 1995). All else being equal, higher correlations in $\boldsymbol{\Sigma}$ and larger p lead to larger inefficiency factors for the discrete data case, whereas the two continuous data panels of Figure 2 reveal dependence of the inefficiency factors on p, but do not exhibit visible inefficiency factor deterioration when the correlations in $\boldsymbol{\Sigma}$ are increased since in those panels $\boldsymbol{y}_i = \boldsymbol{y}_i^*$, $i = 1, \ldots, N$, need not be resampled at every MCMC iteration.

Figure 2: Inefficiency factors in the one-block sampling of correlation matrices of varying sizes in continuous and binary data models with high or low correlations. The number of free parameters in $\boldsymbol{\Sigma}$ is 6 when $p = 4$, 15 when $p = 6$, and 28 when $p = 8$.

To demonstrate the trade-off between numerical and statistical efficiency that is intrinsic in the ARMH algorithm, we also report the acceptance rates in the AR and MH steps. When $p = 4$, the AR acceptance rates were between 0.33 and 0.34, with corresponding MH acceptance rates of 1, indicating that the specific choices of τ, κ, and c produced a pure AR sampler in which the proposal dominates the posterior. When p was increased to 6, the AR rates were in the range 0.52-0.54, whereas the MH rates dropped to 0.74-0.76. When p was further increased to 8, the AR rates were in the range 0.67-0.7, and the MH acceptance rates were between 0.38 and 0.4. This behavior of the acceptance rates is an indication that in high dimensions the proposal density produces smaller regions of domination, that the posterior is not very well-behaved, and that it is only roughly approximated by its mode and modal curvature. However, our implementation shows that the applicability of such approximations can be largely extended because of the adaptability of the ARMH algorithm, which has allowed us to sample the 28 parameters defining the covariance matrix in a single block, despite the irregularities of the posterior surface.

In summary, we mention that in a number of cases the proposed approach has required no MH steps. However, when such steps are required, the framework discussed in this paper leads to a reasonable MH proposal density. When further complications are present (e.g. high dimensionality, high correlations, discrete data), the approach allows for splitting of the parameters in $\boldsymbol{\Sigma}$ into smaller, more manageable blocks that can be sampled sequentially. Grouping of the parameters and extensions to high-dimensional structured matrices can often be facilitated by contextual considerations such as the conditional independence restrictions occurring in graphical models (e.g. Carvalho et al., 2007). Alternatively, single-block simulation can be pursued through a flexible ARMH step when needed. In the next section, we apply these techniques to analyze two real data applications.

5 Applications

5.1 Intertemporal labor force participation of married women

Our first application uses data from Chib and Jeliazkov (2006) to study the labor force participation decisions (1=working, 0=not working) of 1545 married women in the age range 17-66 over a 7-year period (1979-1985). Since the binary decision of whether or not to participate in the labor force is expected to be correlated over time, we consider an MVP model with a 7×7 correlation matrix containing a total of 21 unknown correlation parameters. Under the priors $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_8)$, and

$\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{21})I(\boldsymbol{a} \in \mathcal{C})$, where $\boldsymbol{a} \in \mathcal{C}$ if and only if the implied elements of $\boldsymbol{\lambda}$ are all positive, the MCMC sampler is constructed by sequentially drawing from the distributions for the regression parameters $[\boldsymbol{\beta}|\boldsymbol{y}^*, \boldsymbol{a}]$, the latent data $[\boldsymbol{y}^*|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{a}]$, and the correlation matrix $[\boldsymbol{a}|\boldsymbol{y}^*, \boldsymbol{\beta}]$. The first two steps are identical to those in Chib and Greenberg (1998), whereas in the third step we sample $[\boldsymbol{a}|\boldsymbol{y}^*, \boldsymbol{\beta}]$ and construct $\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}'\boldsymbol{D}^{-1}\boldsymbol{L}$ as in Section 3.2. Since the covariance matrix is large, we estimate it in two ways: (i) we split $\boldsymbol{\Sigma}$ into smaller blocks, following Chib and Greenberg (1998), and sequentially draw from $\boldsymbol{a}_k|\boldsymbol{y}^*, \boldsymbol{\beta}, \boldsymbol{a}_{-k}$, $k = 2, \ldots, 7$, where $\boldsymbol{a}_{-k}$ denotes all parameters in $\boldsymbol{a}$ except $\boldsymbol{a}_k$, and (ii) we sample $\boldsymbol{\Sigma}$ through a single-block ARMH algorithm as discussed in Section 4. The samplers are run for 31000 iterations with the first 1000 discarded as burn-in. Summaries of the posterior distribution for the model parameters are reported in Tables 4 and 5.

Table 4: Parameter estimates in the women's labor force participation application.

		Posterior	
Parameter	Covariate	Mean	SD
β_1	Intercept (column of ones)	-0.636	0.312
β_2	Woman's age in years	0.042	0.015
β_3	Woman's age squared, divided by 100	-0.001	0.000
β_4	Race (1 if black, 0 otherwise)	0.270	0.069
β_5	Attained education (in years) at time of survey	0.103	0.012
β_6	Number of children aged 0-2 in that year	-0.299	0.028
β_7	Number of children aged 3-5 in that year	-0.183	0.025
β_8	Annual labor income of head of household	-0.154	0.025

Table 4 contains the covariate effect estimates for the MVP model. It shows that conditional on the covariates, black women, better educated women, and women whose husbands have low earnings, are more likely to work. After controlling for the effects of the remaining covariates, we see that the presence of children has a larger effect on the probability of working when the children are younger, as expected. In addition, labor force participation is a concave function of age, suggesting varying tastes and trade-offs over a woman's life-cycle, but also capturing the fact that age is revealing of social values, education type, experience, and human capital.

Table 5: Correlation estimates for the women's labor force participation data.

	Posterior			Posterior			Posterior	
Parameter	Mean	SD	Parameter	Mean	SD	Parameter	Mean	SD
σ_{21}	0.844	0.018	σ_{52}	0.696	0.026	σ_{65}	0.856	0.018
σ_{31}	0.745	0.026	σ_{53}	0.812	0.021	σ_{71}	0.570	0.034
σ_{32}	0.854	0.017	σ_{54}	0.865	0.017	σ_{72}	0.586	0.033
σ_{41}	0.693	0.026	σ_{61}	0.637	0.030	σ_{73}	0.670	0.030
σ_{42}	0.758	0.024	σ_{62}	0.659	0.029	σ_{74}	0.729	0.027
σ_{43}	0.883	0.014	σ_{63}	0.697	0.028	σ_{75}	0.825	0.021
σ_{51}	0.645	0.030	σ_{64}	0.786	0.023	σ_{76}	0.897	0.014

As can be seen from Table 5, the correlation parameters are all quite large and relatively precisely estimated. The correlations decline as the distance between time periods grows larger, which is consistent with the results in Chib and Jeliazkov (2006), who suggest that the correlations

can eventually be explained by allowing for dependence on lagged responses and accounting for heterogeneity in the intercept and the effects of children. The inefficiency factors for the MH and ARMH samplers of $\boldsymbol{\Sigma}$ are shown in Figure 3, and, given the caveats on latent data augmentation discussed in Section 4, indicate a good overall performance of these MCMC samplers. The figure demonstrates that both samplers present viable options for estimation, but that the additional tuning and simulation costs of ARMH sampling pay off through a reduction in inefficiency factors relative to multi-block MH sampling.

Figure 3: Inefficiency factors for $\boldsymbol{\Sigma}$ in the MVP application. The first set of inefficiency factors is obtained from a single-block ARMH sampler, whereas the second set comes from multi-block MH.

5.2 Scheduling of work trips

Our second application deals with the scheduling of work trips by 522 San Francisco Bay Area commuters, which was studied by Small (1982) and Brownstne and Small (1989) using conditional and nested logit models. In this example, we analyze these data using an MNP model and focus on a parsimoniously parameterized covariance structure. The data set consists of commuters' self-reported regular time of arrival relative to the official work start time. For our purposes, the arrival times are grouped into six 10-minute intervals because data deficiencies and parameter proliferation in $\boldsymbol{\Sigma}$ preclude analysis with finer 5-minute arrival intervals. The observed 6×1 multinomial vector $\boldsymbol{y}_i$ is modeled in terms of a 5×1 latent representation

$$\boldsymbol{y}_i^* = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, N,$$

where $\boldsymbol{y}_i[1] = I\left(\max(\boldsymbol{y}_i^*) \leq 0\right)$ and $\boldsymbol{y}_i[j+1] = I\left(\boldsymbol{y}_i^*[j] = \max(\boldsymbol{y}_i^*)\right) I\left(\max(\boldsymbol{y}_i^*) > 0\right)$ and $\boldsymbol{X}_i$ contains the covariates for categories $j = 2, \ldots, 5$ differenced with respect to the baseline category ($j = 1$). In economics, this latent representation is usually given a utility interpretation in which $\boldsymbol{y}_i^*$ represent unobserved utility differences and economic agents choose the alternative that gives the largest utility. To identify the scale of the model, $\boldsymbol{\Sigma}$ incorporates the identification restriction $\sigma_{11} = 1$.

One parsimonious model that is of interest in this application involves a tridiagonal covariance matrix such that $\sigma_{ij} = 0$ for $i \geqslant j + 2$ or $j \geqslant i + 2$. This choice is guided by the potential for respondents to round off reported arrivals to within 10 or 15 minutes, thus creating correlation between adjacent categories (see Small, 1982, who used reporting error dummies to address this possibility). On a practical level, this covariance structure is useful because it captures correlations between "close substitutes" while keeping the number of unknown parameters manageable ($\boldsymbol{\Sigma}$ involves 8, instead of 14, unknown parameters).

To impose the condition $\sigma_{11} = 1$, we let $\lambda_1 = 1$ as in Section 3.1. We then partition $\boldsymbol{a} = (a_{21}, a_{31}, a_{32}, \ldots, a_{10,9})$ into two sets: $\boldsymbol{b} \equiv \{a_{i+1,i} : i = 1, \ldots, 5\}$ and $\boldsymbol{c} \equiv \{a_{ij} : i \geqslant j + 2\}$ and apply the method outlined in Section 3.3 to impose the tridiagonal structure. Specifically, by condition (18), it follows that $\sigma_{ij} = 0 \Leftrightarrow a^{ij} = 0$ for $i \geqslant j+2$, and upon using (19), a_{ij} must satisfy

$$a_{ij} = a_{i,i-1}a_{i-1,j}, \; i = 3, \ldots, 5, \; 1 \leqslant j \leqslant i - 2. \tag{25}$$

Therefore, with a tridiagonal covariance structure, each element in $\boldsymbol{c}$ equals a known function of elements in $\boldsymbol{b}$ with probability 1, so that the role of $\boldsymbol{c}$ is purely notational. For other parameters, we consider proper informative priors centered around an identity matrix $\lambda_k \overset{ind}{\sim} \mathcal{IG}\left((\nu + k - p)/2, \nu/2\right)$, $k = 2, \ldots, 5$, $\boldsymbol{b} \sim \mathcal{N}(\boldsymbol{0}, \nu^{-1}\boldsymbol{I})$, where $\nu = 50$. Prior informativeness is very important given the

small data set ($N = 522$), the potential for poor likelihood identification (*e.g.* Geweke et al., 1997), and the fact that the data set is not well balanced (the first and last categories in this example involve only 2.5% and 1.5% of the outcomes, respectively).

The MCMC sampler is constructed by sequentially drawing from 4 full-conditional densities: $[\boldsymbol{y}^*|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{\lambda}]$, $[\boldsymbol{\beta}|\boldsymbol{a},\boldsymbol{\lambda},\boldsymbol{y}^*]$, $[\boldsymbol{\lambda}|\boldsymbol{\beta},\boldsymbol{a},\boldsymbol{y}^*]$ and $[\boldsymbol{b}|\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{y}^*]$. The first two steps are standard (McCulloch et al., 2000). The third step is obtained by (5) with λ_1 set to 1. Lastly, even though the conditional density $[\boldsymbol{b}|\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{y}^*]$ is nonstandard, a natural proposal density is given by

$$f(\boldsymbol{b}|\boldsymbol{y}^*,\boldsymbol{\beta},\widehat{\boldsymbol{a}}) = \prod_{i=2}^{5} \phi_{i-1}(\boldsymbol{D}_i\boldsymbol{d}_i,\boldsymbol{D}_i),$$

where

$$\begin{aligned}
\boldsymbol{D}_i &= \lambda_i(1+\boldsymbol{U}_{i-1}'\boldsymbol{U}_{i-1})^{-1}, \;\; \boldsymbol{d}_i = -\boldsymbol{U}_{i-1}'(\boldsymbol{U}_i + \sum_{j=1}^{i-2}\widehat{a}_{ij}\boldsymbol{U}_j)/\lambda_i, \\
\boldsymbol{U}_k &= (u_{1k},\ldots,u_{Nk})', \;\; u_{ik} = y_{ik}^* - \boldsymbol{x}_{ik}'\boldsymbol{\beta},
\end{aligned}$$

and $\widehat{a}_{ij}, i = 2,\ldots,5,\ j = 1,\ldots,i-2$, can simply be set to the current value of a_{ij} in the Markov chain. Given a candidate draw $\boldsymbol{b}^c$, elements in $\boldsymbol{c}^c$ are determined by (25). The draw $\boldsymbol{a}^c = (\boldsymbol{b}^c,\boldsymbol{c}^c)$ is accepted with probability

$$\min\left\{1, \frac{\ell(\boldsymbol{y}^*|\boldsymbol{\Sigma}^c,\boldsymbol{\beta})p(\boldsymbol{b}^c)f(\boldsymbol{b}|\boldsymbol{y}^*,\boldsymbol{\beta},\widehat{\boldsymbol{a}})}{\ell(\boldsymbol{y}^*|\boldsymbol{\Sigma},\boldsymbol{\beta})p(\boldsymbol{b})f(\boldsymbol{b}^c|\boldsymbol{y}^*,\boldsymbol{\beta},\widehat{\boldsymbol{a}})}\right\},$$

where $\ell(\cdot)$ is the complete data likelihood function and $\boldsymbol{\Sigma}^c = (\boldsymbol{L}^c)^{-1}\boldsymbol{D}^c(\boldsymbol{L}^{c'})^{-1}$.

Table 6: Parameter estimates in the MNP application. Analysis is based on the variables: schedule delay $SD = \{-40,-30,-20,-10,0,10\}$ (arrival minus work start time rounded to 10 minutes); travel time TIM (in minutes); $SDE = \max\{-SD,0\}$; $SDL = \max\{SD,0\}$; $D1L = 1\{SD \geq 0\}$; reported arrival time flexibility $FLEX$; $D2L = 1\{SD \geq FLEX\}$; $SDLX = \max\{SD-FLEX,0\}$; dummies for one-person household SGL, carpool CP, and white collar worker WC.

		Posterior	
Parameter	Covariate	Mean	SD
β_1	TIM	0.005	0.006
β_2	$TIM \cdot SGL$	-0.003	0.009
β_3	$TIM \cdot CP$	-0.010	0.007
β_4	SDE	-0.007	0.005
β_5	$SDE \cdot SGL$	-0.005	0.008
β_6	$SDE \cdot CP$	0.009	0.006
β_7	SDL	-0.958	0.552
β_8	$SDL \cdot WC$	0.797	0.552
β_9	$SDLX$	-0.093	0.054
β_{10}	$D1L \cdot WC$	0.968	0.150
β_{11}	$D2L$	-0.551	0.155

Summaries of the posterior distribution for the model parameters are reported in Tables 6 and 7. The strongest effects in Table 6 indicate that white collar workers are more likely to report arriving

at work late, even after accounting for any flexibility in the starting time. Moreover, Table 7 shows that two correlation parameters are relatively large, suggesting a reasonably strong degree of substitutability between two of the arrival categories and presenting an interesting item for future research.

Table 7: Parameter estimates for the work trip scheduling data.

	posterior			posterior		
parameter	mean	SD	parameter	mean	SD	$\Pr(\sigma_{i+1,i} > 0 \mid Data)$
σ_{22}	1.211	0.199	σ_{21}	-0.025	0.158	0.431
σ_{33}	1.178	0.215	σ_{32}	-0.179	0.157	0.125
σ_{44}	1.479	0.343	σ_{43}	-0.241	0.183	0.084
σ_{55}	1.124	0.242	σ_{54}	0.019	0.206	0.529

The inefficiency factors for the MH sampler of the restricted $\mathbf{\Sigma}$ are shown in Figure 4 together with inefficiency factors from a Gibbs sampler for the unrestricted (except for $\sigma_{11} = 1$) MNP model as in Section 3.1 and McCulloch et al. (2000) under comparable priors. The figure illustrates similar overall performance of the MCMC algorithms, but also shows that the more profligately parameterized MNP model exhibits slightly slower mixing in this setting. One possible reason is that, given the small sample size, identification may deteriorate somewhat as the number of unknown parameters in $\mathbf{\Sigma}$ is increased; another possibility is that with a tridiagonal $\mathbf{\Sigma}$, the full-conditional distribution $[y^*_{ij} | \boldsymbol{y}, \boldsymbol{y}^*_{i,-j}]$ is determined only by latent data that are adjacent to y^*_{ij}, whereas with non-zero covariances $[y^*_{ij} | \boldsymbol{y}, \boldsymbol{y}^*_{i,-j}]$ is determined by the entire vector $\boldsymbol{y}^*_{i,-j}$, which increases the serial dependence in the latent data draws and subsequently slows down the mixing of the entire Markov chain.

Figure 4: Inefficiency factors for the restricted and unrestricted versions of $\mathbf{\Sigma}$ in the MNP application.

6 Concluding Remarks

This article has studied a parameterization of covariance matrices that allows for flexible modeling and straightforward MCMC-based estimation. The proposed approach is related to standard prior-posterior modeling and MCMC sampling methods in the unrestricted case, where simple conjugate priors on the elements of the alternative parameterization can lead to the usual conjugate Wishart prior on the precision matrix. This link is then exploited to facilitate simulation-based inference when covariance restrictions are imposed. Several illustrations with simulated data and two applications from economics demonstrate the handling of various diagonal and off-diagonal restrictions that are frequently encountered in practice, and show that the proposed methods are practical and can help address important problems in modeling and estimation.

Supplemental Materials

Supplemental materials for this article are available online. All of these materials are contained in a zip archive that can be obtained in a single download.

CJ-supplement: The package contains Appendices A and B, as well as the data sets and computer code used in the examples. (Zipped file)

References

A. Atay-Kayis and H. Massam. A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, 92(2):317–335, 2005.

J. Barnard, R. McCulloch, and X. L. Meng. Modelling covariance matrices in terms of standard deviations and correlations with applications to shrinkage. *Statistica Sinica*, 10:1281–1311, 2000.

D. Brownstne and K. Small. Efficient estimation of nested logit models. *Journal of Business & Economic Statistics*, 7:67–74, 1989.

C. Carvalho, H. Massam, and M. West. Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika*, 94(3):647–659, 2007.

S. Chib. Analysis of treatment response data without the joint distribution of potential outcomes. *Journal of Econometrics*, 140:401–412, 2007.

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49:327–335, 1995.

S. Chib and E. Greenberg. Analysis of multivariate probit models. *Biometrika*, 85:347–361, 1998.

S. Chib and I. Jeliazkov. Accept-reject Metropolis-Hastings sampling and marginal likelihood estimation. *Statistica Neerlandica*, 59:30–44, 2005.

S. Chib and I. Jeliazkov. Inference in semiparametric dynamic models for binary longitudinal data. *Journal of the American Statistical Association*, 101:685–700, 2006.

S. Chib, E. Greenberg, and I. Jeliazkov. Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics*, forthcoming, 2009.

M. J. Daniels and M. Pourahmadi. Modeling covariance matrices via partial autocorrelations. `University of Florida working paper`, 2008.

J. H. Dreze and J.-F. Richard. Bayesian analysis of simultaneous equation systems. *Handbook of Econometrics*, 1:517–598, 1983.

P. J. Everson and C. N. Morris. Simulation from wishart distributions with eigenvalue constraints. *Journal of Computational and Graphical Statistics*, 9:380–389, 2000.

J. F. Geweke. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, pages 571–578, 1991.

J. F. Geweke, M. P. Keane, and D. E. Runkle. Statistical inference in the multinomial multiperiod probit model. *Journal of Econometrics*, 80:125–165, 1997.

G. H. Golub and C. F. van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, 1983.

K. Imai and D. van Dyk. A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of Econometrics*, 124:311–334, 2005.

T. Leonard and J. S. Hsu. Bayesian inference for a covariance matrix. *The Annals of Statistics*, 20:1669–1696, 1992.

X. Liu and M. J. Daniels. A new algorithm for simulating a correlation matrix based on parameter expansion and reparameterization. *Journal of Computational and Graphical Statistics*, 15(4): 897–914, 2006.

R. E. McCulloch, N. G. Polson, and P. E. Rossi. A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99:173–193, 2000.

M. K. Munkin and P. K. Trivedi. Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: an application to the demand for healthcare. *Journal of Econometrics*, 114:197–220, 2003.

A. Nobile. Comment: Bayesian multinomial probit models with a normalization constraint. *Journal of Econometrics*, 99:334–345, 2000.

M. Pitt, D. Chan, and R. Kohn. Efficient bayesian inference for gaussian copula regression models. *Biometrika*, 93:537–554, 2006.

M. Pourahmadi. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 87:425–435, 2000.

M. Pourahmadi. Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance correlation parameters. *Biometrika*, 94:1006–1013, 2007.

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

C. Ritter and M. A. Tanner. Facilitating the gibbs sampler: The gibbs stopper and the griddy-gibbs sampler. *Journal of the American Statistical Association*, 87:861–868, 1992.

C. P. Robert. Simulation of truncated normal variables. *Statistics and Computing*, 5:121–125, 1995.

K. Small. The scheduling of consumer activities: Work trips. *American Economic Review*, 72: 467–479, 1982.

L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1761, 1994.

F. Wong, C. Carter, and R. Kohn. Efficient estimation of covariance selection models. *Biometrika*, 90:809–830, 2003.