MCMC Estimation of Restricted Covariance Matrices

Joshua Chi-Chun Chan*
University of Queensland

Ivan Jeliazkov†
University of California, Irvine

February 20, 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.

Appendix A: Proofs

To prove Theorem 1, we begin by computing the determinant of the Jacobian of the transformation considered in Section 2. The result is recorded in the following lemma:

Lemma 1 *Suppose* $\boldsymbol{W}$ *is a* $p \times p$ *positive definite matrix and* $\boldsymbol{W} = \boldsymbol{T}'\boldsymbol{A}\boldsymbol{T}$*, where* $\boldsymbol{T}$ *is a lower triangular matrix whose diagonal elements are all ones and* $\boldsymbol{A}$ *a diagonal matrix with positive diagonal elements. Denote the lower diagonal elements of* $\boldsymbol{T}$ *by* t_{ij}*,* $1 \leqslant j < i \leqslant p$*, and the diagonal elements of* $\boldsymbol{A}$ *by* t_{ii}*,* $i = 1, \ldots, p,$*. Let* $(d\boldsymbol{W})$ *denote the differential form* $(d\boldsymbol{W}) \equiv \bigwedge_{i \geqslant j} dw_{ij}$ *and*

*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.

similarly $(d\boldsymbol{T}) \equiv \bigwedge_{i \geqslant j} dt_{ij}$. *Then we have*

$$(d\boldsymbol{W}) = \prod_{i=1}^{p} t_{ii}^{i-1} (d\boldsymbol{T}).$$

In other words, the determinant of the Jacobian of the transformation from $\boldsymbol{T}'\boldsymbol{A}\boldsymbol{T}$ *to* $\boldsymbol{W}$ *is* $\prod_{i=1}^{p} t_{ii}^{-i+1}$.

Proof of Lemma 1: By definition, we have

$$\begin{pmatrix} w_{11} & w_{21} & \dots & w_{p1} \\ w_{21} & w_{22} & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ w_{p1} & w_{p2} & \dots & w_{pp} \end{pmatrix} = \begin{pmatrix} 1 & t_{21} & \dots & t_{p1} \\ 0 & 1 & \dots & t_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} t_{11} & 0 & \dots & 0 \\ 0 & t_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & t_{pp} \end{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 \\ t_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ t_{p1} & t_{p2} & \dots & 1 \end{pmatrix}.$$

Now we express each w_{ij} in terms of t_{ij}'s and then take differentials (for an introduction to the differential forms approach, see Muirhead, 1982). Since we are going to take the exterior product of these differentials and the exterior products of repeated differentials are zero, there is no need to keep track of differentials in t_{ij} which have previously occurred. In general we get

$$w_{ii} = t_{ii} + \sum_{j=i+1}^{p} t_{ji}^2 t_{jj}, \quad i = 1, \dots, p, \tag{1}$$

$$w_{ij} = t_{ij} t_{ii} + \sum_{k=i+1}^{p} t_{ki} t_{kj} t_{kk}, \quad 1 \leqslant j < i \leqslant p. \tag{2}$$

Taking differentials and ignoring those which have previously occurred, we have

$$\begin{aligned} dw_{pp} &= dt_{pp} \\ dw_{p,p-1} &= t_{pp} dt_{p,p-1} + \dots \\ &\vdots \\ dw_{p1} &= t_{pp} dt_{p1} + \dots \\ dw_{p-1,p-1} &= dt_{p-1,p-1} + \dots \\ &\vdots \\ dw_{11} &= dt_{11} + \dots \end{aligned}$$

Hence taking exterior products gives

$$(d\boldsymbol{W}) \equiv \bigwedge_{i \geqslant j} dw_{ij} = t_{pp}^{p-1} t_{p-1,p-1}^{p-2} \dots t_{22} \bigwedge_{i \geqslant j} dt_{ij}$$

as desired. □

Proof of Theorem 1: Assume the same notation as above. To prove Theorem 1, it is sufficient to consider the case where $t_{ii} \overset{ind}{\sim} \mathcal{G}(\frac{\nu+i-p}{2}, 2)$, $\nu > p$, $i = 1, \ldots, p$, and for each i, we have $t_{ij}|t_{ii} \overset{iid}{\sim} \mathcal{N}(0, t_{ii}^{-1})$, $1 \leqslant j < i \leqslant p$. First note that since $\det \boldsymbol{T} = 1$, we have

$$\det \boldsymbol{W} = \det \boldsymbol{A} = \prod_{i=1}^{p} t_{ii}.$$

Moreover, by (1), we also have

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{W}) &= \sum_{i=1}^{p} w_{ii} \\
&= \sum_{i=1}^{p} t_{ii} + \sum_{i=1}^{p} \sum_{j=i+1}^{p} t_{ji}^2 t_{jj} \\
&= \sum_{i=1}^{p} t_{ii} + \sum_{j=2}^{p} \sum_{i=1}^{j-1} t_{ji}^2 t_{jj} \\
&= \sum_{i=1}^{p} t_{ii} + \sum_{i=2}^{p} \sum_{j=1}^{i-1} t_{ij}^2 t_{ii}
\end{aligned}$$

where we change the order of the double summations in the third equality and interchange the dummy indices $i \leftrightarrow j$ in the last equality. Now, the kernel of the joint density of $\boldsymbol{T}$ and $\boldsymbol{A}$ is

$$\begin{aligned}
&\left(\prod_{i=1}^{p} t_{ii}^{\frac{\nu+i-p}{2}-1} \exp\{-\frac{1}{2} t_{ii}\} \right) \left(\prod_{i=2}^{p} t_{ii}^{\frac{i-1}{2}} \exp\{-\frac{1}{2} \sum_{j=1}^{i-1} t_{ij}^2 t_{ii}\} \right) \\
= &\left(\prod_{i=1}^{p} t_{ii}^{\frac{\nu-p-3}{2}+i} \right) \exp\left\{ -\frac{1}{2} \left(\sum_{i=1}^{p} t_{ii} + \sum_{i=2}^{p} \sum_{j=1}^{i-1} t_{ij}^2 t_{ii} \right) \right\}.
\end{aligned}$$

By Lemma 1, the determinant of the Jacobian is $\prod_{i=1}^{p} t_{ii}^{-i+1}$. Substituting $\operatorname{tr}(\boldsymbol{W})$ and $\det(\boldsymbol{W})$ into the above expression and multiplying the Jacobian, the kernel of the density of $\boldsymbol{W}$ is

$$(\det \boldsymbol{W})^{\frac{\nu-p-1}{2}} \exp\{-\frac{1}{2} \operatorname{tr}(\boldsymbol{W})\},$$

which is the kernel of the Wishart density $\mathcal{W}_p(\nu, \boldsymbol{I}_p)$. □

Proof of Corollary 1: The proof will proceed by construction. By the re-scaling property of the Wishart distribution, if $\boldsymbol{C}'\boldsymbol{C} = \boldsymbol{R}$ and $\boldsymbol{W} \sim \mathcal{W}(\nu, \boldsymbol{I}_p)$, then $\boldsymbol{C}'\boldsymbol{W}\boldsymbol{C} \sim \mathcal{W}_p(\nu, \boldsymbol{R})$. By Theorem 1,

$\boldsymbol{C}'\boldsymbol{W}\boldsymbol{C}$ can be written as $\boldsymbol{C}'\boldsymbol{L}'\boldsymbol{D}^{-1}\boldsymbol{L}\boldsymbol{C}$, where the elements of $\boldsymbol{L}$ and $\boldsymbol{D}$ are distributed as in (1) and (2). At this point, transparency of the construction is greatly improved by choosing $\boldsymbol{C}$ to be a lower triangular matrix, for example by taking $\boldsymbol{C} = \left(\boldsymbol{P}^{-1}\right)'$, where $\boldsymbol{P}$ is the Cholesky factor of $\boldsymbol{R}^{-1}$ such that $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{R}^{-1}$. Since both $\boldsymbol{L}$ and $\boldsymbol{C}$ are lower triangular and the main diagonal of $\boldsymbol{L}$ contains ones, the product $\boldsymbol{L}\boldsymbol{C}$ is a lower triangular matrix with a main diagonal equal to the main diagonal of $\boldsymbol{C}$ and subdiagonal entries that are linear functions of the entries of $\boldsymbol{L}$. This implies that by simple rescaling of the rows of $\boldsymbol{L}\boldsymbol{C}$ by the values on the main diagonal of $\boldsymbol{C}$, we can write $\boldsymbol{C}'\boldsymbol{L}'\boldsymbol{D}^{-1}\boldsymbol{L}\boldsymbol{C} = \widetilde{\boldsymbol{L}}'\widetilde{\boldsymbol{D}}^{-1}\widetilde{\boldsymbol{L}}$, where $\widetilde{\boldsymbol{L}}$ is lower unitriangular and $\widetilde{\boldsymbol{D}}$ is diagonal. Furthermore, because $\widetilde{\boldsymbol{L}}$ contains linear combinations of normal random variables and $\widetilde{\boldsymbol{D}}$ contains rescaled inverse gamma random variables, their elements have Gaussian and inverse gamma distributions, respectively. In essence, by starting with the priors in (1) and (2), transforming them by $\boldsymbol{C}$ we can derive the hyperparameters in priors (3) and (4) necessary to obtain $\boldsymbol{\Sigma}^{-1} \sim \mathcal{W}(\nu, \boldsymbol{R})$. It then goes without saying that any prior hyperparameters in (3) and (4) that do not boil down to those in (1) and (2) after the reverse transformation will produce a $\boldsymbol{\Sigma}^{-1}$ that does not follow the Wishart distribution. Examples include parameters ν_{k0} that vary differently (or not at all) with k, $\{\delta_{k0}\}$ that are not equal to 1, $\boldsymbol{a}_0$ that is distinct from zero, and $\boldsymbol{A}_0$ that does not depend on $\boldsymbol{\lambda}$. □

Appendix B: The ARMH Algorithm

To introduce the ARMH algorithm (Tierney, 1994; Chib and Greenberg, 1995), let $\boldsymbol{\theta}$ be a parameter vector whose density, $\pi(\boldsymbol{\theta})$, is the target density of interest, which is known only up to a normalizing constant and is not easy to simulate. Let $h(\boldsymbol{\theta})$ denote a source (or proposal) density for the ARMH algorithm and let the constant c define the region of domination

$$\mathcal{D} = \{\boldsymbol{\theta} : \pi(\boldsymbol{\theta}) \leqslant ch(\boldsymbol{\theta})\}$$

which is a subset of the support Θ of the target density. Because the domination condition need not be satisfied for all $\boldsymbol{\theta} \in \Theta$, the source density $h(\boldsymbol{\theta})$ is often called a *pseudo-dominating density*. The choice of a pseudo-dominating density is commonly determined by conditioning on the data, other blocks of parameters and latent data as discussed in Section 3; such dependence is implicit

in our discussion, but is suppressed for notational simplicity. Let $\mathcal{D}^c$ be the complement of $\mathcal{D}$, and suppose that the current state of the chain is $\boldsymbol{\theta}$. Then the ARMH algorithm proceeds as follows.

Algorithm 1 ***The accept-reject Metropolis-Hastings (ARMH) algorithm***

1. *A-R step: Generate a draw* $\boldsymbol{\theta}' \sim h(\boldsymbol{\theta})$*; accept* $\boldsymbol{\theta}'$ *with probability* $\alpha_{AR}(\boldsymbol{\theta}') = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')}{ch(\boldsymbol{\theta}'|\boldsymbol{y})}\right\}$. *Continue the process until a draw* $\boldsymbol{\theta}'$ *has been accepted.*

2. *M-H step: Given the current value* $\boldsymbol{\theta}$ *and the proposed value* $\boldsymbol{\theta}'$*:*

   (a) if $\boldsymbol{\theta} \in \mathcal{D}$*, set* $\alpha_{MH}(\boldsymbol{\theta}, \boldsymbol{\theta}') = 1$*;*

   (b) if $\boldsymbol{\theta} \in \mathcal{D}^c$ *and* $\boldsymbol{\theta}' \in \mathcal{D}$*, set* $\alpha_{MH}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{ch(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})}$*;*

   (c) if $\boldsymbol{\theta} \in \mathcal{D}^c$ *and* $\boldsymbol{\theta}' \in \mathcal{D}^c$*, set* $\alpha_{MH}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')h(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})h(\boldsymbol{\theta}')}\right\}$.

   Return $\boldsymbol{\theta}'$ *with probability* $\alpha_{MH}(\boldsymbol{\theta}, \boldsymbol{\theta}')$*; otherwise return* $\boldsymbol{\theta}$.

The ARMH algorithm is an MCMC sampling procedure which nests the accept-reject and MH algorithms when $\mathcal{D}^c$ or $\mathcal{D}$ become empty sets, respectively. But even in the intermediate case when both $\mathcal{D}$ and $\mathcal{D}^c$ are non-empty, ARMH has several attractive features that make it a useful choice for our setting. First, the algorithm is well suited to problems that do not require conjugacy and result in non-standard full-conditional densities, which is the case for the elements of a restricted covariance matrix $\boldsymbol{\Sigma}$. Second, the tuning of an ARMH algorithm can be less demanding and it works reasonably well even if the proposal density $h(\boldsymbol{\theta})$ is only a rough approximation of the target density, as may be the case with standard asymptotic approximating densities (*e.g.*, Zellner and Rossi, 1984). Third, ARMH can produce draws that are closer to iid than those from a similarly constructed MH simulator, but without requiring global domination as the simple accept-reject algorithm. Fourth, the algorithm is useful in sampling covariance matrices because only draws that satisfy positive definiteness, or more stringent eigenvalue constraints as in Everson and Morris (2000), pass through the accept-reject step and continue to the MH step of the sampler. Finally, the building blocks of the ARMH algorithm provide a straightforward way to estimate the marginal likelihood as discussed in Chib and Jeliazkov (2005).

References

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

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

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

R. J. Muirhead. *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc., Baltimore, 1982.

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

A. Zellner and P. E. Rossi. Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics*, 25:365–393, 1984.