Efficient Posterior Simulation for Cointegrated Models with Priors On the Cointegration Space

Gary Koop*
Roberto León-González
Rodney W. Strachan

ABSTRACT: A message coming out of the recent Bayesian literature on cointegration is that it is important to elicit a prior on the space spanned by the cointegrating vectors (as opposed to a particular identified choice for these vectors). In previous work, such priors have been found to greatly complicate computation. In this paper, we develop algorithms to carry out efficient posterior simulation in cointegration models. In particular, we develop a collapsed Gibbs sampling algorithm which can be used with just-identifed models and demonstrate that it has very large computational advantages relative to existing approaches. For over-identifed models, we develop a parameter-augmented Gibbs sampling algorithm and demonstrate that it also has attractive computational properties.

*Gary Koop (Gary.Koop@strath.ac.uk) is at the University of Strathclyde. Roberto León-González (rlg@grips.ac.jp) is at National Graduate Institute of Policy Studies and Rodney Strachan (r.strachan@uq.edu.au) is at University of Queensland. All authors are Fellows of the Rimini Centre for Economic Analysis. Address for correspondence:
Roberto León-González,
National Graduate Institute of Policy Studies (GRIPS)
7-22-1, Roppongi, Minato-Ku, 106-8677, Tokyo, Japan.

1 Introduction

Early Bayesian work on cointegration used Vector Autoregressive (VAR) representations (e.g. DeJong (1992), Dorfman (1994), Koop (1994)), for which simple and standard methods of posterior simulation were available. This early work was criticized by subsequent authors for ignoring the reduced rank structure implied by the cointegrating restrictions. Accordingly, the Vector Error Correction Model (VECM), was increasingly adopted for Bayesian work (see, e.g., Bauwens and Lubrano (1996) and Geweke (1996)). For an $n-$vector of unit root variables, w_t, we write the VECM for $t = 1, .., T$ as:

$$\Delta w_t = \Pi w_{t-1} + \sum_{h=1}^{l} \Psi_h \Delta w_{t-h} + \Phi d_t + \varepsilon_t \tag{1}$$

where ε_t is i.i.d. $N(0, \Sigma)$, the $n \times n$ matrix Π is defined as $\Pi = \alpha\beta'$, α and β are $n \times r$ full rank matrices and d_t denotes deterministic terms (see, e.g., Johansen (1995) pages 81-84 for a commonly-used set of choices). The value of r determines the number of cointegrating relationships. Since crucial issues of identification, prior elicitation and posterior simulation discussed in this paper all relate solely to Π, we will focus on the restricted version of (1):

$$y_t = \alpha\beta' x_t + \varepsilon_t, \tag{2}$$

where $y_t = \Delta w_t$, $x_t = w_{t-1}$. Although this paper discusses a cointegrated model, (2) makes clear that the ideas are of relevance for any model with such a reduced rank structure.

Relative to the VAR, Bayesian inference in the VECM is complicated by the fact that $\Pi = \alpha\beta'$ involves a product of parameters. This precludes direct use of analytical or Monte Carlo integration results for the multivariate linear model. However, once we condition on the cointegrating vectors, β, the otherwise nonlinear VECM becomes a linear one. This means that, under suitable informative priors (e.g. Normal priors of the form used in Geweke (1996)), standard Bayesian analysis of the multivariate linear model applies (conditional on β). This suggests posterior simulation can be done in a straightforward fashion if, as in Geweke (1996), the posterior distribution of β (either a marginal distribution or a distribution conditional on α or on α,Ψ, Φ and Σ where $\Psi = (\Psi_1, .., \Psi_l)$) can be drawn from.

However, the VECM suffers from both a global and local identification problem. A global identification issue can be seen by noting that $\Pi = \alpha\beta'$ and $\Pi = \alpha CC^{-1}\beta'$ are identical for any nonsingular C. This indeterminacy is commonly surmounted by imposing the so-called *linear normalization* where $\beta = \begin{pmatrix} I_r \\ \beta_0 \end{pmatrix}$. Even if global identification is imposed, a local identification issue occurs at the point $\alpha = \mathbf{0}$ (i.e. at this point β does not enter the model). A large literature (e.g. Kleibergen and van Dijk (1994, 1998), Kleibergen and Paap (2002)) discusses problems arising from local non-identification (e.g. lack of existence of posterior moments and lack of convergence of Gibbs samplers using common noninformative priors) and develops other approaches to prior elicitation which surmount these problems. However, the posterior simulation methods used in these papers become more complicated. Furthermore, they all impose global identification through restrictions analogous to the linear normalization.

A literature has recently emerged which argues that it is only the cointegrating space which is identified (see Strachan (2003), Strachan and Inder (2004), Strachan and van Dijk (2004) and Villani (2005, 2006)) and this should be the focus of interest (rather than a particular identified parameter such as β_0). For instance, Strachan and Inder (2004) show how the use of linear identifying restrictions places a restriction on the estimable region of the cointegrating space. Strachan and van Dijk (2004) show that a flat and apparently "noninformative" prior on β_0 in the linear normalization favors regions of the cointegration space near where the linear normalization is invalid. Hence, the linear normalization is used under the assumption that it is valid while at the same time the prior puts weight near the region where the normalization is likely to be invalid.

This recent literature has begun to develop ways of eliciting priors over spaces[1] (as opposed to parameters) and deriving corresponding posterior simulation methods. However, this literature is in its infancy and a good understanding of prior elicitation and efficient posterior simulation have been elusive. A recent survey paper, Koop, Strachan, van Dijk and Villani (2005) describes the development of the Bayesian cointegration literature in detail. The main purpose of our paper is to develop new methods for posterior computation, although we also shed further insight on prior elicitation in the case of the cointegrated model subject to over-identification restrictions. We develop efficient posterior simulation algorithms using a collapsed Gibbs sampler, which adapts the algorithm of Liu (1994) to the present context, and a parameter-augmented Gibbs sampler. The latter is designed for use with the over-identified cointegration model but will also work with the just-identified model. An illustration with simulated data demonstrates the large computational gains achieved by these algorithms.

2 Normalization and Priors in the Just Identified Model

Let $\mathfrak{p} = sp(\beta)$ denote the cointegration space which is an r-dimensional hyperplane in a n-dimensional space. We wish to carry out Bayesian inference relating to this space without imposing identification in such a way as to restrict this space. Furthermore, we wish to develop sensible informative and noninformative priors on this space. To illustrate a key basic idea in a simple case, suppose $n = 2$ and a single cointegrating vector exists. We can parameterize the latter in polar coordinates $\beta = (\cos\theta \ \sin\theta)'$,where $\theta \in [-\pi/2, \pi/2)$. It is only θ

[1] In this paper, a prior over spaces is a prior defined in the Grassmann manifold (Chikuse, (2003), p. 9).

which determines the cointegration space and, thus, we can restrict the length of β to be unity for identification without restricting the cointegration space. A candidate for a noninformative distribution on $\mathfrak{p}$ is the Uniform distribution on θ and this indeed has sensible properties. These points (and many more) are made in Strachan and Inder (2004) and extended to the case where n and r are of higher dimensions. In this general case, the cointegrating space is an element of the Grassmann manifold and, thus, a Uniform prior for the cointegration space is given by the Uniform distribution on the Grassmann manifold. An identification restriction which does not restrict the possible cointegration space is:

$$\beta'\beta = I_r \tag{3}$$

Formally, this restricts the matrix of cointegrating vectors to the Stiefel manifold. These spaces are compact and, hence, a Uniform distribution over them is proper (the integrating constant which ensures propriety is given in Strachan and Inder (2004)). Thus, a noninformative prior for β which does not restrict the cointegrating space is simply equal to an integrating constant with (3) imposed.

To develop an informative prior for the cointegrating space, we introduce a semi-orthogonal matrix H for the purposes of eliciting a prior and adopt the standard notation where $H_\perp$ lies in the orthogonal complement of the space of H. Prior information about the cointegration space can be expressed by eliciting a value for H which spans the space felt, *a priori*, to be most plausible. To obtain H, the researcher will typically first specify a matrix H^g containing desired coefficient values and then use the transformation[2] $H = H^g \left(H^{g\prime}H^g\right)^{-1/2}$. The matrix H constructed in this way will span the same space as H^g but is semi-orthogonal.

For instance, if w_t contains three interest rates of different maturities, then theories of the term structure suggest pairs of them should be cointegrated and, thus:

$$H^g = \begin{pmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}$$

H^g is not semi-orthogonal but $H = H^g \left(H^{g\prime}H^g\right)^{-1/2}$ will be (and will span the same space).

Following Strachan and Inder (2004), as a prior for β and, thus, $\mathfrak{p} = sp(\beta)$ we use a matrix angular central Gaussian distribution with parameter P_τ (i.e. $MACG(P_\tau)$, Chikuse (1990)):

$$p(\beta) \propto |P_\tau|^{-r/2} \left|\beta'(P_\tau)^{-1}\beta\right|^{-n/2} \tag{4}$$

[2] The matrix square root is defined, e.g., in Abadir and Magnus (2005), pages 220-221. This reference also describes a method for its practical calculation.

The $n \times n$ matrix P_τ determines the central location of $\mathfrak{p} = sp(\beta)$ and also the dispersion around the central location. In particular, if we define[3] $P_\tau = HH' + \tau H_\perp H'_\perp$, where τ is a scalar between 0 and 1, the central location of $\mathfrak{p}$ is $\mathfrak{p}^H = sp(H)$ and the dispersion is controlled by the scalar τ. Details are provided in Strachan and Inder (2004), here we note that if $\tau = 1$, then $P_\tau = I_n$ and we have a flat non-informative prior on β (Chikuse, (1990), p. 270). Thus, $\tau = 0$ implies a dogmatic belief that the cointegrating space is $\mathfrak{p}^H$ and $\tau = 1$ implies a noninformative prior (i.e. the prior for $\mathfrak{p}$ is Uniform in the Stiefel manifold when $\tau = 1$). Should the researcher not wish to subjectively elicit a value for τ, she could instead treat τ as an unknown parameter and specify a prior density for it. In practice, such a prior density would typically allocate most weight to values of τ near zero and be restricted to $[0, 1]$.

As a prior for α, in line with previous literature (e.g. Geweke, (1996), Strachan and Inder (2004) and Villani (2005)), we choose a shrinkage prior with zero prior mean:

$$\alpha|\beta, \tau, \Sigma, \nu \sim MN\left(0, \nu\left(\beta' P_{1/\tau}\beta\right)^{-1} \otimes G\right), \tag{5}$$

where MN denotes the matricvariate-Normal distribution (see, e.g., Bauwens, Lubrano and Richard (1999), Appendix A), $P_{1/\tau} = HH' + \tau^{-1} H_\perp H'_\perp$, G is a $n \times n$ matrix[4] and ν is a scalar which controls the degree of shrinkage. The matrix G could be chosen to be equal to Σ, as in Strachan and Inder (2004) and Villani (2005). However, any choice of G is possible should the researcher desire this added flexibility. The scalar ν can either be selected by the researcher subjectively or (5) can be treated as a hierarchical prior with ν being an unknown parameter. Finally, we will use the standard noninformative prior for Σ:

$$p(\Sigma) \propto |\Sigma|^{-(n+1)/2}, \tag{6}$$

although an inverted-Wishart prior can easily be accommodated.

Thus, we have established that the prior specified by (4) through (6) has many sensible properties. Strachan and Inder (2004) provides further discussion and motivation of this prior for the case where $G = \Sigma$. As we shall see in the next section, this prior also allows for simplified computation. Another advantage of this formulation is that the standard noninformative prior for α is found by simply setting $\frac{1}{\nu} = 0$ and, thus, the posterior simulator described in the next section also works with this prior.

[3] Note that this definition of P_τ shrinks the probability mass of $sp(\beta)$ towards $sp(H)$ uniformly in every direction. It does not, for example, allow for tight beliefs about some cointegrating vectors, and vague beliefs about the others. A more general prior can be obtained by defining P to be an unrestricted symmetric positive definite matrix. The modal location of the $sp(\beta)$ is given by the space spanned by the r eigenvectors of P associated with the r largest singular values. The concentration around the mode is controlled by the singular values of P.

[4] Note that $P_{1/\tau} = (P_\tau)^{-1}$

3 Posterior Inference in the Just-Identified Model

A big advantage of the prior given in (4) and (5) is that it allows for efficient and simple posterior computation through use of a collapsed Gibbs sampler (see Liu (1994) and Liu, Wong and Kong (1994)). Note that standard results imply that MCMC draws for Σ can always be obtained from an inverted Wishart:

$$\Sigma|\alpha,\beta,Data \sim IW\left(\left[y - X\beta\alpha'\right]'\left[y - X\beta\alpha'\right], T\right), \tag{7}$$

if G is a fixed known matrix, or if $G = \Sigma$ then the appropriate distribution is:

$$\Sigma|\alpha,\beta,Data \sim IW\left(\left[y - X\beta\alpha'\right]'\left[y - X\beta\alpha'\right] + \nu\alpha\left(\beta' P_{1/\tau}\beta\right)^{-1}\alpha', T + r\right), \tag{8}$$

where y and X are $T \times n$ matrices with t^{th} row given by y_t' and x_t', respectively. IW denotes the inverted-Wishart distribution (see, e.g., Bauwens, Lubrano and Richard (1999), Appendix A). Alternatively, Σ can be integrated out if $G = \Sigma$ (with minor alterations to the algorithm described below). If τ or ν are treated as unknown parameters (as opposed to having values selected for them), then a prior must be selected and an MCMC step which draws τ and ν has to be added. For simplicity, we omit the conditioning arguments Σ, τ and ν in this section and focus on the key issues relating to drawing α and β.

The basic motivation underlying our MCMC algorithm is that computational inefficiencies arise from trying to impose the semi-orthogonality restriction on β. Note that even though α can be easily generated from its Gaussian conditional posterior, the semi-orthogonality restriction implies that the conditional posterior of β is non-standard. To overcome this problem, we introduce the following transformation:

$$\beta\alpha' = (\beta\kappa)\left(\alpha\kappa^{-1}\right)' = \left[\beta\left(\alpha'\alpha\right)^{\frac{1}{2}}\right]\left[\alpha\left(\alpha'\alpha\right)^{-\frac{1}{2}}\right]' \equiv BA', \tag{9}$$

where κ is a positive definite matrix and $A = \alpha\kappa^{-1}$ is semi-orthogonal.

For future reference, we write various relations between the parameters in (9):

$$\begin{aligned} \kappa &= \left(\alpha'\alpha\right)^{\frac{1}{2}} \\ \beta &= B\left(B'B\right)^{-\frac{1}{2}} \\ B'B &= \alpha'\alpha. \end{aligned} \tag{10}$$

Crucially, in the first of these parameterizations (i.e. involving α and β), β is semi-orthogonal while α is unrestricted, whereas in the second (i.e. involving A and B) it is B which is unrestricted whereas A is semi-orthogonal. Our collapsed Gibbs sampler proceeds by switching between these two parameterizations and, hence, it proves useful to express the prior in terms of both. The following proposition, whose proof is in the appendix, states the prior in terms of (A, B).

Proposition 1 *The prior for (α, β) expressed in (4) and (5) implies that the prior for (A, B) is given by:*

$$B|A \sim MN\left(0, \left(A'G^{-1}A\right)^{-1} \otimes \nu P_\tau\right), \tag{11}$$

$$p(A) \propto |G|^{-r/2} \left|A'G^{-1}A\right|^{-n/2} \tag{12}$$

Hence, it can be seen that combining priors (11) and (12) with the likelihood yields a conditional posterior of B that is Normal. After choosing an initial value, $\beta^{(0)}$, our MCMC algorithm repeats the following steps for $s = 1, .., S$:

1. Draw $\alpha^{(*)}$ from $p\left(\alpha|\beta, Data\right)$ and transform this to obtain a draw $A^{(*)} = \alpha^{(*)}\left(\alpha^{(*)\prime}\alpha^{(*)}\right)^{-\frac{1}{2}}$.
2. Draw $B^{(s)}$ from $p\left(B|A^{(*)}, Data\right)$ and then transform this to obtain $\beta^{(s)} = B^{(s)}\left(B^{(s)\prime}B^{(s)}\right)^{-\frac{1}{2}}$ and $\alpha^{(s)} = A^{(*)}\left(B^{(s)\prime}B^{(s)}\right)^{\frac{1}{2}}$.

To see why these conditionals define a collapsed Gibbs sampler, note that (A, κ) is the unique polar decomposition of α (e.g. Cadet (1996)), and therefore the draw of $\alpha^{(*)}$ in step (1) is a draw of $(A^{(*)}, \kappa^{(*)})$ from the joint density[5] $p\left(A, \kappa|\beta, Data\right)$. Similarly, the draw of $B^{(s)}$ in step (2) is a draw of $\left(\beta^{(s)}, \kappa^{(s)}\right)$ from $p\left(\beta, \kappa|A^{(*)}, Data\right)$. Therefore, $A^{(*)}$ in step (1) is a draw from $p\left(A|\beta, Data\right)$ (i.e. obtained marginally on κ), and $\beta^{(s)}$ in the second step is a draw from $p\left(\beta|A^{(*)}, Data\right)$ (i.e. obtained again marginally on κ). Therefore, our sampling strategy is equivalent to the collapsed Gibbs sampler proposed by Liu (1994) and Liu, Wong and Kong (1994, scheme 1), who show that this algorithm is more efficient than a standard Gibbs sampling algorithm (i.e. one which simply draws sequentially from the conditional posteriors $p\left(\alpha|\beta, Data\right)$ and $p\left(\beta|\alpha, Data\right)$). We stress that, because of the semi-orthogonality restriction on β, the latter algorithm does not exist, so our algorithm is likely to be much faster than the existing choices which involve Metropolis-Hastings steps (e.g. Strachan and van Dijk (2004)).

An advantage of this algorithm is that it only involves drawing from the Normal distribution. In particular, in Step 1 $vec\left(\alpha^{*}\right)$ is drawn from the Normal with mean $\overline{\alpha}$ and variance $\overline{\Omega}_\alpha$ where

$$\overline{\Omega}_\alpha = \left(\left[\beta' X' X \beta\right] \otimes \Sigma^{-1} + \frac{1}{\nu}\left[\beta' P_\tau^{-1} \beta\right] \otimes G^{-1}\right)^{-1}$$

and

[5] The polar decomposition decomposes the multivariate variable α into two multivariate variables of smaller dimension each. For example, this is analogous to decomposing a 4 × 1 random vector in $\mathbb{R}^4$ into 2 random vectors of dimension 2 × 1 each.

$$\overline{\alpha} = \overline{\Omega}_{\alpha} vec\left(\Sigma^{-1} y' X \beta\right)$$

In Step 2 $vec\left(B^{(s)}\right)$ is drawn from the Normal with mean $\overline{B}$ and variance $\overline{\Omega}_B$ where

$$\overline{\Omega}_B = \left(\left[A'\Sigma^{-1}A\right] \otimes [X'X] + \left[A'G^{-1}A\right] \otimes \frac{1}{\nu} P_{\tau}^{-1}\right)^{-1}$$

and

$$\overline{B} = \overline{\Omega}_B vec\left(X'y\Sigma^{-1}A\right).$$

Should the researcher wish to treat τ and ν as unknown parameters, then sensible priors for them would be inverted Gamma-2 $IG_2(s_\tau, n_\tau)$ (see, e.g., Bauwens, Lubrano and Richard (1999), Appendix A) for τ and $IG_2(s_\nu, n_\nu)$ for ν. These priors result in simple posterior conditionals: $IG_2(tr[\nu^{-1}A'G^{-1}AB'H_\perp H_\perp' B]+s_\tau, (n-r)r+n_\tau)$ for τ and $IG_2\left(s_\nu + tr\left[A'(G)^{-1}AB'P_{1/\tau}B\right], n_\nu + nr\right)$ for ν. For τ, it will typically make sense to elicit s_τ and n_τ such that virtually all of the prior probability is allocated between 0 and 1.

4 Prior Specification and Posterior Inference in the Over-identified model

It is often the case that there is interest in imposing an over-identifying restriction on the cointegrating space. In this section we consider the restriction $\mathfrak{p} \subseteq \mathfrak{p}^H$ (Johansen 1995, p. 73) and the appendix outlines how to consider other restrictions. This restriction can be imposed by writing $\beta = F\varphi$, where F is a known $n \times s$ semi-orthogonal matrix and φ is an unknown $s \times r$ full rank semi-orthogonal matrix, with $r \leq s \leq n$. The model with this restriction can be written as:

$$y_t = \alpha\varphi'\widetilde{x}_t + \varepsilon_t$$

where $\widetilde{x}_t = F'x_t$. Because α and φ have different dimensions, the algorithm in Section 3 is not applicable. To overcome this problem, let us rewrite the matrix of long-run multipliers as:

$$\varphi\alpha' = \varphi D D^{-1}\alpha' \equiv \widetilde{\varphi}\widetilde{\alpha}'$$

where D is a $r \times r$ symmetric positive definite matrix. We stress that unlike κ, which was defined as one of the components of the polar decomposition of α (Golub and van Loan, 1996, p. 149), the matrix D is not identified. However, the introduction of D facilitates posterior computations because neither $\widetilde{\varphi}$ nor $\widetilde{\alpha}$ are subject to restrictions.

A convenient prior density results from assuming that $\widetilde{\varphi}|\left(\nu,\tau\right)$ follows a priori a $MN(0,s^{-1}I_r\otimes P_\tau)$, and that $vec(\widetilde{\alpha})|\left(\nu,\tau\right)$ is a $N\left(0,\nu I_r\otimes G\right)$. The following proposition summarizes the properties of the prior for the identified parameters (α,φ) and for the non-identified D.

Proposition 2 *The assumptions $\widetilde{\varphi}|\widetilde{\alpha},\nu,\tau\sim MN(0,s^{-1}I_r\otimes P_\tau)$ and $vec(\widetilde{\alpha}|\nu,\tau)\sim N\left(0,\nu I_r\otimes G\right)$ imply that $\varphi|\left(\nu,\tau\right)$ follows a MACG(P_τ) and that $D^2|\left(\varphi,\nu,\tau\right)$ is a Wishart $W_r(s,\left(s\varphi' P_\tau^{-1}\varphi\right)^{-1})$. In addition, the conditional prior mean and variance of α given (φ,ν,τ) are:*

$$\begin{aligned} E(vec(\alpha)|\left(\varphi,\nu,\tau\right)) &= 0 \\ var(vec(\alpha)|\left(\varphi,\nu,\tau\right)) &= \left(\nu\left(\varphi' P_\tau^{-1}\varphi\right)^{-1}\otimes G\right) \end{aligned}$$

Note that the prior mean and variance of $\alpha|\varphi$ is the same as in the prior in Section 2. However, the prior density of α given (φ,ν,τ) is no longer a Normal. In fact, for $r=1$ this distribution is the multivariate Variance-Gamma analyzed by Madan and Seneta (1990). They show that compared to the Normal, the Variance-Gamma distribution puts higher probability near the origin and at the tails, at the expense of probability in the intermediate range.

Note also that if we write $P_\tau=HH'+\tau H_\perp H_\perp^{'}$, with H being a known $s\times r$ semi-orthogonal matrix and $0<\tau<1$, then the prior for $sp(\varphi)$ is centered on $sp(H)$ and the dispersion is controlled by τ (Strachan and Inder, 2004). The following proposition shows that the posterior density of $(\widetilde{\alpha},\widetilde{\varphi})$ is proper if the prior for $(\widetilde{\varphi},\widetilde{\alpha})$ is proper.

Proposition 3 *Let $\widetilde{X}$ be a $T\times s$ matrix containing $\widetilde{x}_t$. Suppose the prior for Σ is given by (6), the prior for $(\widetilde{\varphi},\widetilde{\alpha})$ given G is proper and $\left(\widetilde{X}'\widetilde{X}\right)$ has full rank. Suppose that G is either a fixed known matrix or is equal to Σ. Then the marginal posterior of $(\widetilde{\varphi},\widetilde{\alpha})$ is proper and has at least as many moments as the conditional prior for $(\widetilde{\varphi},\widetilde{\alpha})$ given G.*

However, it can be shown that fixing $1/\nu=0$ results in an improper posterior for the non-identified parameter D. Hence, this framework does not allow for the prior $\pi(\alpha)\propto 1$ in the over-identified case.

The vector $vec(\widetilde{\alpha})$ is drawn conditional on $\widetilde{\varphi}$ from the Normal with mean $\overline{\widetilde{\alpha}}$ and variance $\overline{\Omega}_{\widetilde{\alpha}}$ where

$$\overline{\Omega}_{\widetilde{\alpha}}=\left(\left[\widetilde{\varphi}'\widetilde{X}'\widetilde{X}\widetilde{\varphi}\right]\otimes\Sigma^{-1}+\frac{1}{\nu}I_r\otimes G^{-1}\right)^{-1}$$

and

$$\overline{\widetilde{\alpha}}=\overline{\Omega}_{\widetilde{\alpha}}vec\left(\Sigma^{-1}y'\widetilde{X}\widetilde{\varphi}\right)$$

Similarly, $vec\left(\widetilde{\varphi}\right)$ is drawn from the Normal with mean $\overline{\Phi}$ and variance $\overline{\Omega}_{\Phi}$ where

$$\overline{\Omega}_{\Phi} = \left(\left[\widetilde{\alpha}'\Sigma^{-1}\widetilde{\alpha}\right] \otimes \left[\widetilde{X}'\widetilde{X}\right] + I_r \otimes nP_{\tau}^{-1}\right)^{-1}$$

and

$$\overline{\Phi} = \overline{\Omega}_{\Phi} vec\left(\widetilde{X}'y\Sigma^{-1}\widetilde{\alpha}\right)$$

A sample from the posterior of (α, φ) can be obtained using the following transformations:

$$D = \left(\widetilde{\varphi}'\widetilde{\varphi}\right)^{1/2} \qquad \varphi = \widetilde{\varphi}D^{-1} \qquad \alpha = \widetilde{\alpha}D$$

Note that although this Gibbs algorithm could also be used for the just-identified case, the '$\kappa-$algorithm' of Section 3 would be more efficient for two reasons. Firstly, the $\kappa-$algorithm (implicitly) integrates out the parameter D, and it thereby achieves a comparative advantage (Liu, 1994). Secondly, the $\kappa-$algorithm draws β and A marginally of κ, which is likely to result in smaller autocorrelations in the Markov Chain. We give evidence of the relative performance of these two strategies in the next section.

Note also that the method described in this section is designed to handle over-identifying restrictions on the cointegration space. However, with the obvious alterations, it can also be used to carry out posterior simulation in the VECM subject to linear restrictions on α such as weak exogeneity restrictions.

5 Empirical Evidence on Computational Efficiency

In order to simulate data from model (1) we decompose w_t into two components $w_t = \left(w_{1,t}' \; w_{2,t}'\right)'$ with $w_{1,t} : r\times 1$, $w_{2,t} : (n-r) \times 1$ and use the cointegrating system $w_{1,t} = \beta_0' w_{2,t} + z_{1,t}$, $\Delta w_{2,t} = z_{2,t}$, with the errors generated according to $z_{j,t} = \rho_j z_{j,t-1} + \varepsilon_{j,t}, \; j = 1, 2$ where $\varepsilon_{j,t} \backsim iidN\left(0, (1.5)^2\right)$, $\rho_1 = \rho I_r$, $\rho_2 = 0$ and β_0 is a $(n-r)\times r$ matrix of ones. This specification corresponds to (1) with $\alpha = \left[(\rho - 1) I_r \quad 0_{r\times(n-r)}\right]'$, $\beta = \left[\begin{array}{cc} I_r & -\beta_0' \end{array}\right]'$, $\Phi = 0$ and $\varepsilon_t = \left(\varepsilon_{1,t}' + \varepsilon_{2,t}'\beta_0, \quad \varepsilon_{2,t}'\right)'$.

We first compare the collapsed Gibbs sampler of Section 3 with a Metropolis-Hasting algorithm used in previous literature (Strachan and van Dijk (2004)). We then compare the collapsed Gibbs sampler of Section 3 with the (parameter augmented) Gibbs sampler of Section 4.

5.1 Collapsed Gibbs versus Metropolis-Hasting

For the sake of comparison, consider the following algorithm:

1. Generate $\beta^{(s)}$ from $p\left(\beta|Data\right)$ using a Metropolis-Hasting (MH) algorithm.

2. Draw $\alpha^{(s)}$ from $p\left(\alpha|\beta, Data\right)$.

3. Draw Σ^{s} from $p\left(\Sigma|\alpha, \beta, Data\right)$.

Following Strachan and van Dijk (2004), we specify a random walk proposal density to generate candidates for β. In particular, they suggest a $MACG(P_{z^2})$ (see Chikuse (1990)) proposal density with parameter $P_{z^2} = \beta^{(s-1)}\beta^{(s-1)\prime} + z^2\beta_{\perp}^{(s-1)}\beta_{\perp}^{(s-1)\prime}$, where $\beta^{(s-1)}$ is the value of β obtained in the previous iteration. This proposal density is therefore of the same type as the prior discussed above. For values of z smaller than 1 this density gives more weight to the space spanned by $\beta^{(s-1)}$, and this weight is greater the closer z is to zero (Strachan and Inder (2004)). The variable z is specified to follow a $N(0, \sigma^2)$ and we adjust σ^2 to obtain acceptance rates between 20% and 50% (Chib and Greenberg (1995)). Steps 2 and 3 imply drawing from standard distributions (multivariate Student and inverted Wishart, respectively). For this comparison, we use the noninformative version of the prior.

We compare the efficiency of these algorithms using the effective sample size (e.g. Brooks (1999)) and the average update distance between iterations (Holmes and Held (2005)). The effective sample size measures the number of independent draws from the posterior that is equivalent to n^* draws from an MCMC algorithm. For $n^* = 1$ the effective sample size is defined by:

$$ESS = \frac{1}{1 + 2\sum_{t=1}^{\infty} a(t)}$$

where $a(t)$ is the autocorrelation function at lag t. A value near to one means that the algorithm is as efficient as independent sampling, whereas a value near to zero implies that the algorithm is very inefficient (compared to independent sampling). Since the estimated autocorrelations become imprecise as t gets large, we truncate the sum in the denominator using the initial monotone sequence estimator proposed by Geyer (1992). ESS is calculated for $d(\beta, \beta^*)$, where β^* spans the true cointegrating space (specified above) and $d(\beta, \beta^*)$ is the distance measure between cointegrating spaces proposed by Larsson and Villani (2001). We also report ESS for the $(1, 1)$ element of β_0 (labelled ESS^{Lin} in the tables), where β_0 is the unknown $(n-r) \times r$ matrix when the linear normalization is imposed.

We also calculate the average distance between subspaces sampled in consecutive iterations, defined as:

$$AD = \frac{1}{N-1}\sum_{s=1}^{N-1} d(\beta^s, \beta^{s-1})$$

where N is the number of iterations after the burn-in. Our choice for the burn-in is 300 iterations and we fix N at 15000.

We measure the relative efficiency of our algorithm with respect to the MH algorithm with the ratio of their corresponding ESS and AD values. In particular, we define this ratio as $RESS = ESS_G/ESS_{MH}$, where ESS_G is the ESS value for our algorithm. Similarly, we define the relative gains in update distance as $RAD = AD_G/AD_{MH}$. We compare the algorithms for a range of values of n and r. For each pair (n, r) we generate 100 fictitious samples and run both algorithms for each of these samples. The average values of $RESS$, $RAD, ESS_G, ESS_{MH}, ESS_G^{Lin}, ESS_{MH}^{Lin}$ and their standard deviations together with other summaries are reported in Table 1 for $\rho = 0.3$ and in Table 2 for $\rho = 0.98$. From Table 1, when $n = 2$ and $r = 1$, our algorithm is almost as efficient as independent sampling ($ESS = 0.95$), whereas the MH algorithm is about 8.4 times less efficient (in terms of ESS). Moreover, the relative gains increase as n gets larger. For $n = 6$ and $r = 3$ our algorithm is on average 666 times more efficient, and for $n = 9$ and $r = 5$ it is about 1270 times more efficient. A similar pattern is observed when efficiency measures are calculated for the parameters of the linear normalization (ESS_G^{Lin} and ESS_{MH}^{Lin}) and substantial improvements are also observed in terms of the average update distance. Furthermore, our algorithm computes 15000 iterations slightly more quickly than the MH algorithm (RT), suggesting that the gains adjusted for computation time are even higher. Table 2 shows that when $\rho = 0.98$, which implies that α is close to zero and thus β is weakly identified, the results are similar. Thus, weak identification does not seem to have a significant impact on the speed of convergence. Recall that β is restricted to be semi-orthogonal, and note that the space of semi-orthogonal matrices has a finite measure. Thus, the posterior for β conditional on $\alpha = 0$ is proper, and thus no problems of local non-identification arise. From the computational side, a draw of $\alpha^{(*)}$ (or $B^{(s)}$) close to zero could potentially cause numerical problems at the time of calculating the inverse of $(\alpha'\alpha)$ (or the inverse of $(B'B)$). We did not find this problem in our calculations, but note that A and β can be calculated without using inverses. In particular, if $U : n \times r$, $S : r \times r$, and $V : n \times r$ form the singular value decomposition of $\alpha = USV'$, then A can be calculated as[6] $A = UV'$. Similarly, if U, S, V form instead the singular value decomposition of B, then β can be calculated as $\beta = UV'$.

5.2 Collapsed Gibbs versus parameter augmented Gibbs

We use the same prior for Σ, but use the proper prior for (α, β) described in Sections 3 and 4, with $\tau = 1$, $G = I_n$ and $n_\nu = 2, s_\nu = 1$. This prior allows for a large prior variance for α, since *a priori* $Pr(\nu > 49.75) = 0.01$. Table 3 shows efficiency measures for these two algorithms when $\rho = 0.8$. The collapsed Gibbs is slightly more efficient for some values of (n,r) according to some indicators, but overall it seems that the parameter augmented Gibbs is almost as efficient in practice as the collapsed Gibbs sampler. In our implementation, however, the collapsed

[6] Note that because U and V are semi-orthogonal, $\alpha = (UV')(VSV')$, where UV' is semi-orthogonal and (VSV') is a symmetric positive definite matrix. From the uniqueness of the polar decomposition (e.g. Cadet (1996)), it follows that $A = UV'$ and $\kappa = VSV'$.

Gibbs sampler needed slightly less computing time to do the same number of iterations.

When ρ is close to one, there is little information about the cointegrating space in the data. Thus, the algorithms will explore large regions of the parameter space. In order to illustrate this, we simulate 500 artificial datasets as described above for ($n = 2$,$r = 1$) and several values of ρ. In each of them we calculate the maximum distance (MD) between the draws and the posterior mean of $\mathfrak{p} = sp(\beta)$, where the posterior mean of the space is defined as in Villani (2006). Note that in the case ($n = 2$,$r = 1$) the distance measure is bounded between 0 and 1 (Larsson and Villani (2001)). Table 4 shows that when $\rho = 0.3$ MD is small on average, but when $\rho = 1$ MD is always equal to one, which implies that the algorithm visits all regions of the parameter space.

As an aside, which is unrelated to computational properties, Table 4 also calculates the frequentist coverage of 95% and 99% credible intervals. That is, it calculates the proportion of times that a 95% (or 99%) credible interval contains the true value of the cointegrating space. Note that Bayesian credible intervals are designed to have correct Bayesian coverage (e.g. Raftery and Zheng (2003)), but they often also have correct frequentist coverage (Bernardo and Smith (1994, p. 359)). We construct credible intervals as the set of spaces whose distance to the mean is smaller than a given distance x. For a 95% credible interval, the distance x is the 95% percentile of the distances between the draws and the posterior mean of $\mathfrak{p}$. Table 4 suggests that Bayesian credible intervals have correct frequentist coverage when $\rho = 0.3$ and $\rho = 0.9$. However, when $\rho = 1$ the frequentist coverage of a 95% (or 99%) credible interval is smaller than 95% (or 99%).

6 Conclusions

We have developed efficient and simple algorithms for cointegration models in a framework that allows putting a prior, possibly Uniform, directly on the cointegrating space. This approach avoids the problem that arises when putting a commonly used and tractable prior on the linearly normalized cointegrating vectors (Strachan and van Dijk (2004)), which implies an awkward prior for the cointegrating space. In addition, the approach avoids the problem of non-convergence in the Gibbs sampler caused by local non-identification (Kleibergen and van Dijk (1998)).

References

Abadir, K. and Magnus, J., 2005, *Matrix Algebra.* Cambridge: Cambridge University Press.

Bauwens, L. and Lubrano, M., 1996, Identification restrictions and posterior densities in cointegrated Gaussian VAR systems, in *Advances in Econometrics* 11, Part B, JAI Press, pages 3-28.

Bauwens, L., Lubrano, M. and Richard, J.-F., 1999, *Bayesian Inference in Dynamic Econometric Models.* Oxford: Oxford University Press.

Bernardo, J.M. and Smith, A.F.M. 1994. *Bayesian Theory*, Wiley, Chichester.

Brooks, S., 1999, Bayesian analysis of animal abundance data via MCMC, in *Bayesian Statistics 6* (ed. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith). Oxford: Clarendon Press.

Cadet, A., 1996, Polar coordinates in the R^{np}; Application to the computation of the Wishart and Beta Laws, *Sankhya: The Indian Journal of Statistics* 58, 101-114.

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

Chikuse, Y., 1990, The matrix angular central Gaussian distribution, *Journal of Multivariate Analysis* 33, 265-274.

Chikuse, Y., 2003, Statistics on special manifolds, volume 174 of *Lecture Notes in Statistics*, Springer-Verlag, New York.

DeJong, D., 1992, Co-integration and trend-stationarity in macroeconomic time series," *Journal of Econometrics* 52, 347-370.

Dorfman, J., 1994, A numerical Bayesian test for cointegration of AR processes, *Journal of Econometrics* 66, 289-324.

Geweke, J., 1996, Bayesian reduced rank regression in econometrics, *Journal of Econometrics* 75, 121-146.

Geyer, C.J., 1992, Practical Markov chain Monte Carlo, *Statistical Science* 7, 473-511.

Holmes, C.C. and Held, K., 2004, Bayesian auxiliary variable models for binary and polychotomous regression, forthcoming in *Bayesian Analysis.*

Johansen, S., 1995, *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models.* Oxford: Oxford University Press.

Kleibergen, F. and Paap, R., 2002, Prior, posteriors and Bayes factors for a Bayesian analysis of cointegration, *Journal of Econometrics* 111, 223-249.

Kleibergen, F. and van Dijk, H.K., 1994, On the shape of the likelihood/posterior in cointegration models, *Econometric Theory* 10, 514-551.

Kleibergen, F. and van Dijk, H.K., 1998, Bayesian simultaneous equations analysis using reduced rank structures, *Econometric Theory* 14, 701-743.

Koop, G., 1994, An objective Bayesian analysis of common stochastic trends in international stock prices and exchange rates, *Journal of Empirical Finance* 1, 343-364.

Koop, G., Strachan, R., van Dijk, H.K. and Villani, M., 2005, Bayesian approaches to cointegration. To appear in T.C. Mills and K. Patterson (eds.). *Palgrave Handbook of Theoretical Econometrics*, manuscript available at http://www.le.ac.uk/economics/research/RePEc/lec/leecon/dp04-27.pdf.

Larsson, R. and Villani, M., 2001, A distance measure between cointegration spaces, *Economics Letters* 70, 21-27.

Liu, J.S., 1994, The collapsed Gibbs sampler with applications to a gene regulation problem, *Journal of the American Statistical Association* 89, 958-966.

Liu, J.S., Wong, W.H., and Kong, A., 1994, Covariance structure of the Gibbs sampler with applications to comparisons of estimators and augmentation schemes, *Biometrika* 81, 27-40.

Madan, D.B. and Seneta, E., 1990, The Variance-Gamma (V.G) Model for Share Market Returns, *Journal of Business* 63, 511-524.

Muirhead, R. J. (2005) *Aspects of Multivariate Statistical Theory,* New Jersey: Wiley.

Raftery, A. E., Zheng, Y. 2003. Discussion: Performance of Bayesian Model Averaging. *Journal of the American Statistical Association*, 98, 931-938.

Strachan, R., 2003, Valid Bayesian estimation of the cointegrating error correction model, *Journal of Business and Economic Statistics* 21, 185-195.

Strachan, R. and Inder, B., 2004, Bayesian analysis of the error correction model, *Journal of Econometrics* 123, 307-325.

Strachan, R. and van Dijk, H., 2004, Valuing structure, model uncertainty and model averaging in vector autoregressive processes, Econometric Institute Report EI 2004-23, Erasmus University Rotterdam.

Villani, M., 2005, Bayesian reference analysis of cointegration, *Econometric Theory* 21, 326-357.

Villani, M., 2006, Bayesian point estimation of the cointegration space, *Journal of Econometrics* 134, 645-664.

	$RESS$	RAD	RT	ESS_G	ESS_{MH}	AR	σ	ESS_G^{Lin}	ESS_{MH}^{Lin}
$n = 2,\ r = 1$	8.4 (1.09)	3.7 (0.290)	0.96 (0.0313)	0.943 (0.0523)	0.115 (0.0158)	0.501 (0.110)	0.025	0.93 (0.0703)	0.113 (0.0163)
$n = 3,r = 2$	19.4 (3.99)	6.36 (1.13)	0.97 (0.035)	0.94 (0.069)	0.0504 (0.0109)	0.39 (0.12)	0.020	0.923 (0.0846)	0.0565 (0.012)
$n = 3,r = 1$	31.1 (16.4)	8.59 (3.39)	0.977 (0.038)	0.865 (0.0942)	0.0341 (0.015)	0.341 (0.072)	0.025	0.821 (0.116)	0.0427 (0.0244)
$n = 4,r = 3$	35.7 (8.94)	9.54 (2.45)	0.996 (0.014)	0.938 (0.062)	0.028 (0.006)	0.308 (0.102)	0.020	0.935 (0.080)	0.033 (0.00881)
$n = 4,r = 2$	139 (91.1)	22.9 (13.4)	0.99 (0.013)	0.84 (0.11)	0.009 (0.006)	0.251 (0.068)	0.02	0.802 (0.114)	0.0151 (0.0114)
$n = 4,r = 1$	72 (36.2)	14.8 (5.32)	0.996 (0.077)	0.728 (0.119)	0.0129 (0.0068)	0.296 (0.058)	0.020	0.677 (0.152)	0.0166 (0.0137)
$n = 5,r = 3$	343 (270)	41.1 (25.2)	0.994 (0.0136)	0.81 (0.134)	0.0040 (0.0030)	0.213 (0.0658)	0.016	0.806 (0.133)	0.00973 (0.00757)
$n = 5,r = 2$	353 (253)	41.3 (25.2)	0.991 (0.0076)	0.692 (0.149)	0.00315 (0.0025)	0.235 (0.0642)	0.015	0.66 (0.159)	0.0072 (0.00712)
$n = 6,r = 4$	507 (402)	60.4 (52.0)	1.00 (0.0176)	0.818 (0.123)	0.0027 (0.0018)	0.217 (0.075)	0.012	0.803 (0.113)	0.00648 (0.00502)
$n = 6,r = 3$	688 (490)	65.1 (33.7)	0.995 (0.0158)	0.669 (0.131)	0.00140 (0.00083)	0.235 (0.0625)	0.010	0.703 (0.141)	0.00471 (0.0045)
$n = 9,r = 5$	1220 (918)	191 (124)	1.01 (0.0244)	0.465 (0.150)	0.000521 (0.00027)	0.184 (0.0582)	0.007	0.799 (0.180)	0.00138 (0.000875)

Table 1: Performance measures of our algorithm and the MH algorithm, averaged over 100 fictitious samples for each value of (n, r) $(\rho = 0.3)$. The columns $RESS, RAD, ESS_G, ESS_{MH}, ESS_G^{Lin}, ESS_{MH}^{Lin}$ are defined in the text and standard deviations are in parentheses. RT is the computation time needed for our algorithm to perform 15000 iterations divided by the time needed by the MH algorithm. AR is the acceptance rate in the MH algorithm, which is controlled by σ.

	$RESS$	RAD	RT	ESS_G	ESS_{MH}	AR	σ	ESS_G^{Lin}	ESS_{MH}^{Lin}
$n=2,\ r=1$	7.63 (2.84)	3.70 (1.08)	0.95 (0.055)	0.838 (0.127)	0.122 (0.0367)	0.475 (0.179)	0.5	0.996 (0.0109)	0.725 (0.290)
$n=3$,$r=2$	35.7 (17.9)	9.64 (4.29)	0.98 (0.0293)	0.876 (0.0894)	0.0316 (0.0176)	0.288 (0.105)	0.4	0.999 (0.0108)	0.45 (0.235)
$n=3$,$r=1$	21.2 (11.9)	9.15 (4.60)	0.963 (0.0872)	0.57 (0.162)	0.035 (0.0228)	0.251 (0.12)	0.5	0.988 (0.0429)	0.402 (0.257)
$n=4$,$r=3$	114 (100)	18 (8.58)	0.992 (0.013)	0.916 (0.06)	0.0117 (0.007)	0.231 (0.0837)	0.3	0.999 (0.0105)	0.350 (0.232)
$n=4$,$r=2$	152 (133)	31.1 (23.7)	0.992 (0.016)	0.706 (0.13)	0.0083 (0.0062)	0.196 (0.0810)	0.22	0.999 (0.0226)	0.229 (0.188)
$n=4$,$r=1$	53.4 (33.3)	16.8 (8.14)	0.971 (0.0538)	0.471 (0.152)	0.0112 (0.0060)	0.202 (0.0752)	0.3	0.978 (0.0726)	0.260 (0.223)
$n=5$,$r=3$	345 (222)	43.7 (25.3)	0.994 (0.0168)	0.746 (0.102)	0.00312 (0.00207)	0.219 (0.0744)	0.10	0.998 (0.0269)	0.227 (0.195)
$n=5$,$r=2$	267 (240)	41.2 (29.7)	0.987 (0.0279)	0.576 (0.118)	0.0035 (0.0023)	0.226 (0.081)	0.09	0.999 (0.0123)	0.195 (0.154)
$n=6$,$r=4$	636 (768)	50.3 (28.2)	0.996 (0.034)	0.815 (0.086)	0.0021 (0.0014)	0.300 (0.0847)	0.04	0.999 (0.0079)	0.316 (0.23)
$n=6$,$r=3$	512 (414)	56.9 (38.4)	0.998 (0.015)	0.652 (0.104)	0.0019 (0.0011)	0.274 (0.0811)	0.040	0.999 (0.0067)	0.272 (0.221)
$n=9$,$r=5$	1300 (784)	150 (86.8)	1.00 (0.0082)	0.659 (0.0712)	0.00067 (0.00038)	0.225 (0.0603)	0.040	0.999 (0.0063)	0.178 (0.172)

Table 2: Performance measures when $\rho = 0.98$. Labels defined in Table 1.

	ESS_G	ESS_{PAG}	RT	ESS_G^{Lin}	ESS_{PAG}^{Lin}
$n=3$,$r=2$	0.752 (0.133)	0.424 (0.072)	0.638 (0.018)	0.989 (0.036)	0.993 (0.057)
$n=4$,$r=2$	0.477 (0.136)	0.316 (0.088)	0.624 (0.022)	0.991 (0.038)	1.008 (0.108)
$n=6,\ r=4$	0.573 (0.092)	0.538 (0.091)	0.606 (0.006)	0.999 (0.005)	0.999 (0.006)
$n=9,\ r=5$	0.575 (0.072)	0.595 (0.068)	0.648 (0.006)	0.996 (0.037)	0.999 (0.005)

Table 3: Performance measures averaged over 100 samples: collapsed Gibbs versus parameter augmented Gibbs ($\rho = 0.8$). ESS_G and ESS_G^{Lin} refer to the collapsed Gibbs of Section 3 and ESS_{PAG} and ESS_{PAG}^{Lin} refer to the parameter augmented Gibbs of Section 4. RT is the ratio of computation time of collapsed Gibbs over the other Gibbs. Standard deviations are in parentheses.

	MD_G	MD_{PAG}	COV_G^{95}	COV_{PAG}^{95}	COV_G^{99}	COV_{PAG}^{99}
$\rho = 0.3$	0.124	0.141	0.94	0. 954	0.988	0.988
$\rho = 0.9$	0.987	0.998	0.932	0.962	0.986	0.994
$\rho = 1$	1.000	1.000	0.594	0.694	0.852	0.944

Table 4: Maximum distance and frequentist coverage. The subindex G refers to the algorithm of Section 3, and PAG refers to the algorithm of Section 4. The columns labelled with MD give the average value of the maximum distance. COV^{95} gives the frequentist coverage for 95% credible intervals, and COV^{99} corresponds to 99% credible intervals.

Appendix for Proofs:

Proof of Proposition 1. The prior for (α, β) is proportional to:

$$\left|\beta' P_{1/\tau}\beta\right|^{-n/2} \left|\beta' P_{1/\tau}\beta\right|^{n/2} \exp\left[-\frac{1}{2} tr\left(\nu^{-1}\beta' P_{1/\tau}\beta\alpha' (G)^{-1}\alpha\right)\right]$$

where $\left|\beta' P_{1/\tau}\beta\right|^{-n/2}$ is part of the prior for β and $\left|\beta' P_{1/\tau}\beta\right|^{n/2}$ is part of the conditional prior of α given β. Recall that $\alpha = A\kappa$, where κ is a symmetric positive definite matrix, and let $K = \kappa^2$. Using Lemma 2.1 in Chikuse (1990), this prior for (α, β) implies that the prior for (A, K, β) is proportional to:

$$\exp\left[-\frac{1}{2} tr\left(\nu^{-1}\beta' P_{1/\tau}\beta\kappa A' (G)^{-1} A\kappa\right)\right] |K|^{\frac{n-r-1}{2}}$$

where $|K|^{-\frac{n-r-1}{2}}$ is the Jacobian of the transformation (Muirhead 2005, Theorem 2.1.14). Using the properties of the trace of a matrix, and noting that $B = \beta\kappa$, this expression can be written as:

$$\exp\left[-\frac{1}{2} tr\left(\nu^{-1}A' (G)^{-1} AB' P_{1/\tau}B\right)\right] |K|^{\frac{n-r-1}{2}}$$

Therefore, taking into account that the Jacobian from (A, K, β) to (A, B) is $|K|^{-\frac{n-r-1}{2}}$, the prior for (A, B) is proportional to:

$$\exp\left[-\frac{1}{2} tr\left(\nu^{-1}A' (G)^{-1} AB' P_{1/\tau}B\right)\right]$$

Since $(P_\tau)^{-1} = P_{1/\tau}$, this shows that the prior of B given A is given by (11). If we integrate with respect to B, we get the prior for A :

$$\left|\nu\left(A' (G)^{-1} A\right)^{-1} \otimes P_\tau\right|^{1/2} = |P_\tau|^{r/2} \left|\nu\left(A' (G)^{-1} A\right)^{-1}\right|^{n/2} \propto \left|A' (G)^{-1} A\right|^{-n/2}$$

which is proportional to a $MACG(G)$. ■

Proof of Proposition 2. From Theorem 2.2 in Chikuse (1990), $\varphi|(\nu, \tau)$ follows a MACG(P_τ). To prove that $D^2|(\varphi, \nu, \tau)$ follows a Wishart, note that the joint prior of $(\widetilde{\varphi}, \widetilde{\alpha})|(\nu, \tau)$ is proportional to:

$$\pi(\widetilde{\varphi}, \widetilde{\alpha}|\nu, \tau) \propto \exp\left(-\frac{1}{2} tr\left(s\widetilde{\varphi}' P_\tau^{-1}\widetilde{\varphi}\right)\right) \exp\left(-\frac{1}{2} tr\left(\nu^{-1}\widetilde{\alpha}' G^{-1}\widetilde{\alpha}\right)\right) \tag{13}$$

Note that $\widetilde{\varphi} = \varphi D$ and that $tr\left(s\widetilde{\varphi}' P_\tau^{-1}\widetilde{\varphi}\right) = tr\left(s\widetilde{\varphi}\widetilde{\varphi}' P_\tau^{-1}\right) = tr\left(sD^2\varphi' P_\tau^{-1}\varphi\right)$. Recall also that the Jacobian from $(\widetilde{\varphi}, \widetilde{\alpha})$ to $(\varphi, D^2, \widetilde{\alpha})$ is $\left|D^2\right|^{(s-r-1)/2}$ (Muirhead 2005, Theorem 2.1.14). Thus, (13) implies that $\pi(\varphi, D^2, \widetilde{\alpha}|\nu, \tau)$ is proportional to:

$$\pi(\varphi, D^2, \widetilde{\alpha}|\nu, \tau) \propto \left|D^2\right|^{(s-r-1)/2} \exp\left(-\frac{1}{2} tr\left(sD^2\varphi' P_\tau^{-1}\varphi\right)\right) \exp\left(-\frac{1}{2} tr\left(\nu^{-1}\widetilde{\alpha}' G^{-1}\widetilde{\alpha}\right)\right)$$

which shows that $D^2|(\varphi, \nu, \tau)$ is a $W_r(s, \left(s\varphi' P_\tau^{-1}\varphi\right)^{-1})$. To calculate the first moments of α, note that $\alpha = \widetilde{\alpha}D$, and that $E(\widetilde{\alpha}|D) = 0$. Thus, $E(\widetilde{\alpha}D|D) = E(\alpha|D) = 0$. By the law of iterated expectations, $E(\alpha) = 0$. Next from $vec(\alpha) = vec(\widetilde{\alpha}D) = (D' \otimes I_n)vec(\widetilde{\alpha})$ and $var(vec(\widetilde{\alpha})|D, \varphi, \tau, \nu) = \nu(I_r \otimes G)$, it follows that $var(vec(\alpha)|D, \varphi, \tau, \nu) = \nu(D^2 \otimes G)$. By the law of iterated expectations, and the properties of the Wishart distribution, $var(vec(\alpha)|\varphi, \tau, \nu) = (\nu(E(D^2|\varphi, \tau, \nu) \otimes G)) = \nu(\left(\varphi' P_\tau^{-1}\varphi\right)^{-1} \otimes G)$ ■

Proof of Proposition 3. The proof consists in showing that the integral of the posterior with respect to Σ is never larger than a constant times the prior. Hence, a proper prior results in a proper posterior. Let us first consider the case in which G is a known matrix. Let $\pi(\widetilde{\alpha}, \widetilde{\varphi})$ be the prior of $(\widetilde{\alpha}, \widetilde{\varphi})$. The integral of the posterior with respect to Σ is:

$$\pi(\widetilde{\alpha}, \widetilde{\varphi}|Data) \propto \left| \left[y - \widetilde{X}\widetilde{\varphi}\widetilde{\alpha}'\right]' \left[y - \widetilde{X}\widetilde{\varphi}\widetilde{\alpha}'\right] \right|^{-T/2} \pi(\widetilde{\alpha}, \widetilde{\varphi})$$

which can be written as:

$$\left| \left(\widetilde{\varphi}\widetilde{\alpha}' - \widehat{\Pi}\right)' \Omega^{-1} \left(\widetilde{\varphi}\widetilde{\alpha}' - \widehat{\Pi}\right) + y'y - \widehat{\Pi}'\Omega^{-1}\widehat{\Pi} \right|^{-T/2} \pi(\widetilde{\alpha}, \widetilde{\varphi}) \tag{14}$$

where:

$$\begin{aligned} \Omega &= (\widetilde{X}'\widetilde{X})^{-1} \\ \widehat{\Pi} &= \Omega\widetilde{X}'y \end{aligned}$$

Expression (14) is never greater than:

$$\left| y'y - \widehat{\Pi}'\Omega^{-1}\widehat{\Pi} \right|^{-T/2} \pi(\widetilde{\alpha}, \widetilde{\varphi}) \tag{15}$$

Thus, if $\pi(\widetilde{\alpha}, \widetilde{\varphi})$ is proper then the posterior is proper. When $G = \Sigma$ the integral of the posterior with respect to Σ is:

$$\pi(\widetilde{\alpha}, \widetilde{\varphi}|Data) \propto \left| \left[y - \widetilde{X}\widetilde{\varphi}\widetilde{\alpha}'\right]' \left[y - \widetilde{X}\widetilde{\varphi}\widetilde{\alpha}'\right] + \widetilde{\alpha}'\widetilde{\alpha} \right|^{-(T+r)/2} \pi(\widetilde{\varphi})$$

which, using the same reasoning as before, is never greater than:

$$\left| y'y - \widehat{\Pi}'\Omega^{-1}\widehat{\Pi} + \widetilde{\alpha}\widetilde{\alpha}' \right|^{-(T+r)/2} \pi(\widetilde{\varphi}) \tag{16}$$

which is the product of a matricvariate Student Distribution (Bauwens et al. 1999, p. 307) with T degrees of freedom times the prior of $\widetilde{\varphi}$. Hence, the integral of (16) is finite, and thus the posterior is proper. ■

Appendix for other over-identifying restrictions

Consider the restriction (e.g. (Johansen 1995, p. 73)) $\beta = \begin{pmatrix} F_1\varphi_1 & F_2\varphi_2 \end{pmatrix}$, where F_1 and F_2 are known matrices. We can partition α conformably as: $\alpha = \begin{pmatrix} \alpha_1 & \alpha_2 \end{pmatrix}$, so that

$$\beta\alpha' = F_1\varphi_1\alpha_1' + F_2\varphi_2\alpha_2'$$

In a similar manner to Section 4, we introduce non-identified conformable squared matrices D_1 and D_2 as follows:

$$\begin{aligned} F_1\varphi_1\alpha_1' + F_2\varphi_2\alpha_2' &= F_1\varphi_1 D_1 D_1^{-1}\alpha_1' + F_2\varphi_2 D_2 D_2^{-1}\alpha_2' \\ &= F_1\widetilde{\varphi}_1\widetilde{\alpha}_1' + F_2\widetilde{\varphi}_2\widetilde{\alpha}_2' \end{aligned}$$

If Normal priors are chosen for $(\widetilde{\alpha}_1, \widetilde{\alpha}_2, \widetilde{\varphi}_1, \widetilde{\varphi}_2)$ then the Gibbs sampling algorithm consists in sampling alternatively $(\widetilde{\alpha}_1, \widetilde{\alpha}_2)|(\widetilde{\varphi}_1, \widetilde{\varphi}_2)$ and $(\widetilde{\varphi}_1, \widetilde{\varphi}_2)|(\widetilde{\alpha}_1, \widetilde{\alpha}_2)$, with both conditionals being Normal.