BAYESIAN POINT ESTIMATION OF THE COINTEGRATION SPACE

MATTIAS VILLANI

ABSTRACT. A neglected aspect of the otherwise fairly well developed Bayesian analysis of cointegration is point estimation of the cointegration space. It is pointed out here that, due to the well known non-identification of the cointegration vectors, the parameter space is not Euclidean and the loss functions underlying the conventional Bayes estimators are therefore questionable. We present a Bayes estimator of the cointegration space which takes the curved geometry of the parameter space into account. This estimate has the interpretation of being the posterior mean cointegration space and is invariant to the order of the time series, a property not shared with many of the Bayes estimators in the cointegration literature. An overall measure of cointegration space uncertainty is also proposed. Australian interest rate data are used for illustration. A small simulation study shows that the new Bayes estimator compares favorably to the maximum likelihood estimator.

KEYWORDS: Bayesian inference, Cointegration analysis, Estimation, Grassman manifold, Subspaces.

JEL CLASSIFICATION: C11, C13, C32.

1. INTRODUCTION

Building on the work of Kleibergen and van Dijk (1994), Bauwens and Lubrano (1996) and Geweke (1996), the Bayesian analysis of cointegration has recently been developed by, for example, Kleibergen and Paap (2002), Strachan (2003), Strachan and Inder (2004), Strachan and van Dijk (2004) and Villani (2000, 2005) into a fairly complete alternative to the more established classical approaches in *e.g.* Phillips (1991) and Johansen (1995).

The focus in the above mentioned works is on developing suitable prior distributions and deriving the corresponding posterior distributions. The efforts put into this activity have diverted attention from other important aspects of the analysis, some of which have been largely unexplored. One such aspect is the topic here: how to best summarize the posterior distribution of the cointegration vectors by a small number of quantities, such as measures of location and spread. This problem usually goes under the heading of point estimation.

One obvious reason for such a reduction in information is that it can be useful as a reporting device. The ideal Bayesian analysis displays the posterior distribution under a wide range of priors, which is very effective in bringing out the full inferential content in the data, and at the same time allows the user to use his own prior in processing the results. The mapping from the set of prior distributions to the set of posterior distributions is of course too complex to handle unless both the prior and posterior distribution are summarized compactly.

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden and Sveriges Riksbank, SE-103 37 Stockholm, Sweden. E-mail: mattias.villani@riksbank.se. This work has benefitted from valuable discussions with Daniel Thorburn and comments from two anonymous referees. The author gratefully acknowledges financial support from the Swedish Council of Research in Humanities and Social Sciences (HSFR), grant no. F0582/1999 and the Swedish Research Council (Vetenskapsrådet) grant no. 412-2002-1007. The views expressed in this paper are solely the responsibility of the author and should not be interpreted as reflecting the views of the Executive Board of Sveriges Riksbank.

In many models, the location of the posterior distribution is adequately summarized by, for example, the posterior mean or median. The rationale underlying these estimates is that they minimize posterior expected loss, each for a particular type of loss function (Bernardo and Smith, 1994). The loss function penalizes the discrepancy between the unknown model parameter and its estimate. This type of *estimation loss* may be motivated by the reporting objective stated above. In line with much of the Bayesian literature in this area, the term *Bayes estimator* will be used to describe the rule for computing the optimal posterior estimate under a particular estimation loss, despite the firm roots of the term 'estimator' in non-Bayesian statistics.

It is well known that the cointegration vectors are only identified up to arbitrary linear combinations; that is, it is only the space spanned by them, the so called *cointegration space*, which we are able to infer from data. This paper argues that, since we are estimating subspaces rather than real numbers, it no longer makes sense to use the traditional loss functions, like the quadratic loss, on the unrestricted parameters of the cointegration vectors. Thus, traditionally used loss functions should be replaced by functions measuring the distance between subspaces, and the traditional estimators replaced by new ones derived from these subspace loss functions.

We present results for one such subspace distance: the projective Frobenius distance. The optimal measure of location for this loss is very easily computed and has the nice interpretation of being the posterior mean cointegration space, rather than the posterior mean of the elements of the cointegration vectors under a more or less arbitrarly chosen identification scheme. Furthermore, a scalar-valued measure of overall cointegration space variation is proposed. An interesting feature of this uncertainty measure is that it is independent of both the dimension of the system and the number of cointegration vectors, which makes it comparable across model specifications.

There is an earlier literature on the effect of normalization on point estimates in simultaneous equations models. Fisher (1976) gives an example where Bayesian point estimates depend on the choice of normalization. In a comment on Fisher's paper, Kadane (1978) notes that this non-invariance stems from using a normalization dependent prior distribution, which in turn implies that also the posterior depends on the normalization. So in Fisher's example the Bayes estimates are non-invariant to normalization as a result of the posterior lacking this invariance property. One of the points in this paper is that many of the Bayes estimates used in the cointegration literature are *not* normalization invariant *even if* the posterior distribution is.

The results in this paper are not tied to a particular identifying scheme or prior distribution. This is important as both these two choices are somewhat controversial and in some respects still open issues (see *e.g.* Kleibergen and Paap (2002) and Villani (2005) on the choice of prior, and Strachan's (2003) discussion of different identifying schemes).

The paper is organized as follows. The next section gives the notation for the cointegrated VAR model and discusses the geometry of the parameter space of the cointegration vectors. The problems with the currently available Bayes estimators are then discussed and illustrated by some simple examples in Section 3. Section 4 proposes a new Bayes estimator of the cointegration space and the following section gives corresponding measures of cointegration space variation. Section 6 contains an empirical illustration and Section 7 a small simulation study where the new Bayes estimator is compared to the maximum likelihood estimator in Johansen (1995). The last section gives some concluding remarks.

2. The geometry of the parameter space in cointegration models

The *error correction model* is of the form

$$\Delta x_t = \Pi x_{t-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta x_{t-i} + \varepsilon_t, \tag{2.1}$$

where x_t is a vector of observations on p time series at time t, Π, $\Gamma_1, ..., \Gamma_{k-1}$ are $p \times p$ matrices of dynamic coefficients and ε_t is a sequence of disturbances. The time series are assumed to be integrated of order 1 such that $\{x_t\}$ contains unit roots, but $\{\Delta x_t\}$ is stationary. All results here holds for more general specifications of the model with additional exogenous variables and deterministic components possibly restricted to the cointegration space, and under any assumption about the disturbances.

Under the assumption of independent $\varepsilon_t \sim N_p(0, \Sigma)$ and the usual vague Jeffreys prior for the multivariate regression, the marginal posterior distribution of Π is a matrix t density (Zellner, 1971), *if* Π in (2.1) is assumed to be of full rank. In this case every element of Π is identified and the usual loss functions may be used to derive optimal posterior location estimates of Π, such as the posterior mean for a quadratic loss. A key assumption is the full rank of Π. Relaxing this requirement opens up a range of useful models, but at the same time prompts a rethinking of many aspects of the analysis, such as loss functions for point estimation.

Let us now assume that Π is of reduced rank $r < p$. Then, by the Granger representation theorem, there exists r stationary linear combinations of the otherwise non-stationary time series in x_t (Engle and Granger, 1987). Furthermore, this allows us to decompose Π as $\Pi = \alpha\beta'$, where α and β are both $p \times r$ matrices. The model in (2.1) may now be rewritten as

$$\Delta x_t = \alpha\beta' x_{t-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta x_{t-i} + \varepsilon_t,$$

where $\beta' x_{t-1}$ is an r-vector of stationary departures from the r long run equilibria and α governs the adjustment back to equilibrium after a disturbance. The columns of β determine the long run equilibria and are called *cointegration vectors*.

It is well known that the likelihood function is invariant under the class of transformations $T_Q : (\alpha, \beta) \rightarrow (\alpha Q'^{-1}, \beta Q)$, where Q is a non-singular $r \times r$ matrix; this is usually phrased as: only the space spanned by the columns of β, the *cointegration space*, is identified. In order to obtain a unique estimate of β, some of its elements have to be restricted to known values, usually zero or one, in such a way that the mapping from the remaining unrestricted elements of β to the cointegration spaces is one-to-one. A particularly simple set of restrictions is the *linear normalization* $\beta' = (I_r, B')$, where B is a $(p-r) \times r$ matrix of unrestricted elements. It is clear that the identification problem is symmetric in α and β. Exactly the same analysis applies if identifying restrictions are placed on α rather than β, which is a natural choice in simultaneous equations models (Geweke, 1996). The case of identifying restrictions on both α and β is not treated here.

Focusing on the unrestricted parameters in a given identifying scheme is convenient in that the parameter space is the familiar Euclidean space. It should be clear, however, that the use of identifying restrictions does not change the fact that the underlying parameter of the model is the cointegration space as a whole. Thus, the 'deep' parameter space is not Euclidean, but the abstract space of all r-dimensional subspaces of $\mathbb{R}^p$, the *Grassman manifold* $\mathbb{G}_{r,p-r}$. It is of course possible to work with the unrestricted parameters under an identifying scheme while at the same keeping track of the implied inferences in $\mathbb{G}_{r,p-r}$ space. For example, Villani (2005) develops a prior in the linear normalization which implies that β is marginally uniformly distributed on the Grassman manifold. A similar approach to point estimation would require the development of loss functions on B which remain sensible when transformed to $\mathbb{G}_{r,p-r}$ (see the last paragraph of Section 4). The reason for using the linear normalization is that the posterior distribution may then be evaluated by simple Gibbs sampling. When it comes to developing point estimators of β, however, there are no advantages in confining the analysis to a particular identification scheme (the posterior may still be obtained under such an identification scheme, see Section 4). In fact, identifying restrictions and normalizations tend to complicate the geometry of the parameter space, thereby making it hard to come up with suitable loss functions for the unrestricted parameters. Working directly on the Grassman manifold, *i.e.* in terms of subspaces, gives a clearer understanding of the problem and we will do so in this paper. The relevant statistical theory on the Grassman manifold can be found in *e.g.* James (1954), Eaton (1983) and Wijsman (1990). Edelman et al. (1998) give an enlightening discussion of the geometry of the Grassman manifold.

$\mathbb{G}_{r,p-r}$ is a (analytic) manifold of dimension $r(p-r)$. An important property of a manifold is that it behaves locally like the Euclidean space. A useful mental picture is the unit sphere (a two-dimensional manifold embedded in $\mathbb{R}^3$) which at every point may be locally approximated by a plane. The current state of maximum likelihood analysis in cointegration models relies almost exclusively on asymptotic results. The fact that the Grassman manifold is locally Euclidean justifies the use of statistical theory for Euclidean spaces in the asymptotic analysis of β. Bayesian analysis, on the other hand, claims to be applicable for all sample sizes and as such is concerned with global properties for which we are no longer able to rely on theory developed for Euclidean spaces.

3. Traditional Bayes estimators of the cointegration space

Despite the curved geometry of the parameter space, the large majority of Bayesian applications have estimated the cointegration space by the posterior mean, mode or median of the free elements in β after the identifying restrictions have been imposed. The most common approach is to insert the posterior median of B into $\beta' = (I_r, B')$; the posterior mean of B does not exist for some priors (Bauwens and Lubrano, 1996; Kleibergen and van Dijk, 1994; Villani (2005)).

Many of these plug-in estimators are not invariant to the way the variables are ordered. This is true even if the posterior distribution itself exhibits this invariance property (Kleibergen and Paap (2002) and Villani (2005) prove that their posteriors are invariant). This is illstrated in the following example.

Example 1. Figure 1 contains two cointegration vectors, β_1 and β_2 which are assumed to have the same posterior probability. Let us further assume that all other vectors have zero posterior probability. The argument which we are about to make is easily seen to hold for other, more realistic, posteriors. The thick vector β_m in Figure 1 is a reasonable location estimate for this two-point distribution. Consider now the linear normalization $\beta = (1, B^I)'$, normalized on the first variable. This normalization restricts the cointegration vectors to the vertical dashed line in Figure 1, and is labeled normalization I. The two vectors β_1 and β_2 in normalization I are given by their projections on the vertical dashed line (points B_1^I and B_2^I, respectively). The mean and median of B^I are both equal to $B_m^I = (B_1^I + B_2^I)/2$, using the usual convention to handle ties, and the corresponding plug-in estimator of β is β_m^I. If, on the other hand, the linear normalization $\beta = (B^{II}, 1)'$ had been chosen (*i.e.* normalizing on the second variable, restricting the cointegration vector to the horizontal dashed line), the posterior mean or median of B^{II} would instead be $B_m^{II} = (B_1^{II} + B_2^{II})/2$ and β_m^{II} the plug-in estimator. Since

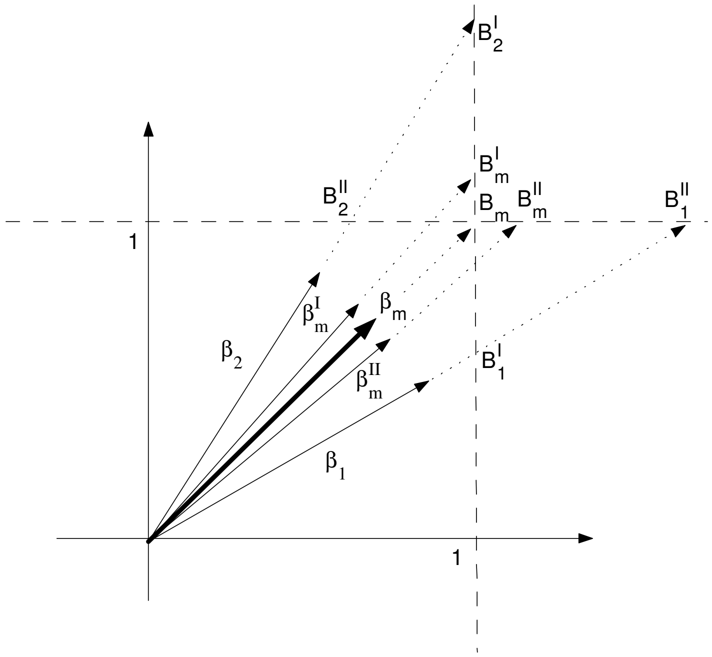

FIGURE 1. Lack of invariance of point estimates to variable order: illustration for Example 1.

$\beta_m^I \neq \beta_m^{II}$, the plug-in estimators of the cointegration space are not invariant to normalization. Furthermore, both of them differs from the estimate β_m suggested by intuition. The trouble with the plug-in estimators is of course that rely on Euclidean metrics, like the quadratic and the absolute value, for measuring distances between cointegration vectors. They therefore judge *e.g.* that the distance from B_1^I to B_m is smaller than that of B_2^I to B_m, which clearly is inappropriate (the angle between β_1 and β_m is the same as the angle between β_2 and β_m).

Another possibility is to use the posterior mode of B as a plug-in estimate in $\beta = (I_r, B')'$. Rather than maximizing the full joint posterior, B is sometimes estimated by numerical maximization of the fairly low-dimensional marginal posterior of B (which is available in closed form under certain priors, see e.g. Bauwens and Lubrano (1996) and Villani (2005)) or directly from the posterior sample of B. The empirical example in Section 6 shows, however, that this *marginal mode* plug-in estimator is, in general, not invariant to variable order.

The estimator in Strachan (2003) is based on the identifying scheme in Anderson (1951) and Johansen (1995), which he terms the *non-ordinal normalization*. In the case of a single cointegrating relation, the non-ordinal normalization restricts β to lie on a hemisphere with a fixed radius. Strachan (2003) estimates the cointegration vector with the posterior mean of β, which exists in the non-ordinal normalization. This estimator is clearly invariant to variable order, but, as the next example illustrates, may generate counterintuitive results.

Example 2. Assume that we have the situation depicted in Figure 2, where all posterior mass is distributed equally on the two vectors $\beta_1 = (b, \sqrt{1-b^2})'$ and $\beta_2 = (b, -\sqrt{1-b^2})'$, where $0 \leq b \leq 1$; the linear normalization $\beta = (1, B)'$ is also given in the figure by the vertical line tangent to the circle. Such a posterior distribution would clearly not be encountered in practise, but it caricatures less extreme situations that do occur in applications, at least when b is small. For small values of b, we clearly have $\beta_1 \approx (0, 1)'$ and $\beta_2 \approx (0, -1)'$, which both say that the second variable is stationary, since the sign of the second coefficient in β

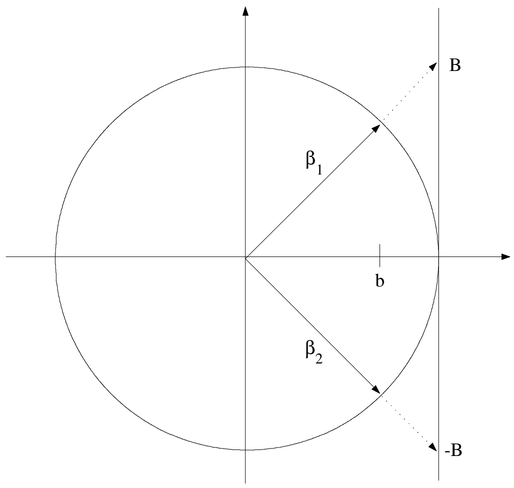

FIGURE 2. Implications of the non-standard geometry of the parameter space for point estimators of the cointegration vectors: illustration for Example 2.

does not matter when the first coefficient is zero. On the other hand, as $b \to 1$, we have $\beta_1 \to (1,0)'$ and $\beta_2 \to (1,0)'$, both implying that the first variable is stationary. It is easy to see that the median and mean plug-in estimators of β are all equal to $(1,0)'$. Strachan's (2003) posterior mean estimate of β in the non-ordinal normalization is $(\beta_1+\beta_2)/2 = (b,0)'$, which after normalization becomes $(1,0)'$. Thus, previously suggested estimators estimates β with $(1,0)'$, regardless of b, which goes against one's intuition. The reason for the counterintuitive behavior of the traditional estimators of the cointegration space is of course that they fail to acknowledge that two cointegration vectors (or two matrices of cointegration vectors) may have widely differing unrestricted coefficients (*e.g.* B in the linear normalization) and still span cointegration spaces which are very close to each other.

An objection to the above reasoning about the central role of the cointegration space may be that in some situations the unrestricted elements of β *are* the object of analysis, *e.g.* B in the linear normalization $\beta' = (I_r, B')$, and that it therefore makes sense to report summarizing measures directly on them; Campbell and Shiller (1987) and Paap and van Dijk (2003) analyze models with this property. Standard loss function penalizes the distance between the estimate of B and the true B and may make sense locally, *i.e.* when the posterior is concentrated in a small neighbourhood of the parameter space, but will of course suffer globally from the same problems discussed above. It is not an easy task to come up with loss functions directly on B which avoid the above mentioned problems (see the end of the next section, however). This is especially true for $r > 1$. The next section describes an approach where losses are instead defined directly on the cointegration space. This leads to an estimate of the cointegration space which may then be mapped to the unrestricted parameters in any normalization so that, for example, the posterior estimate of B may be reported.

4. THE POSTERIOR MEAN COINTEGRATION SPACE ESTIMATOR

We shall assume that a posterior distribution $p(\beta|\mathcal{D})$ for the matrix of cointegration vectors is available, perhaps obtained from one of the numerical algorithms proposed in the references in the introduction. The ensuing analysis works directly on the Grassman manifold, *i.e.* the object of analysis is the subspace spanned by β, the cointegration space. Technically this means that β will only appear as a part of the projection matrix $\beta(\beta'\beta)^{-1}\beta'$ defining the subspace

$\text{sp}\,\beta$. The choice of normalization in the computation of the posterior distribution of β is therefore immaterial. We will however assume, for clarity and without any loss of generality, that $\beta'\beta = I_r$. The projection matrix for $\text{sp}\,\beta$ then simplifies to $\beta\beta'$ and unnecessary clutter in the formulae is avioded. If the posterior of β has been obtained using a different normalization, one may simply make the transformation $\beta \to \beta(\beta'\beta)^{-1/2}$. The posterior distributions in all Bayesian analyses of cointegration to date have been evaluated numerically by Monte Carlo sampling methods, so the transformation to orthonormality is conveniently performed on each draw. We hasten to note that the restriction to orthonormal β does not entirely identify the model (the matrix βQ, where Q is an $r \times r$ orthonormal matrix, is orthonormal and determines the same cointegration space as β). Since our analysis works with subspaces (or projection matrices), this lack of identification does not matter here.

We shall now consider the question: given a (posterior) distribution for β, what single point $\hat{\beta} \in \mathbb{G}_{r,p-r}$ is the, in some sense, best summary of this distribution? The Bayesian solution to this problem is given by (Bernardo and Smith, 1994)

$$\hat{\beta} \stackrel{def}{=} \underset{\tilde{\beta} \in \mathbb{G}_{r,p-r}}{\operatorname{argmin}} E[l(\beta, \tilde{\beta})],$$

where $E(\cdot)$ denotes the posterior expectation and $l(\beta, \tilde{\beta})$ is a loss function on $\mathbb{G}_{r,p-r} \times \mathbb{G}_{r,p-r}$. Although there are many distances on the Grassman manifold (Edelman et al, 1998) which may be used to construct a loss function, we shall here restrict attention to the square of the *projective Frobenius distance*

$$l(\beta, \tilde{\beta}) = ||\beta\beta' - \tilde{\beta}\tilde{\beta}'||^2,$$

where $||A|| = \text{tr}(A'A)^{1/2}$ is the usual Frobenius norm for matrices. Note that $\beta\beta'$ is the projection matrix for the subspace $\text{sp}\,\beta$ and recall that a subspace is uniquely determined by the orthogonal projection onto it. $l(\beta, \tilde{\beta})$ is therefore obtained by embedding the Grassman manifold in the set of $p \times p$ projection matrices of rank r and then using the Frobenius norm. The projective Frobenius distance is one of the most widely used distances between subspaces and has the additional advantage of leading to a simple analytical expression for $\hat{\beta}$, as the next result shows. We will refer to $\hat{\beta}$ under the projective Frobenius distance as the posterior mean cointegration space (PMCS) estimator. The following theorem was proved independently by Srivastava (2000) and Villani (2000).

Theorem 4.1. *The posterior mean cointegration space (PMCS) estimator is*

$$\hat{\beta} = (v_1, ..., v_r),$$

where v_i is the eigenvector of $E(\beta\beta')$ corresponding to the ith largest eigenvalue.

We want to stress that the restriction to orthonormal β is made for clarity in presentation and does not tie the above results to a particular normalization. If the posterior draws of β are not transformed to be orthonormal, then the PMCS estimator is the matrix of eigenvectors of $E[\beta(\beta'\beta)^{-1}\beta']$ corresponding to the r largest eigenvalues.

It is well known that the posterior distribution of B in the linear normalization is integrable, but lacks integer moments of any order (Bauwens and Lubrano, 1996; Kleibergen and van Dijk, 1994). The reason for this is that B is a matrix quotient of submatrices of β and therefore has a posterior distribution with heavy tails. The existence of the PMCS estimator, on the other hand, hinges upon the existence of $E(\beta\beta')$, where β is an orthonormal matrix. Using that the Grassman manifold is a compact space, Strachan and van Dijk (2004) show that all finite moments of the elements of β exist in the orthonormal normalization, which of course implies the existence of $E(\beta\beta')$. This result is proved for a particular prior (uniform distribution on the Grassman manifold for β and a normal distribution for the adjustment coefficients in α), but holds generally as long as the marginal posterior of β is bounded above almost everywhere (this is evident from the proof in Strachan and van Dijk (2004)). In particular, it holds for the prior $p(\alpha, \beta, \Gamma, \Sigma) \propto |\Sigma|^{-(p+1)/2}$, which is the limit of the prior in Strachan and van Dijk (2004) when $\tau \to 0$ (their notation).

A closed form expression for $E(\beta\beta')$ may not be available, but a numerical approximation may be used in its place. For example, importance sampling (Kloek and van Dijk, 1978; Geweke, 1989) or the Gibbs sampler (Tierney, 1994) are two common algorithms used to generate draws from the distribution of β. These generated matrices can subsequently be made orthonormal and the following well-known result (Tierney, 1994) may be used to estimate $E(\beta\beta')$

$$\frac{1}{N}\sum_{i=1}^{N}\beta^{(i)}\beta^{(i)\prime} \overset{a.s.}{\to} E(\beta\beta'),$$

where N denotes the number of posterior draws, $\beta^{(i)}$ denotes the ith sampled matrix after the transformation to orthonormality and $\overset{a.s.}{\to}$ denotes almost sure convergence.

We now return to the two examples of the previous section. In the first example (Figure 1), it is easily established that the PMCS estimate equals the common sense estimate β_m. In the second example (Figure 2), we recall that all previously proposed estimators returned the estimate $(1,0)'$ regardless of the value of b. The PMCS estimator tells a completely different story. The two eigenvalues of $E(\beta\beta')$ are $2(1-b^2)$ and $2b^2$, with corresponding eigenvectors $(0,1)'$ and $(1,0)'$. Thus, $\hat{\beta} = (0,1)'$ if $b < 1/2$ and $\hat{\beta} = (1,0)'$ if $b > 1/2$, exactly as suggested by intuition.

It is illuminating to look at the implied loss function on B in the linear normalization when the projective Frobenius loss is used for β. This may be derived in the general case, but is more transparent in the simple case with $p = 2$ and $r = 1$. Straightforward algebra shows that the implied distance on B is of the form

$$d_B(B, \tilde{B}) = \frac{2(B-\tilde{B})^2}{(1+B^2)\,(1+\tilde{B}^2)}.$$

The numerator of $d_B(B, \tilde{B})$ is proportional to the quadratic loss for B. The denominator of the implied loss makes sure that a given discrepancy $B - \tilde{B}$ is less penalized if B and $\tilde{B}$ are large in absolute values compared to when they are small. This is very natural as a given discrepancy when both B and $\tilde{B}$ are large (in absolute value) does not produce as large discrepancy between the cointegration spaces generated by B and $\tilde{B}$ as when B and $\tilde{B}$ are small.

5. Measures of cointegration space variation

The usual measures of spread of the unrestricted coefficients in β are easily computed numerically by sampling from the distribution of β. Due to the curved geometry of the parameter space in cointegration models, they may however be of limited value in assessing the variation of $\operatorname{sp}\beta$, at least for moderately informative distributions.

A quite different measure of variation suggests itself from Theorem 4.1. Let $\lambda_1 \geq \cdots \geq \lambda_p$ denote the eigenvalues of $E(\beta\beta')$. Since λ_i measures the variation of $\operatorname{sp}\beta$ in the direction determined by v_i, $\lambda_1, ..., \lambda_r$ can be used to assess the uncertainty regarding $\operatorname{sp}\beta$. A natural suggestion for an *overall* measure of variation of $\operatorname{sp}\beta$ to accompany the PMCS estimate is given in the following definition.

Definition 5.1. *The projective Frobenius span variation is defined as*

$$\tau^2_{\mathrm{sp}\,\beta} \stackrel{def}{=} \frac{E[l(\beta,\hat{\beta})]}{r(p-r)/p},$$

where $\hat{\beta}$ *is the PMCS estimate of* β *and* $l(\cdot,\cdot)$ *is the projective Frobenius distance.*

Lemma 5.1.

$$\tau^2_{\mathrm{sp}\,\beta} = \frac{r - \sum_{i=1}^{r} \lambda_i}{r(p-r)/p},$$

where λ_i *is the* i*th largest eigenvalue of* $E(\beta\beta')$.

Proof. Follows from Proposition A.4 in Lütkepohl (1993, Section A.14). □

The following theorem shows that the lower and upper bound of $\tau_{\mathrm{sp}\,\beta}$ do not depend on either the number of time series in the system or the cointegration rank, which facilitates comparisons across studies. In addition, the maximal value of $\tau_{\mathrm{sp}\,\beta}$ is obtained under what is usually referred to as the uniform distribution on the Grassman manifold (James, 1954; Mardia and Khatri, 1977).

Theorem 5.1. $\tau_{\mathrm{sp}\,\beta}$ *satisfies*

$$0 \leq \tau_{\mathrm{sp}\,\beta} \leq 1$$

and the upper bound is obtained when $\mathrm{sp}(\beta)$ *is distributed according to the (unique) invariant Haar distribution on* $\mathbb{G}_{r,p-r}$.

Proof. The non-negativity of $\tau_{\mathrm{sp}\,\beta}$ follows directly from Definition 5.1 and the non-negativity of the projective Frobenius distance. From Lemma 5.1, $\tau_{\mathrm{sp}\,\beta}$ is maximal when $\sum_{i=1}^{r} \lambda_i$ is minimal. Note that $\lambda_1, ..., \lambda_p$ are constrained to satisfy the equation

$$\sum_{i=1}^{p} \lambda_i = \mathrm{tr}[E(\beta\beta')] = E[\mathrm{tr}(\beta'\beta)] = E[\mathrm{tr}(I_r)] = r,$$

in addition to the order restriction. One immediate candidate as a minimizer of $\sum_{i=1}^{r} \lambda_i$ is

$$\lambda_i = \frac{r}{p}, \text{ for } i = 1, ..., r,$$

resulting in $\sum_{i=1}^{r} \lambda_i = r^2/p$; we shall now prove that this is indeed the minimum by first proving that $\sum_{i=1}^{r} \lambda_i > r^2/p$ for all ordered sequences $\lambda_1 \geq ... \geq \lambda_p$ of eigenvalues where $\lambda_{r+1} \neq r/p$. First, if $\lambda_{r+1} < r/p$, then $\lambda_i < r/p$ for $i = r+2, ..., p$, by the order restriction, with the result that $\sum_{i=1}^{r} \lambda_i = r - \sum_{i=r+1}^{p} \lambda_i > r - (p-r)r/p = r^2/p$. On the other hand, if $\lambda_{r+1} > r/p$, then $\lambda_i > r/p$ for $i = 1, ..., r$, again using the order restriction, with the result that $\sum_{i=1}^{r} \lambda_i > r^2/p$. Having established that λ_{r+1} must equal r/p in the minimizing sequence, it follows that $\lambda_i = r/p$, for $i = 1, ..., r$, minimizes $\sum_{i=1}^{r} \lambda_i$. The upper bound now follows from Lemma 5.1

$$\max \tau^2_{\mathrm{sp}\,\beta} = \frac{r - \sum_{i=1}^{r} r/p}{r(p-r)/p} = 1.$$

In remains to be shown that the upper bound of $\tau_{\mathrm{sp}\,\beta}$ is obtained when $\mathrm{sp}(\beta)$ follows the Haar invariant distribution on $\mathbb{G}_{r,p-r}$. Mardia and Khatri (1977) show that $E(\beta\beta') = (r/p)I_p$ if $\mathrm{sp}(\beta)$ follows the Haar invariant distribution on $\mathbb{G}_{r,p-r}$ and thus $\lambda_i = r/p$, for $i = 1, ..., r$, which in turn implies that $\tau_{\mathrm{sp}\,\beta} = 1$. □

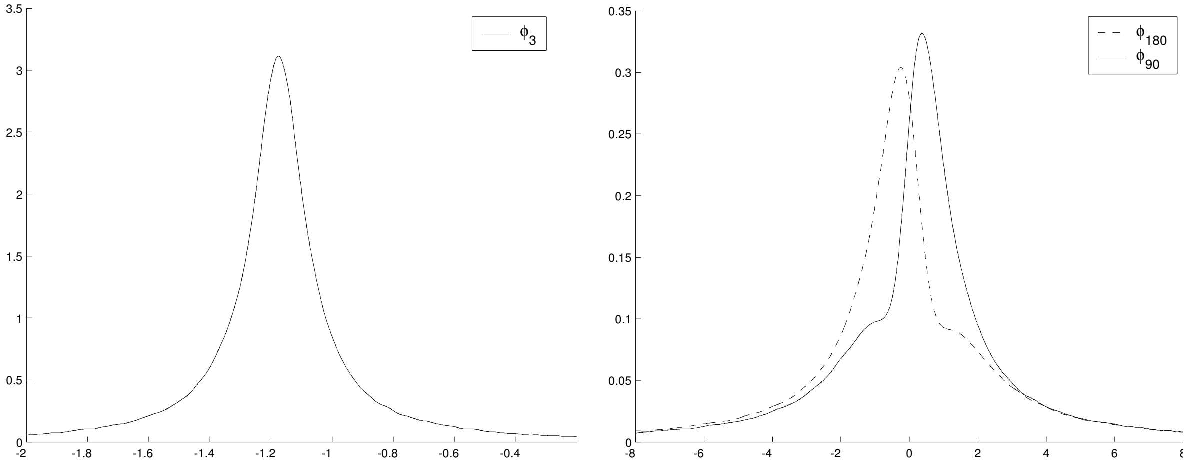

FIGURE 3. Posterior distribution of the cointegration vector $\beta = (1, \phi_3, \phi_{90}, \phi_{180})$.

6. Empirical illustration

The Australian interest rates data in Strachan (2003) will be used for illustration. The data consist of 94 monthly observations on four Australian interest rates of different maturity during the time period 1994:1-2001:10. Two of the rates are taken from the longer part of the yield curve with 5 and 3 years to maturity (i_5 and i_3) and the other two are shorter rates with 180 and 90 days to maturity (i_{180} and i_{90}), respectively. Following Strachan (2003), we condition our analysis on two lagged differences in the error correction model and a single cointegration vector. The cointegrating relation is normalized on the five year bond rate, and is of the form

$$i_5 + \phi_3 i_3 + \phi_{180} i_{180} + \phi_{90} i_{90}.$$

The model also contains an unrestricted constant term. We refer to Strachan's paper for more details and some discussion of the theory of the term structure of interest rates which motivates the empirical analysis.

The posterior distribution of the unrestricted elements of $\beta = (1, \phi_3, \phi_{180}, \phi_{90})'$ is computed using the marginal Gibbs sampler in Villani (2005, Theorem 4.6) under the prior $p(\alpha, \beta, \Gamma, \Sigma) \propto |\Sigma|^{-(p+1)/2}$. The results based on $300,000$ draws (including $10,000$ burn-in iterations) are displayed in Figure 3. Note how the marginal posterior of ϕ_{180} and ϕ_{90} are nearly mirror images of each other, reflecting that the two short rates most probably enter the cointegrating relation as a difference, but the coefficient on this spread relative the longer rates is much less clearly determined in the data. The strong correlation between ϕ_{180} and ϕ_{90} slows down the convergence of the Gibbs sampler. The numerical stability is characterized in Figure 4, where the evolution of the median estimates is plotted as function of the number of Gibbs iterations. The precision in the estimates is sufficient for our purposes here. Given the strong linear dependence between ϕ_{180} and ϕ_{90}, a natural way forward would be to restrict the two short rates to have the same coefficient with opposite signs. This restriction is testable and easily analyzed by posterior odds ratios (Villani, 2000; Strachan, 2003), but this analysis lies outside the focus of this paper and will not be pursed here.

Table 1 gives the maximum likelihood (ML) and the PMCS estimates along with the median and marginal mode plug-in estimates (see Section 3). The median is computed from the Gibbs sample whereas the marginal mode is computed by numerical optimization (using the Matlab

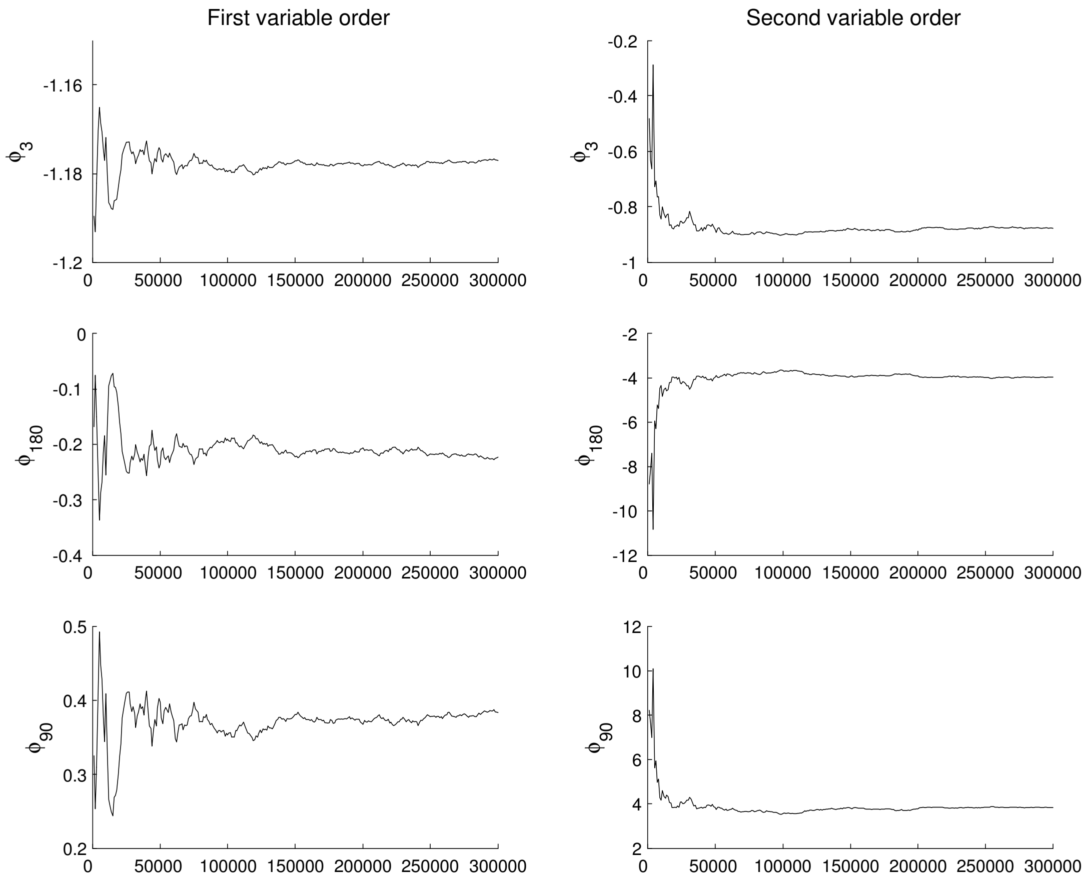

FIGURE 4. Cumulative estimates of the posterior median as a function of the number of Gibbs sampling draws.

routine *fmincon*) of the analytical expression for the marginal posterior of B given in Bauwens and Lubrano (Theorem 3.1). Several different sets of initial values gave essentially the same results. The posterior mode plug-in estimate (obtained by numerical maximization of the joint posterior mode of α and B) is equal to the ML estimate under this particular prior. Strachan's (2003) estimate under a vague prior is also presented. To show the effects of a different order of the time series, the estimates for the reversed order are also displayed. The PMCS estimator is clearly invariant to the way the variables have been ordered, but has been computed under both variable order to indicate the precision in the Gibbs sampling. The last row of Table 1 contains the projective Frobenius distance between each of the estimates and the ML estimate.

From Table 1 it is seen that the PMCS and maximum likelihood estimates are very close to each other and relatively far off the other three estimates. The discrepancy between the median and marginal mode estimators in the second variable order is smaller than appears at first sight (the projective Frobenius distance between the two is 0.467); both estimates convey more or less that the spread between the short rates is stationary. The complete opposite message is obtained from both the median and marginal mode estimates under the first variable order. Here the estimates point toward the conclusion that the two long rates are cointegrated.

			Order $i_5, i_3, i_{180}, i_{90}$			Order $i_{90}, i_{180}, i_3, i_5$		
Coefficient	ML	Strachan	PMCS	Mode	Median	PMCS	Mode	Median
ϕ_3	-1.026	-1.100	-1.014	-1.179	-1.177	-1.011	-2.879	-0.878
ϕ_{180}	-2.022	-12.376	-2.183	-0.005	-0.223	-2.202	22.731	-3.970
ϕ_{90}	2.040	12.263	2.186	0.180	0.383	2.207	-20.983	3.826
$l(\hat{\beta}, \hat{\beta}_{ML})$	-	0.522	0.044	1.210	1.054	0.050	0.745	0.320

Table 1. Estimates of the cointegration vector normalized on i_5. The column named 'mode' is the point estimate based on the mode in the marginal posterior of B. The last row of the table displays the projective Frobenius distance between the estimate and the ML estimate.

The projective Frobenius span variation is $\tau_{\mathrm{sp}\,\beta} = 0.757$, roughly three quarters between the degenerate and the uniform distribution on the Grassman manifold, indicating a rather uninformative posterior distribution of the cointegration vector.

7. A small simulation study

This section investigates the performance of the PMCS estimator in repeated sampling by simulation methods. In order to attract the attention of practitioners with preferences toward likelihood procedures, we will compute the expectation of $E(\beta\beta')$ (see Theorem 4.1) with respect to the normalized likelihood function. The resulting PMCS estimator may be called the generalized likelihood estimator or the mean likelihood estimator.

The data generating process is chosen to be the bivariate VAR(1) with a single cointegration vector

$$\begin{aligned}\Delta x_{1t} &= \alpha_1(\beta_1 x_{1,t-1} + \beta_2 x_{2,t-1}) + \varepsilon_{1t} \\ \Delta x_{2t} &= \alpha_2(\beta_1 x_{1,t-1} + \beta_2 x_{2,t-1}) + \varepsilon_{2t},\end{aligned}$$

where $(\varepsilon_{1t}, \varepsilon_{2t})'$ are independent bivariate normal vectors with zero mean and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}.$$

The cointegration vector is parameterized in polar coordinates

$$\beta = \begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} = \begin{pmatrix}\cos\theta \\ \sin\theta\end{pmatrix}, \quad -\pi/2 < \theta \leq \pi/2.$$

This simple process has two related advantages. First, the number of parameters is small enough to enable a large part of the parameter space to be covered in the simulations. Second, the marginal normalized likelihood function of the unrestricted elements of the cointegration vector is one-dimensional and may therefore be evaluated numerically over a grid without having to recourse to more advanced numerical procedures which would be prohibitively time-consuming in a simulation study.

We consider all possible combinations of $\theta \in \{0, \pi/4, \pi/2\}$, $\alpha_1 \in \{-.25, -.20, -.15, -.10, -.05\}$, $\alpha_2 \in \{-.25, -.15, -.10, -0.075, -.05, -.025\}$, $\sigma \in \{.25, 1, 3\}$ and $\rho \in \{-.7, 0, .7\}$, all in all 810 combinations of parameters values.

Note that the single non-zero eigenvalue of $\Pi = \alpha\beta'$ is $\alpha_1 \cos\theta + \alpha_2 \sin\theta$. When $\alpha_1 = \alpha_2 = 0$, the two time series are $I(1)$, but not cointegrated ($r = 0$). Table 2 lists some important characteristics of the three cointegration vectors. The last column in particular gives the

θ	β	eig(Π)	$I(2)$-ray in (α_1, α_2)-space
0	$(1,0)'$	α_1	$\alpha_1 = 0$ and $\alpha_2 \neq 0$
$\pi/4$	$(1,1)'$	$2^{-1/2}(\alpha_1+\alpha_2)$	$\alpha_1 = -\alpha_2$ and $\alpha_2 \neq 0$
$\pi/2$	$(0,1)'$	α_2	$\alpha_1 \neq 0$ and $\alpha_2 = 0$

Table 2. Characteristics of the three cointegration vectors used in the simulation study.

subspace in (α_1, α_2)-space where the process is integrated of second order $I(2)$, *i.e.* where it requires differencing twice to become stationary.

200,000 data sets are simulated for each parameter setting and the maximum likelihood (ML) and PMCS estimates computed for each generated data set. Two different sample sizes, $n = 25$ and $n = 50$ are used. These sample sizes are probably a fair representation of the information typically available in empirical studies where the sample sizes are usually larger, but the data much less 'tidy' than those resulting from our generating models.

The efficiency of the PMCS estimator relative the maximum likelihood estimator is measured by

$$RE = \frac{\text{Mean distance between PMCS estimate and true cointegration vector}}{\text{Mean distance between ML estimate and true cointegration vector}},$$

where the distance between an estimate of the cointegration vector and the true value is measured by the arc length distance (Edelman et al, 1998); other distances, including the projective Frobenius distance, led to essentially the same results. $RE < 1$ indicates that the PMCS estimator outperforms the ML estimator and for $RE > 1$ the opposite holds.

We will only present the simulation results for the case $\rho = 0$; the results for $\rho = -0.7$ and $\rho = 0.7$ are qualitatively similar and may be obtained from the author by request. A graphical overview of the results is presented in Figure 5 ($n = 25$) and Figure 6 ($n = 50$); detailed information may be obtained from the author. Each subgraph in Figure 5 and 6 gives the simulation results for a particular combination of cointegration vector (θ) and error standard deviation (σ). The horizontal axis in each subgraph measures α_1 and the vertical axis measures α_2. The grayness of a rectangle indicates the gain in estimation precision from using the PMCS estimator rather the ML estimator. Note that the settings where the ML estimator is better than the PMCS estimator ranges from very light gray to white.

A general observation from Figure 5 is that the PMCS estimator outperforms the ML estimator in a large majority of the parameter settings, sometimes to the extent of a 40% improvement in RE. Furthermore, in those cases where the ML estimator performs better than the PMCS estimator, the improvement is always modest (RE is always smaller than 1.07). The differences between the two estimators diminishes as the sample increases from 25 to 50 observations (Figure 6), but the PMCS estimator is still substantially better in some parameter settings with the larger sample size.

The PMCS estimator performs particularly well compared to the ML estimator in the upper right and lower left subgraphs of Figure 5 and 6. Both these settings correspond to a data generating process where data are relatively informative regarding θ. To see this, note that the cointegration vector with $\theta = 0$ is the vector $(1, 0)$, implying that the first variable in the system, x_1, is stationary by itself. Also, σ controls the variation in x_2 relative to that of x_1. A large value of σ therefore makes it relatively easy to separate the stationary x_1 from the non-stationary x_2, which in turn leads to a precise estimate of the cointegration vector.

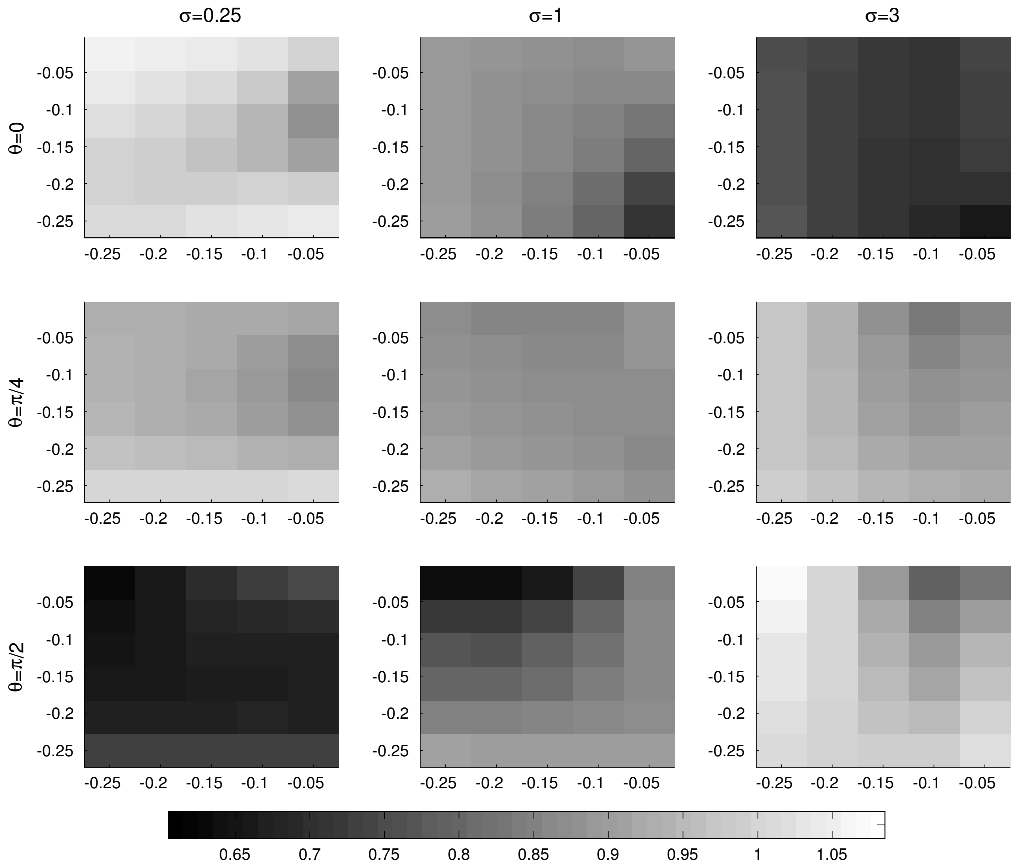

FIGURE 5. Relative efficiency (RE) of the PMCS estimator compared to the maximum likelihood estimator for $\rho = 0$ and $n = 25$. $RE < 1$ indicates a superior performance of the PMCS estimator. Each subgraph displays the RE for a particular combination of cointegration vector (θ) and error standard deviation σ. In each subgraph, α_1 is measured along the horizontal axis and α_2 along the vertical axis.

When $\theta = \pi/2$, on the other hand, the cointegration vector is $(0, 1)$ implying that the second variable is stationary. Here the opposite applies: a small value of σ produces informative data on average.

Another observation from Figure 5 and 6 is that the best performance of the PMCS compared to the ML estimator occurs for the parameter settings where the process is close to being $I(2)$ (see Table 2; note that when $\theta = \pi/4$, the upper right corner of the plots lies closest to the $I(2)$-ray). This is an interesting result in the light of the empirical findings that some commonly used macroeconomic variables, such as price indicies, are close to $I(2)$ (*e.g.* Juselius, 1992). Future research should explore if the same result holds in larger systems.

In the region near the point where the process is $I(1)$ without cointegration ($\alpha = (0, 0)'$), the PMCS estimator is more accurate than the ML estimator for all but a few of the (θ, σ)-pairs. This effect is not as important as the $I(2)$ effect, however. Thus, the local identification problem in the point $\alpha = (0, 0)'$ and the resulting non-standard behavior of the likelihood (first documented by Kleibergen and van Dijk, 1994) seems to have only a minor effect on the relative performance of the two estimators.

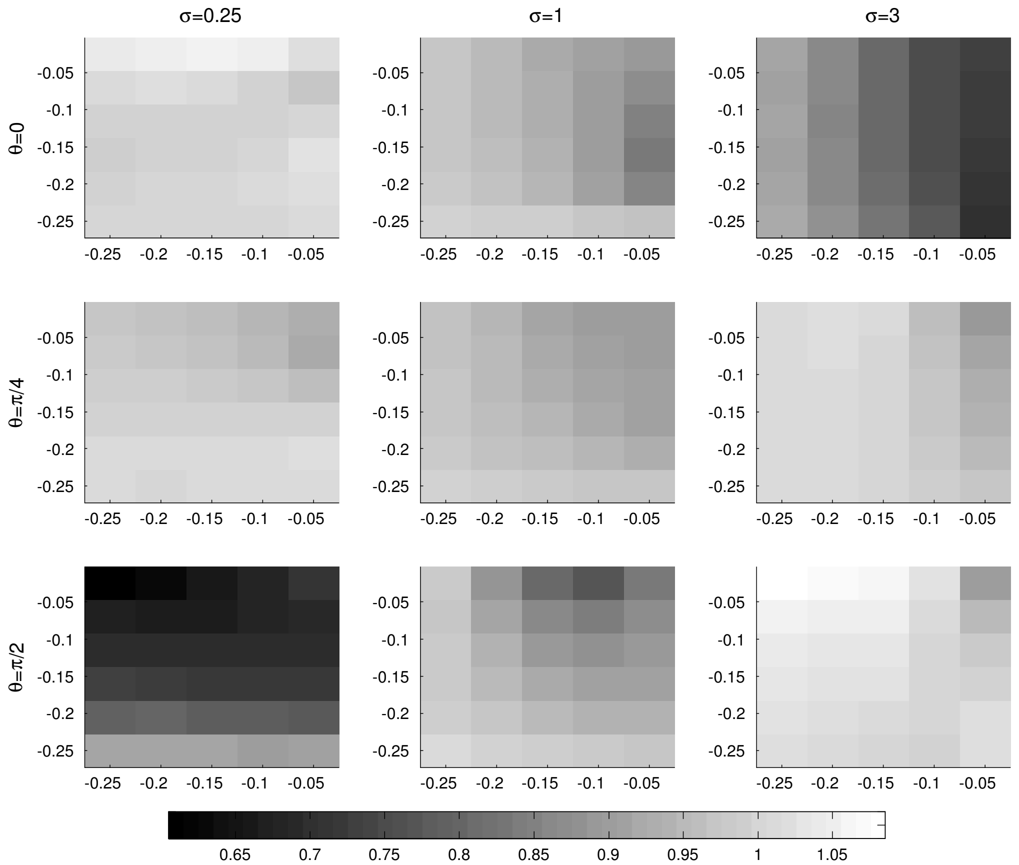

FIGURE 6. Relative efficiency (RE) of the PMCS estimator compared to the maximum likelihood estimator for $\rho = 0$ and $n = 50$. $RE < 1$ indicates a superior performance of the PMCS estimator. Each subgraph displays the RE for a particular combination of cointegration vector (θ) and error standard deviation σ. In each subgraph, α_1 is measured along the horizontal axis and α_2 along the vertical axis.

8. Concluding remarks

It should be clear that the results here apply to any situation where a part of the parameter space is the Grassman manifold, *i.e.* subspace estimation problems. This includes of course the reduced rank regression model in Anderson (1951), but also the common factor model (Anderson, 1984) extensively used in psychometrics, the simultaneous equations model (Geweke, 1996; Kleibergen and van Dijk, 1998), error in variable models (Zellner, 1971) and many other widely used models in multivariate analysis.

We have focused here on the just-identified case, which is the starting point of most analyses. When the same over-identifying restrictions are imposed on all r cointegration vectors, the parameter space of the remaining unrestricted elements is a Grassman manifold of smaller dimension and the results here apply directly. This is a situation of substantial practical interest; it covers all linear restrictions when $r = 1$, the frequently occurring case where one or several variables are assumed not to enter any of the r cointegrating relations and many other situations. It would of course be nice to extend the results to general linear over-identifying restrictions, but this leads to complicated optimization problems which do not seem to have a closed form solution. It should be kept in mind, however, that it is less important to take the correct geometry into account when the cointegration vectors are heavily restricted.

It would be interesting to conduct the type of analysis presented here for other distance measures on the Grassman manifold and compare the resulting estimators and variation measures. Edelman *et al* (1998) lists six common distances, including the projective Frobenius, but, at least in some cases, approximate or numerical solutions may be needed to solve the optimization problems.

References

[1] Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions, *Annals of Mathematical Statistics*, **22**, 327-351.
[2] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd ed.), Wiley, New York.
[3] 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, 3-28.
[4] Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*, Wiley, New York.
[5] Campbell, J. Y. and Shiller, R. J. (1987). Cointegration and tests of present value models, *Journal of Political Economy*, **95**, 1062-1088.
[6] Edelman, A., Arias, T. A. and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints, *SIAM Journal of Matrix Analysis and Applications*, **20** (2), 303-353.
[7] Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration, *Econometrica*, **57**, 1317-40.
[8] Geweke, J. (1996). Bayesian reduced rank regression in econometrics, *Journal of Econometrics*, **75**, 121-46.
[9] Eaton, M. L. (1983). *Multivariate Statistics - A Vector Space Approach*, New York: Wiley.
[10] Fisher, W. D. (1976). Normalization in point estimation, *Journal of Econometrics*, **4**, 243-252.
[11] James, A. T. (1954). Normal multivariate analysis and the orthogonal group, *Annals of Mathematical Statistics*, **25**, 40-74.
[12] Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.
[13] Juselius, K. (1992). Domestic and foreign effects on prices in an open economy: the case of Denmark, *Journal of Policy Modelling*, **14**, 401-428.
[14] Kadane, J. B. (1978). A comment on "Normalization in point estimation", *Journal of Econometrics*, **7**, 123-125.
[15] Kleibergen, F. and Paap, R. (2002). Priors, posteriors and bayes factors for a Bayesian analysis of cointegration. *Journal of Econometrics*, **111**, 223-249.
[16] Kleibergen, F. and van Dijk, H. K. (1994). On the shape of the likelihood/posterior in cointegration models, *Econometric Theory*, **10**, 514-51.
[17] Kleibergen, F. and van Dijk, H. K. (1998). Bayesian simultaneous equations analysis using reduced rank structures, *Econometric Theory*, **14**, 701-743.
[18] Kloek, T. and van Dijk, H. K. (1978). Bayesian estimates of equation system parameters: an application of integration by Monte Carlo, *Econometrica*, **46**, 1-19.
[19] Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
[20] Mardia, K. V. and Khatri, C. G. (1977). Uniform distribution on a Steifel manifold, *Journal of Multivariate Analysis*, **7**, 468-473.
[21] Mardia, K. V. and Jupp (2000). *Directional Statistics*, Wiley.
[22] Paap, R. and van Dijk, H. K. (2003). Bayes estimates of Markov trends in possibly cointegrated series: an application to U.S. consumption and income, *Journal of Business and Economic Statistics*, **21**, 547-563.
[23] Phillips, P. C. B. (1991). Optimal inference in cointegrated systems, *Econometrica*, **59**, 283-306.
[24] Srivastava, A. (2000). A Bayesian approach to geometric subspace estimation, *IIIE Transactions on Signal Processing*, **48** (5), 1390-1400.
[25] Strachan, R. W. (2003). Valid Bayesian estimation of the cointegrating error correction model, *Journal of Business and Economic Statistics*, **21**, 185-195.
[26] Strachan, R. W. and Inder, B. (2004). Bayesian analysis of the error correction model, forthcoming in *Journal of Econometrics*.

[27] Strachan, R. W. and van Dijk, H. K. (2004). The value of structural information in the VAR model, Econometric Institute Report EI 2003-17, Erasmus University Rotterdam.
[28] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*, **22**, 1701-1762.
[29] Villani, M. (2000). *Aspects of Bayesian Cointegration.* Ph.D. thesis, Department of statistics, Stockholm University.
[30] Villani, M. (2005). Bayesian reference analysis of cointegration, forthcoming in *Econometric Theory.*
[31] Wijsman, R. A. (1990). *Invariant Measures on Groups and Their Use in Statistics*, IMS Lecture Notes-Monograph Series vol **14**.
[32] Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.