ESAIM: PROCEEDINGS AND SURVEYS, Vol. ?, 2017, 1-10
Editors: Will be set by the publisher

SOME RECENT DEVELOPMENTS IN MARKOV CHAIN MONTE CARLO FOR COINTEGRATED TIME SERIES

Maciej Marowka[1], Gareth W. Peters[2], Nikolas Kantas[1 3] and Guillaume Bagnarosa[4]

Abstract. We consider multivariate time series that exhibit reduced rank cointegration, which means a lower dimensional linear projection of the process becomes stationary. We will review recent suitable Markov Chain Monte Carlo approaches for Bayesian inference such as the Gibbs sampler of [41] and the Geodesic Hamiltonian Monte Carlo method of [3]. Then we will propose extensions that can allow the ideas in both methods to be applied for cointegrated time series with non-Gaussian noise. We illustrate the efficiency and accuracy of these extensions using appropriate numerical experiments.

1. Introduction

The study of multivariate times series displaying the feature of reduced rank cointegration is an important topic within a spectrum of fields related to econometrics and statistics; [27], [25], [36] and [24]. The concept of cointegration was generally developed in the works of [27] and [10]. Since these early developments, there has been a wide investigation of cointegration in econometrics and finance, see [11] and more recently [16]. Fundamentally, cointegration is a property of multivariate time series whereby a lower dimensional linear transformation of a non-stationary process becomes stationary. The resulting projected time series are often referred to as "spread series" or the "cointegration portfolio".

However, in practice both the basis of projection that yields a stationary process and its dimension are unknown quantities that need to be estimated. Before presenting any models or estimation techniques a basic notion of cointegration can be described by considering a collection of time series, $y_{i,t}$ where $i = 1, \ldots, n$. If for some i, $y_{i,t}$ are not stationary but there exists real valued coefficients,

[1] Imperial College London, UK
[2] University College London, UK
[3] Imperial College London, UK
[4] ESC Rennes School of Business, France

denoted by $\boldsymbol{b}$, such that the linear combination

$$z_t = \sum_{i=1}^{n} b_i y_{t,i} \tag{1}$$

is stationary, then the series are said to be cointegrated. The linear combination vector $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)^T$ shown in (1) is called a *cointegrating relationship*; [12]. A cointegrating relationship may be seen as a long-term equilibrium phenomenon which allows the cointegrating variables, $y_{i,t}$, to deviate from their relationships in the short term, but retains their long term associations. This property is of particular relevance in financial applications as it allows to construct a mean reverting portfolio by taking positions proportional to the cointegration relationships $\boldsymbol{b}$; for a reference on algorithmic trading applications see [70], [52] and [54].

The main challenge in data analysis for such types of processes lies in the detection and estimation of the cointegration relationships. It is clear that the stationary property is invariant under linear combinations of cointegration relations, i.e. if $\boldsymbol{b}$ and $\boldsymbol{c}$ are valid relations, then so is $a_1\boldsymbol{b} + a_2\boldsymbol{c}$ for any $a_1, a_2 \in \mathbb{R}$. Therefore linearly independent relations form a basis for a subspace which is referred to as a *cointegration space.*

The dimension of the cointegration space will be referred to as a *cointegration rank*, r, and is actually the number of linearly independent vectors $\boldsymbol{b} \in \mathbb{R}^n$ whose inner product with observable series y_t yields stationary process z_t. In this paper, we are particularly interested in the estimation of the cointegration space under various modelling assumptions of the underlying processes $y_{t,i}$. When casting the inference problem for this space, certain likelihood based identification problems can arise ([39, Section 3]). We will elaborate on this issue later in Sections 2.2 and 2.3.

1.1. Contributions and Organization

This paper details the challenges in the estimation of parameters determining cointegration in a time series model. We will present some recent developments on simulation based Bayesian inference for such models and therefore the paper serves as a combination of both a survey on particular aspects of Bayesian estimation and Markov chain Monte Carlo (MCMC) methods for cointegration modelling as well as presenting several novel developments in this setting.

In terms of the Monte Carlo sampler we consider, we note that we have confined our presentation mainly to two MCMC simulation approaches: the Gibbs sampler of [41] and the GMC method of [3]. These methods originate from different research communities in econometrics and statistics, but we believe it is useful for practitioners to combine these different ideas and demonstrate how they can be incorporated for the estimation of Bayesian cointegration models. In addition to this review material, we also propose extensions by combining ideas from both methods. To the best of our knowledge this is the first application of GMC to cointegrated time series. We believe there is great potential in extending GMC for cointegration, in terms of being able to address non-linear, non-Gaussian models and using different specifications for the priors. In the numerical examples we will make comparisons related to the efficiency and accuracy of the basic methodology and proposed extensions.

The contributions developed in this manuscript can be summarised as follows:

- we bridge the gap in the MCMC literature between GMC and Bayesian cointegration models. We provide an explanation of how to define a MCMC sampler for cointegration parameters on a Stiefel manifold and in particular to define the class of Hamiltonian Monte Carlo (HMC) and Geodesic Monte Carlo (GMC) samplers in this context.;

- we develop a cointegration model extension that incorporates into the Vector Error Correction Model (VECM) the ability to work with more flexible multivariate student t errors and to develop a Bayesian model in such a context. Sampling from the resulting generalised VECM Bayesian model is then achieved in Gaussian error and the Student-t error cases via efficient samplers. In the case of the Student-t errors they exploit the scale mixture structure of a random vector with Student-t law.;
- we provide a novel algorithm which we demonstrate equally efficient as current state-of-the art samplers and furthermore can be applied in more generic scenarios for the assumed VECM driving noise random vectors distribution.;
- we consider Singular Value Decomposition (SVD) representations and priors on parameters. These are useful from a financial applications perspective to determine jointly the mean reversion and the projection basis parameters. In addition it enables us to apply random scan Gibbs sampling, which can be used to save computation effort by skipping the computation of certain geodesics for the Stiefel manifold.

Details of these contributions will be made explicit throughout the manuscript.

The organization of this paper is as follows: Section 2 introduces the error correction model (ECM) representation that is widely used in cointegration analysis and formulates the Bayesian inference problem. Then in Section 3 we present MCMC simulation techniques. In Section 4, we discuss the issue of cointegration space point estimation and in Section 5 we present various simulation studies. In Section 6 we conclude with a brief discussion.

1.2. Notation and Background Ancillary Material

This preliminary section serves the purpose of both introducing notations that we will adopt throughout the manuscript and also providing for a general statistical and econometric audience with a brief set of basic principles and background for key quantities upon which the results of this paper are based. Though mildly technical, it is useful as we focus on a particular sub-set of results that pertain directly to the relevant background in forming our model developments. We believe it will be useful to present this basic background as in the statistics literature there is still a need to link some of the concepts starting to be adopted to these general well known results that arise from topology, algebraic topology and measure theory for practitioners. With this objective in mind, we only really present a core set of key results required in this regard.

We denote a Gaussian random $(n \times T)$ matrix by $Y \sim N_{n,T}(\mu, \Sigma, \Psi)$ with row dependence in $(n \times n)$ covariance matrix Σ and column dependence in $(T \times T)$ matrix Ψ which shall be understood as the covariance between the respective rows or columns of Y. $N(\mu, \Sigma)$ will be a multivariate normal with mean vector $\mu \in \mathbb{R}^{n \times 1}$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$; I_r is an identity matrix of dimension $r \times r$.

We say that a time series y_t is integrated of order 1 and denote it as $I(1)$ if Δy_t is weakly stationary process, i.e. integrated of order 0, $I(0)$. Here we assume the weak notion of stationarity, i.e $\mathbb{E}(\Delta y_t) = g$, $Cov(\Delta y_t, \Delta y_{t-h}) = m(h)$, where $\Delta y_t = y_t - y_{t-1}$ and y_t is an $I(1)$ time series. For a time series sequence $y_1, y_2, ..., y_T$, we will also use the concise notation $y_{1:T}$.

$Vec(A)$ denotes the matrix vectorization operator which transforms a matrix A into a column vector in which columns of A are successively stacked. Furthermore, we denote the Kronecker product or tensor product between two matrices by $\otimes$ and Kronecker sum as $\oplus$. The space spanned by the columns of any $n \times r$ matrix A is denoted as $col(A)$; if A is of full column rank $r < n$, then $A_\perp$ denotes $n \times (n-r)$ matrix of full column rank satisfying $A_\perp^T A = 0$. For any square matrix, A, $||A||_F$ is a Frobenius norm, $||A||^2 = tr\{A^T A\}$, and $\rho(A)$ its spectral radius (that is, the maximal absolute value of the eigenvalues of A). The cardinality of a set B is denoted by $|B|$ and $\nabla[\cdot]$ denotes a matrix/vector of partial derivatives of an appropriately defined function.

The next two sections provide core background mathematical details assumed throughout the paper, as such the following two more technical sections in the manuscript may be skipped by practioners as they are not required for applications of our method.

We include these two sections in order to make sure all concepts are clearly and accurately defined and presented from a mathematical perspective. It also serves as a more technical reference to some concepts that are applied in later sections of the manuscript but are often not explained or detailed carefully in other MCMC literature on this topic.

1.2.1. *Set-up and Notation for Manifolds, Metrics and Geodesics.*

In general, in this manuscript we will refer to $\mathcal{M}$ as a differentiable manifold of dimension n. We may then define a Riemannian metric G which for every point on the manifold $q \in \mathcal{M}$ defines the scalar product of tangent vectors in the tangent space, denoted by $T_q\mathcal{M}$, smoothly depending on the point q. This means that in every co-ordinate system $(x^1, \dots, x^n)$ a metric $G = g_{ik}dx^i dx^j$ is defined by a matrix valued smooth function $g_{ik}(x)$ for $i, k \in \{1, \dots, n\}$ such that for any two vectors A, B the i-th component is given by,

$$[A]_i = A^i(x)\frac{\partial}{\partial x^i}, \;\; [B]_i = B^i(x)\frac{\partial}{\partial x^i}, \;\; i \in \{1, \dots, n\}\,, \tag{2}$$

which are tangent to the manifold $\mathcal{M}$ at the point q with co-ordinates $x = (x^1, \dots, x^n)$ i.e. for $A, B \in T_q\mathcal{M}$ the scalar product is equal to

$$\langle A, B\rangle_G\big|_q = G(A,B)\big|_q = \begin{pmatrix} A^1 \dots A^n \end{pmatrix} \begin{pmatrix} g_{11}(x) & \dots & g_{1n}(x) \\ \dots & \dots & \dots \\ g_{n1}(x) & \dots & g_{nn}(x) \end{pmatrix} \begin{pmatrix} B^1 \\ \cdot \\ \cdot \\ B^n \end{pmatrix} \tag{3}$$

where the metric will satisfy that

(1) $G(A, B) = G(B, A)$ a symmetricity condition;
(2) $G(A, A) > 0$ if $A \neq 0$ a positivity condition; and
(3) $G(A, B)\big|_{q=x}$ i.e. $g_{ik}(x)$ are smooth functions.

Hence, we can always consider a Riemannian metric on $\mathcal{M}$ as a family of (positive definite) inner products generically denoted by

$$G_q \colon T_q\mathcal{M} \times T_q\mathcal{M} \longrightarrow \mathbb{R}, \qquad q \in \mathcal{M} \tag{4}$$

such that, for all differentiable vector fields X,Y on $\mathcal{M}$ one has,

$$q \mapsto G_q(X(q), Y(q)), \tag{5}$$

which will define a smooth function between $\mathcal{M}$ and $\mathbb{R}$. Then we can be sure that when endowed with this metric G, we may consider the differentiable manifold $(\mathcal{M}, G)$ as a Riemannian manifold.

Furthermore, the ideas we present in this manuscript will in general be restricted to a sub-space topology in $\mathbb{R}^{n\times d}$ which will correspond to a choice of $\mathcal{M}$ given by the compact Stiefel manifold that we shall denote more specifically in this case by the notation

$$\mathbb{V}_{n,r} := \{V \in \mathbb{R}^{n\times r} : \ V^T V = I_r\} \ \ n \geq r,$$

and when $n = r$ one obtains the orthogonal group.

We will make more explicit reference to the form of G in the case of a Stiefel manifold (for cointegration model settings) in latter sections. We note that it is well known that the Stiefel manifold will become a Riemannian manifold by introducing an inner product in its tangent spaces. In this regard, we have two natural choices that can be considered for the inner products for tangent spaces of Stiefel manifolds: the Euclidean inner product and the canonical inner product. The choice adopted can affect the computational efficiency of the resulting GMC sampler we design and so we explicitly explain this detail in a latter section. Next we briefly provide a remark on defining curves.

Remark 1. *For practitioners the aforementioned material, though important for formal setup of our problem may appear a little abstracted. To aid in gaining some basic intuition for this in order to later set up of the Haussdorf measure structure, we first note the following basic properties for measuring distances on a curve embedded on a Riemannian manifold. To provide such basic intuition for aspects of these quantities defined above it will suffice to introduce a simple illustrative example to calculate relevant quantities explicitly. Consider a curve* $\gamma : x^i = x^i(t)$ *for* $i \in \{1, \dots, n\}$ *with* $a \leq t \leq b$ *on a Riemannian manifold* $(\mathcal{M}, G)$*. This may be a path one wish to follow when proposing a state dynamic in a MCMC sampler for instance. Now, at every point of the curve the velocity vetor (tangent vector) is given by* $v(t) = \left(\frac{dx^1}{dt}, \dots, \frac{dx^n}{dt}\right)$*. Then the length of the velocity vector* $V \in T_x\mathcal{M}$ *at location* x *on the manifold* $\mathcal{M}$ *is given by*

$$\sqrt{\langle v, v\rangle_G\big|_x} = \sqrt{g_{ik}\frac{dx^i(t)}{dt}\frac{dx^k(t)}{dt}\bigg|_x} \tag{6}$$

Then the length of a curve is defined by the integral of the length of velocity vector given by

$$L_\gamma = \int_a^b \sqrt{\langle v, v\rangle_G\big|_{x(t)}}dt. \tag{7}$$

Intuitively, these concepts can then be extended to define the unit of space measure change from Lebesgue measure to Hausdorff measure presented later in (12) when working with particular Riemannian manifolds, such as the case we will develop for Stiefel manifolds.

Now we can consider the formal definition of a geodesic curve on a Riemannian manifold. A geodesic is a curve $\gamma : [t_1, t_2] \mapsto \mathcal{M}$ which is intuitively considered as a trajectory of motion for a point particle without external force, travelling with constant speed. Locally, the geodesic gives the shortest path that will connect two points and it is therefore naturally related to the Riemannian metric and generally can be described by the Euler-Lagrange equations as the solution to the variational problem given by $\delta\frac{1}{2}L_\gamma = 0$ where L_γ is given by (7) such that $a = t_1$ and $b = t_2$ and the solution geodesic curve γ satisfies appropriate initial conditions, see details in [51].

Finally, we complete this section with some fundamentals that allow us to define accurately a distribution on general metric spaces that also applies in our Stiefel manifold setting. This is important, as in later sections of the paper we develop concepts of Markov chain Monte Carlo for target distributions with support on a manifold, with particular application to Stiefel topological structures that arise from the class of cointegration models we study.

1.2.2. *Distributions on Manifolds: the Hausdorff Measure on a Metric Space and its Representative Form for Riemannian Manifolds.*

Consider any metric space denoted generically in this stand alone ancillary sub-section by $(\mathbb{X}, G)$. Then for any subset of this space $S \subset \mathbb{X}$ one may define the diameter according to the metric G as follows:

$$\text{diam } S := \sup\{G(x,y)|x,y \in S\}, \quad \text{diam } \emptyset := 0. \tag{8}$$

If we then consider any subset $S \subset \mathbb{X}$ and real constants $\delta > 0$ and $d \geq 0$, we may construct all countable covers of S by sets satisfying $U_i \subset \mathbb{X}$ and diam$U_i < \delta$ that can then be used to define the metric outer measure as the infimum over all countable covers given by

$$H^d_\delta(S) = \inf\Big\{\sum_{i=1}^{\infty}(\text{diam } U_i)^d : \bigcup_{i=1}^{\infty} U_i \supseteq S, \text{ diam } U_i < \delta\Big\}. \tag{9}$$

One can then define the limiting metric outer measure according to

$$H^d(S) := \sup_{\delta > 0} H^d_\delta(S) = \lim_{\delta \to 0} H^d_\delta(S). \tag{10}$$

This particular definition of the Hausdorff measures is just a special case of a more general construction due to Caratheodory. In this context, if we consider measurable sets in a Caratheodory sense, such that they satisfy Caratheodory's criterion for the Lebesgue outer measure on $\mathbb{R}^n$ denoted by λ, see further details in [13]. That is, we consider sets $E \subseteq \mathbb{R}^n$ satisfying that for Lebesgue outer measure, the sets E will be Lebesgue measurable if and only if

$$\lambda(A) = \lambda(A \cap E) + \lambda(A \cap E^{\mathrm{c}}), \tag{11}$$

for every $A \subseteq \mathbb{R}^n$, where A is not required to be a measurable set itself. Then it will make sense to consider the restriction of this metric outer measure to the σ-field of Caratheodory-measurable sets and the result will be a well defined measure which is generically referred to as the d-dimensional Hausdorff measure of S. Consequently, as a result, the properties of the metric outer measure in this context will indeed ensure that all Borel subsets of $\mathbb{X}$ are H^d measurable. It is important to realise that in this set-up the definition of covering sets is somewhat arbitrary. With this in mind one could of course describe different general forms of Haussdorf measures by considering different restrictions on the class of admissible coverings, see details in [47] and [58]. For instance one can use coverings by balls in such case the resulting outer measure is often referred to as the spherical Hausdorff measure or by cylinders and one obtains the cylindrical Hausdorff measure. In this manuscript we will consider the case preferred in [3] corresponding to open sets of $\mathbb{X}$.

Although brief, this basic background suffices for our applications in this manuscript and we may now consider the crux of the formulation relating to developing a probability measure on a manifold with a well defined metric. In particular we may now refer to a family of probability measures that can be defined on the Riemannian manifold via what is known as the Haussdorff outer measure (henceforth measure) see [26]. For the remainder of the manuscript, unless otherwise specified all densities are w.r.t the Lebesgue measure $\mathcal{L}(\cdot)$ on the appropriate space and $\mathcal{H}(\cdot)$ denotes Hausdorff measure on an appropriate manifold, as detailed above. Now, if we consider a Riemannian manifold, the Hausdorff measure is related to the Lebesque measure according to the expression:

$$\mathcal{H}(dM) = \sqrt{G(M)}\mathcal{L}(dM), \tag{12}$$

which is a Lebesque measure scaled by the volume element $\sqrt{G(M)}$, where G is a Riemannian metric on the manifold.

2. Bayesian Inference for Cointegrated Time Series

A range of different parametrizations and model constructions have previously been proposed in the literature on cointegration time series modelling, these include: the Triangular form (see [56]); the Common Trends representations (see [59]); and the class of Vector Error Correction Models (VECM) models.

In this manuscript we will focus on the class of VECM model structures. Throughout all this section, we assume that the cointegration rank r is fixed and known. In fact, it is a crucial parameter and needs to be estimated. A brief review on the rank estimation techniques is presented subsequently in Section 4.3.

2.1. **Vector Error Correction Models for Cointegration**

In this section we present one of the most widely utilised characterization of multivariate time series models for cointegration settings. This is based on cointegration time series models that are based on Vector Autoregressive (VAR) structures. Such models have been widely studied in the Econometrics literature, see [10] and [65]. A popular representation is the Error Correction Model (ECM), see [20], [63] and the overview of [39]. Due to the substantial popularity of this particular form of cointegration model specification, the remainder of manuscript will focus on this framework in a Bayesian model estimation context. We will first present the structural form of the ECM form of cointegration.

An ECM is written based on a VAR process of order p, as follows:

$$\Delta y_t = \Pi y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \phi D_t + \epsilon_t$$

where $y_t \in \mathbb{R}^n$ denote observed returns, $\Pi \in \mathbb{R}^{n\times n}$ is the long-run multiplier matrix, $\Gamma_i \in \mathbb{R}^{n\times n}$ the i-th lag matrix and $D_t \in \mathbb{R}^n$ is an exogenous covariate for the observation y_t. The appeal of the ECM formulation is that it combines flexibility in dynamic specification with desirable long-run properties; see [7] for a discussion.

The cointegration properties of the ECM depend on the rank r of the long-run multiplier matrix Π. If $r = 0$, then the ECM does not exhibit any cointegration relationship and it can be estimated as a stationary process in first differences. If $r = n$, that is, if the matrix Π is of full rank, then the VAR model itself is stationary and can therefore be estimated via standard stationary process techniques in multivariate settings. If, however, the rank r is intermediate, $0 < r < n$, then the ECM process exhibits cointegration.

In the context of cointegration, we can write the matrix Π as the product $\Pi = \alpha\beta^T$ where both $\alpha,\beta \in \mathbb{R}^{n\times r}$ are full rank matrices. The r columns of the matrix β are the cointegrating vectors of the process. In addition, $\beta^T y_t$ reflects common trends while α contains their loading factors. For simplicity, we will consider a simplified model, where y_t is marginally integrated of order 1, $I(1)$, with r linear cointegration relationships and r is assumed to be known. The observation equation is given by:

$$\Delta y_t = \alpha\beta^T y_{t-1} + \epsilon_t \tag{13}$$

where $t = 1, ..., T$ and ϵ_t form an i.i.d. noise sequence. The most commonly used case for the noise distribution is $\epsilon_t \sim N(0, R)$, however in this paper we will also consider multivariate Student-t distributed noise examples where $\epsilon_t \sim t_\omega(0, R)$. In what follows more types of noises can be treated similarly, as long as the corresponding densities can be evaluated point-wise up to a proportionality constant and the log densities are differentiable.

Remark 2. *It is important to note that the decomposition of the long-run multiplier matrix* Π *into* $\alpha\beta^T$ *(and hence the cointegration relations) are not unique. In fact, for every non-singular matrix* $Q \in \mathbb{R}^{r\times r}$*, we can define* $\alpha^\star = \alpha Q^T$ *and* $\beta^\star = \beta Q^{-1}$ *and get* $\Pi = \alpha^\star \beta^{\star T}$.

If cointegration exists, the ECM representation will generate better forecasts than the corresponding representation in first-differenced form, particularly over medium and long-run horizons. Indeed, under the cointegration property, z_t in (1) will maintain a finite forecast error variance, whereas other linear combinations of the forecasts of the individual series in y_t could have increasing variance, see [11] for examples.

In the remainder of the paper, we focus on the ECM characterization in the simple form of (13). This model will be sufficient to demonstrate estimation properties related to the cointegration rank and basis. For simplicity we do not include autoregressive lags, however we note that analogously to [71], they can be incorporated in the simulation methodology.

2.1.1. *Matrix and Vectorized Representations*

Different matrix or vector representations for the time series model in (13) will be used throughout. These expressions will prove particularly useful (later in Section 3) for performing likelihood evaluations, and manipulating posterior densities and their gradients. One possibility is to write (13) compactly as:

$$Y = \alpha\beta^T Z + E, \tag{14}$$

with $Y = [\Delta y_1, \Delta y_2, ..., \Delta y_T]$, $Z = [y_0, y_1, ..., y_{T-1}]$, $E = [\epsilon_1, ..., \epsilon_T]$. In addition, one can also use vectorization operations and treat Y and $Vec(Y)$ as equivalent random variables and choose the form that is most convenient for implementation. (14) can be expressed as:

$$\begin{aligned} Vec(Y) &= (Z^T\beta \otimes I)Vec(\alpha) + \widetilde{E} \\ &= (Z^T \otimes \alpha)Vec(\beta^T) + \widetilde{E}. \end{aligned}$$

For the Gaussian case, $Y \sim N_{n,T}(\mu, \Sigma, \Psi)$ implies $Vec(Y) \sim N(Vec(\mu), \Sigma \otimes \Psi)$ ([21, Theorem 2.2.1]), so then $\widetilde{E} \sim N(\mathbf{0}, I_T \otimes R)$. Note that this does not hold for the Student-t case as diagonal covariance structure (i.e. no correlation) does not imply cross-sectional independence. Henceforth, we need to resort to the computation of likelihood as a product of independent terms. We will return to this point in Section 2.4.2.

2.1.2. *Cointegration Spread Series*

We may now define the so called "Spread Series" as $z_t = \beta^T y_t$. The dynamics of this process can be trivially derived from the standard ECM representation:

$$z_t = (I + \beta^T\alpha)z_{t-1} + \beta^T\epsilon_t. \tag{15}$$

Therefore, the spread process under the ECM representation (13) is a $r-$dimensional $VAR(1)$ process. Note that the necessary spectral radius condition for the stability of the ECM can be written as $|\rho(I + \beta^T\alpha)| < 1$.

2.2. Identification Considerations

In the Bayesian modelling context of cointegration models there have been a number of papers developing frameworks to deal with the lack of identification in the likelihood and studying its influence in the posterior with different classes of priors on α and β. We refer the reader to [39] for a thorough review.

For a fixed record of observations, it is clear that we can find a pair of parameters $(\alpha, \beta) \neq (\alpha', \beta')$, such that $p(y_1, \dots, y_t|\alpha, \beta) = p(y_1, \dots, y_t|\alpha', \beta')$. This is often referred to as non-identifiability in the literature of point estimation (see [44, pages 24, 57]). This is a general issue that appears often in point estimation or Maximum Likelihood methods and is not unique to cointegration models. Since cointegrating vectors are not unique (see Remark 2), identifying restrictions must be imposed to allow their estimation. This can be achieved either by imposing normalizations on particular coefficients or by using an eigenvalue-eigenvector method of identification first developed by [1] for the reduced rank regression model and then used by [33, 34] for cointegrating ECM.

One standard approach to globally overcome the identification issue illustrated in Remark 2 is to impose constraints in the form of linear normalizations as mentioned earlier: for instance via a non-unique identification constraint of r^2 restrictions as follows $\beta = [I_r, \beta^*]^T$ where $\beta^* \in \mathbb{R}^{r\times(n-r)}$.

This method has been successfully applied in works such as [66], or more recently [52] and [54]. The implementation of linear restrictions is based on the prior knowledge of which r rows of β will be linearly independent. More generally, one can partition $\beta = (\beta_1^T, \beta_2^T)^T$, where $\beta_1 \in \mathbb{R}^{r\times r}$ and impose the normalization by choosing a matrix Q such that $Q\beta$ is invertible; then use $\beta(Q\beta)^{-1}$ instead of β.

A challenge with this approach is that inappropriately specified Q may lead to $Q\beta$ being singular ([39, Section 3]). Furthermore, it might also restrict the number of important models to be considered in the cointegration analysis. A second issue that may occur with such approaches is that at the regime where α is close to 0, β does not enter the model ([38]). This results in a local non-identification and consequently an improper posterior under a diffuse prior for β in the Bayesian setting.

Choosing priors for cointegration is widely studied topic; see [39, Section 4] for a review. In this paper we we will follow a popular approach to dealing with identification issues by imposing priors directly on the cointegration space, which assumes a distribution on a Grassman manifold. This is discussed in [57] and [62], both of which are related to earlier work of [45].

2.3. Background on Bayesian Approaches on Manifolds in Cointegration Contexts

In [60], a Bayesian inference procedure is presented which allows for unconditional inference on the structural features of VAR process. The novelty of this paper was the development of a probability measure on a Grassman manifold to elicit uniform priors on subspaces defined by particular structural features of VARs. [63] developed similar uninformative and informative priors for the cointegrating space, which allows one to develop sampling schemes such as MCMC or approximations schemes such as Laplace approximations to perform numerical integration with respect to such cointegration Bayesian models when undertaking estimation.

Furthermore, these authors provide careful elicitation of the prior distribution on the model coefficients from a prior on the cointegrating space. They also provide an outline of the identification restrictions which will then naturally arise or be implied by their specification of model structure. In [71], a Bayesian reference prior is presented with the property of distributing its probability mass uniformly over all cointegration spaces for a given rank and which is invariant to the choice of normalizing variables for the cointegration vectors. Several methods for computing the posterior distribution of the model parameters conditional on the cointegration rank, r, are proposed, whereby all inferences are determined from approximate samples of the posterior.

Following these works, there has been a shift of paradigm from direct prior specifications on the parameter space to the priors on the space spanned by the cointegrating vectors ([42, 43, 71]) or [61] for some recent extensions. This approach has benefited from efficient posterior simulation methods, such as the Gibbs sampler [41], which will be presented in more detail in Section 3.1.

Given the recent interest in this re-specification of the class of Bayesian cointegration models, we will focus the remainder of the paper on these new classes of model specification. That is, we will follow this approach of modelling priors on the cointegrating space under the ECM. The cointegration space (or equivalently the span of β, $col(\beta)$) is a r-dimensional hyperplane in a n-dimensional space. The identification restriction $\beta^T\beta = I_r$ can be used, where r is the cointegration rank. This restricts β to belong to the following Stiefel manifold:

$$\mathbb{V}_{n,r} := \{V \in \mathbb{R}^{n\times r} :\ V^T V = I_r\},$$

which is compact, so the uniform distribution is a proper prior.

Remark 3. *In the linear normalization case with $\beta = [I_r, \beta^*]$, if β^* follows matrix variate $t-$distribution, then $col(\beta)$ has uniform distribution on the Grassman manifold. A drawback of this prior choice is that the second and higher moments of the posterior distribution do not exist for $r > 1$; see [39, Section 5.1] for details.*

2.4. **Prior Model Consideration**

In this manuscript we are particular interested in the cointegration parameters (α, β), but we note that of-course the ECM contains additional parameters, θ, such as the parameters related to ϵ_t. We will use a Gibbs sampling approach alternating simulation of $p(\alpha, \beta|y_{1:T}, \theta)$ and $p(\theta|y_{1:T}, \alpha, \beta)$.

When direct simulation is possible this will lead to a standard Gibbs sampler, but when this is not possible one could use instead a small number of iterations from MCMC sampler kernels invariant to $p(\alpha, \beta|y_{1:T}, \theta)$ and $p(\theta|y_{1:T}, \alpha, \beta)$ respectively. We will present below independent priors between (α, β) and θ that will be used later for comparisons between different MCMC algorithms.

2.4.1. *Priors for α, β*

We will choose the prior distribution for $\beta \in \mathbb{V}_{n,r}$ to be the matrix angular central Gaussian distribution defined over the Stiefel manifold with respect to the Hausdorff measure:

$$dp(\beta) \propto |P_\tau|^{-r/2}|\beta^T(P_\tau)^{-1}\beta|^{-n/2}d\mathcal{H}(\beta) \tag{16}$$

where $P_\tau = \mathtt{H}\mathtt{H}^T + \tau\mathtt{H}_\perp\mathtt{H}_\perp^T,\ \tau \in [0, 1]$, with $\mathtt{H} \in \mathbb{V}_{n,r}$ acts as a hyper parameter on the cointegration space. In this class of priors, P_τ determines the central location of the distribution on $col(\beta)$, which in this case is $col(H)$ and τ the amount of the dispersion around the central location. If $\tau = 1$, then $P_\tau = I_n$ and (16) defines uniform prior on the manifold. In turn, the value of hyperparameter $\tau = 0$, expresses prior assigning the cointegration space to be $col(\mathtt{H})$. Note that one can introduce a further hierarchical structure using hyperparameters τ and hyper-priors with support restricted to $[0, 1]$.

For α we will use the shrinkage prior used in [71]:

$$\alpha|\beta \sim N_{n\times r}(0, (\nu\beta^T P_{1/\tau}\beta)^{-1}, G), \quad (17)$$

so that $Vec(\alpha)|\beta \sim N(0, \Sigma_\alpha)$ where $\Sigma_\alpha = (\nu\beta^T P_{1/\tau}\beta)^{-1} \otimes G$, where G is a symmetric positive definite matrix that is often chosen to be the noise covariance matrix R. The shrinkage parameter ν can be fixed or random by defining hierarchical priors.

Following, the Bayes rule, the joint posterior density of α, β is given by

$$dp(\alpha, \beta|y_{1:T}, \theta) = \pi(\alpha, \beta)\mathcal{L}(d\alpha)\mathcal{H}(d\beta). \quad (18)$$

with the density being

$$\pi(\alpha, \beta) \propto p(y_{1:T}|\alpha, \beta, \theta)p(\beta)p(\alpha|\beta).$$

Sampling from $p(\alpha, \beta|y_{1:T}, \theta)$ is not possible directly, so in Section 3 we present MCMC methods appropriate for this task.

2.4.2. *Priors and Conditional Posterior Distributions for the Parameters of ϵ_t*

As presented in the ECM framework, the time series is driven by a i.i.d. white-noise sequence of random vectors. In this paper we consider a two classes of driving error stochastic noises: multivariate Gaussian and multivariate Student-t.

In the Gaussian case, $\theta = R$ and standard conjugacy results motivate using an Inverse Wishart distribution for the prior of the covariance matrix R, see details in [52] and [54]. For simplicity in this paper we will use $R \sim \mathcal{W}^{-1}(n+2, I)$ for the prior. As a result for the full conditional we have

$$p(R|y_{1:T}, \alpha, \beta) = \mathcal{W}^{-1}((n+2) + T, \left(Y - \alpha\beta^T Z\right)\left(Y - \alpha\beta^T Z\right)^T + I_n).$$

In the Student-t case $\theta = \{R, \omega\}$, since we will express the driving error random vector according to a well known scale mixture form (see examples in [73] and [2]), which lends itself naturally to a sampling scheme based on a standard data augmentation approach. Let $\epsilon_t = \lambda_t\varepsilon_t$ with $\varepsilon_t \sim N(0, R)$ and $\lambda_t \sim IG(\frac{\omega}{2}, \frac{\omega}{2})$ being i.i.d. In this case, one can show that integrating out λ will give $\epsilon_t \sim t_\omega(0, R)$.

We can consider then instead the parameterization $\theta = \{R, \omega, \lambda_{1:T}\}$ and set priors for R, ω and then derive a full conditionals for each $R, \omega, \lambda_{1:T}$ that should be used in sequence when sampling. Inference for ω whilst possible is challenging and needs careful choice for priors, see [14, 15, 18] for details. Typically, improper priors are used for ω, but for simplicity we will consider here the case where it is fixed and known.

The methodology presented below can be extended using the priors in [14, 15, 18], but due to the full conditional being intractable it needs to be replaced with an appropriate MCMC procedure. For R we will use the same prior as in the Gaussian case, so for the full conditionals we get

$$p(R|y_{1:T}, \alpha, \beta, \lambda_{1:T}, \omega) = \mathcal{W}^{-1}((n+2) + T, \left(Y - \alpha\beta^T Z\right)\left(diag\left(\lambda_{1:T}\right) \otimes I_n\right)^{-1}\left(Y - \alpha\beta^T Z\right)^T + I_n),$$

and

$$p(\lambda_t|y_{1:T}, \alpha, \beta, R, \omega) = IG(\frac{\omega + n}{2}, \frac{\omega + \left(y_t - \left(I + \alpha\beta^T\right)y_{t-1}\right)R^{-1}\left(y_t - \left(I + \alpha\beta^T\right)y_{t-1}\right)}{2}).$$

3. MCMC Approaches to Bayesian Cointegration

In this section we restrict our attention to parameters α, β and present different MCMC approaches for simulating from $p(\alpha, \beta | y_{1:T}, \theta)$ or other equivalent conditional posterior distributions under different parameterizations. Also we will often drop for simplicity the conditioning on θ in the notation.

Below we present the Gibbs sampler proposed in [41], which can be considered the "state of the art" method for this problem in the Gaussian case. Then we extend the Geodesic Monte Carlo (GMC) sampler of [3] to the ECM Bayesian cointegration model framework. This allows us to significantly generalise the class of models for which we may develop efficient MCMC samplers for that extend well beyond the sampler restrictions to the Gaussian case of [41], whilst maintaining the sampler performance of this state of the art method.

We emphasize that these are not the only options available. The problem of sampling from matrices with rank restrictions, has long attracted interest due to its relevance in principal components analysis and other settings, see discussion in [67]. Recent developments that are relevant to sampling from matrices belonging to a Stiefel manifold are [6, 29], where appropriate transformations are combined with column-wise updates in a Gibbs framework. These methods can perform well (see also [3, Section 5]), but in the interest of brevity we will not present them here nor include them in our numerical examples.

Other reasons behind this omission are that the efficiency of the Gibbs sampler in [41] makes the use of columnwise updates in β less desirable in practice and GMC is more generic so has potential to be applied in wider variety of settings.

3.1. Gibbs Sampling for Bayesian ECM Posterior Distributions

From a Gibbs sampling perspective the conjugacy of Gaussian distributions is attractive, but it is hard to derive conditionals for β and α (the semi-orthogonality restriction implies that the conditional posterior of β is non-standard).

In [41] the authors exploit instead the polar decomposition ([5, p. 19]), $\alpha = \mathcal{A}\kappa^{\frac{1}{2}}$, with $\kappa = \alpha^T\alpha$ and $\mathcal{A} = \alpha\kappa^{-\frac{1}{2}}$ being the rotational component ($\mathcal{A} \in \mathbb{V}_{n,r}$). Similarly β can be viewed as the rotational part of a matrix $\mathcal{B} = \beta\kappa^{\frac{1}{2}}$, so that $\kappa = \alpha^T\alpha = \mathcal{B}^T\mathcal{B}$ and $\beta = \mathcal{B}(\mathcal{B}^T\mathcal{B})^{-1/2}$. As a result we end up with various possible parameterizations for Π

$$\Pi = \alpha\beta^T = \mathcal{A}\mathcal{B}^T = \alpha\kappa^{-\frac{1}{2}}\mathcal{B}^T, \tag{19}$$

where it is useful to notice that $\mathcal{B}$ is unrestricted. The following proposition establishes equivalent priors for $(\mathcal{A}, \mathcal{B})$.

Lemma 1. *Given the hierarchical prior on* (α, β) *as in* (16)-(17). *Then the prior for* $\mathcal{A}$ *and* $\mathcal{B}$ *is given by:*

$$dp(\mathcal{A}) \propto |G|^{-r/2}|\mathcal{A}^T G^{-1}\mathcal{A}|^{-n/2} d\mathcal{H}(\mathcal{A}), \tag{20}$$
$$Vec(\mathcal{B})|\mathcal{A} \sim N(0, \Sigma_{\mathcal{B}}) \tag{21}$$

where $\Sigma_{\mathcal{B}} = (\mathcal{A}^T G^{-1}\mathcal{A})^{-1} \otimes \nu P_\tau$.

The proof can be found in technical appendix of [41].

In Algorithm 1 we present the Gibbs sampler developed in [41] for the Gaussian case when $\epsilon_t \sim N(0, R)$. To simulate a MCMC transition leaving $p(\alpha, \kappa, \mathcal{B}|y_{1:T}, R)$ invariant, they alternate sampling between distributions p_α and $p_\mathcal{B}$ defined below in (22)-(23) and update the value for κ deterministically.

In [41] the authors justify the method as a partially collapsed Gibbs sampler, which could explain the resulting efficiency. Given the Lemma 1 and the choice of model Gaussian conjugate priors (16)-(17),(20)-(21), the conditional posteriors of $Vec(\alpha)$ and $Vec(\mathcal{B}^T)$ are distributed as vector variate normals, $p_\alpha,\ p_\beta$ respectively:

$$p_\alpha(\cdot|\beta, y_{1:T}) \sim N(\mu^\alpha_{post}, \Sigma^\alpha_{post}), \tag{22}$$

$$p_\mathcal{B}(\cdot|\tilde{\mathcal{A}}, \mathcal{D}) \sim N(\mu^\mathcal{B}_{post}, \Sigma^\mathcal{B}_{post}). \tag{23}$$

The respective means and variances of multivariate normals are:

$$\Sigma^\alpha_{post} = (M_\alpha^T \tilde{V}^{-1} M_\alpha + (\Sigma_\alpha)^{-1})^{-1}, \quad \mu^\alpha_{post} = \Sigma^\alpha_{post}(M_\alpha^T \tilde{V}^{-1}\tilde{y}) \tag{24}$$

and

$$\Sigma^\mathcal{B}_{post} = (M_\mathcal{B}^T \tilde{V}^{-1} M_\mathcal{B} + (\Sigma_\mathcal{B})^{-1})^{-1}, \quad \mu^\mathcal{B}_{post} = \Sigma^\mathcal{B}_{post}(M_\mathcal{B}^T \tilde{V}^{-1}\tilde{y}) \tag{25}$$

where $\tilde{V} = \oplus_{t=1}^T R$, $\tilde{y} = Vec(Y)$, $M_\alpha = (Z^T\beta \otimes I)$, $M_\mathcal{B} = (Z^T \otimes \alpha)$. All the technical derivations and based on the simple multivariate Gaussian conjugate properties and can be found in [41]. The algorithm describing this sampler is provided below.

Algorithm 1 One iteration of the Gibbs sampler of [41] for cointegration ECM in (13).

Starting from (α, β):

(1) Sample $Vec(\tilde{\alpha})$ from $p_\alpha(\cdot|\beta, y_{1:T})$ and transform to $\tilde{\mathcal{A}} = \tilde{\alpha}(\tilde{\alpha}^T\tilde{\alpha})^{-1/2}$.
(2) Sample $Vec(\tilde{\mathcal{B}^T})$ from $p_\mathcal{B}(\cdot|\tilde{\mathcal{A}}, y_{1:T})$ and transform $\beta^* = \tilde{\mathcal{B}}(\tilde{\mathcal{B}}^T\tilde{\mathcal{B}})^{-1/2}$ and then $\alpha^* = \tilde{\mathcal{A}}(\tilde{\mathcal{B}}^T\tilde{\mathcal{B}})^{1/2}$.
(3) Return (α^*, β^*).

Remark 4. *Whilst the Gibbs sampler is intended for $\epsilon_t \sim N(0, R)$, obtaining a Gibbs sampler for the Student-t case is trivial. One can use Algorithm 1 and then alternate between the full conditionals in Section 2.4.2.*

3.2. Geodesic and Hamiltonian Monte Carlo

In this section we will begin by explaining how to utilise the formulation of Hamiltonian equations of motion to develop a proposal kernel for a Markov chain sampler that will produce a sampler known in the literature as Hamiltonian Monte Carlo (HMC). We will first explain the development of this sampler in the familiar case of Euclidean space, then we will consider the generalization of this framework to a manifold in a Riemannian metric space, for which we may define and evaluate geodesics to define movements on the manifold and quantify the distance over such geodesics via a well defined Riemannian metric that will allows us to work with distributions on such spaces via the specification of the Hausdorff measure.

3.2.1. *Basics of Hamiltonian Monte Carlo (HMC) on Euclidean Space*

Hamiltonian dynamics were originally introduced in molecular simulation and later used within a MCMC framework in [8] leading to the so-called *Hybrid or Hamiltonian Monte Carlo*, which was popularized in Statistics in [48] and [49]. Here we present a very short summary of the method. Following the standard HMC convention, let q be the variable we wish to infer (in our case it is (α, β)). HMC is a a MCMC method to sample from a posterior distribution with density $\pi(q)$ by introducing an auxiliary variable, p called momentum variable, and then targeting the joint distribution:

$$\begin{aligned} \pi(q,p) &\propto \pi(q) \exp\left(-\frac{1}{2} p^T \mathsf{P}(q)^{-1} p\right) \\ &\propto \exp\left(-H(q,p)\right), \end{aligned} \tag{26}$$

where $\mathsf{P}(q)$ is a positive definite matrix. The log of this joint posterior is interpreted as a the Hamiltonian function, that is a sum of a potential and kinetic energy function given by $-\log \pi(q)$ and $\frac{1}{2} p^T \mathsf{P}(q)^{-1} p$ respectively. The advantage of this interpretation is that one can design proposals within a MCMC algorithm using artificial Hamiltonian dynamics with respect to a fictitious time τ:

$$\dot{q} = \nabla_p H, \quad \dot{p} = -\nabla_q H. \tag{27}$$

When this system of motion equations can be solved exactly (27) the resulting solution will leave H invariant. Furthermore, the solution will possess interesting properties such as volume preservation and time reversibility. These two particular properties make this dynamic framework particularly relevant for MCMC sampling methods as these properties will guarantee the resulting constructed Markov chain will satisfy detailed balance when used within MCMC, see [49] for details.

The problem is that typically one cannot solve the system of equations for the Hamiltonian dynamics in (27) exactly. Consequently, one typically resorts to the use a numerical integration method to find a solution. Consquently, H will no longer be invariant. However, in the context of MCMC sampler proposal design, this can be overcome through the use of a Metropolis accept-reject correction step.

Let (q', p') be the numerical solution of equation (27) after some chosen time and starting from (q, p), then this numerical solution will be accepted as the new proposed movement for the Markov chain state via a Metropolis-Hastings Accept-Reject probability with acceptance probability

$$\min(1, \exp(-H(q', p') + H(q, p)), \tag{28}$$

otherwise the proposed numerical solution movement according to the Hamiltonian motion is rejected and the Markov chain remains at state (q, p).

It is important to note that in order for this MCMC scheme to preserve detailed balance the numerical integration method needs to be time reversible and volume preserving (or *symplectic*). Different choices of integration method may be considered to achieve this objective such as: the Newmark-beta method [50], Stormers method [22], Verlet's method [69], the Velocity Verlet method or the closely related leapfrog integration framework, see discussions in surveys such as [64].

The most popular of these in MCMC applications involves the class of leapfrog integration methods. A leapfrog integration method is a numerical approach to integrating differential equations of the form $\ddot{x} = \frac{d^2 x}{dt^2} = f(x)$ such as the system of Hamiltonian dynamics described in (27).

The approach of leapfrog integration takes its name from the fact that the method interleaves or alternates between two solution steps: an updating of positions $x(t)$ and then an updating of velocities $\dot{x}(t)$ at interleaved time points. These points of update are set-up in such a manner that the solution systematically "leapfrogs" over each previous solution in pursuit of the next solution point.

Note, in particular, when P is a constant matrix (i.e. not a function of q) it is common in MCMC literature to utilise the leapfrog method and then optimize performance by tuning the choice of P, the number of steps and step size of the leapfrog integration. If however, we consider the more general context in which P is no-longer a constant and in particular we consider $\mathsf{P}(q)$ to vary with q we may construct a more general framework for MCMC sampler design in this class of methods, which becomes in some sense locally adaptive. That is, taking into account local behaviour of π brings more flexibility in terms of tuning and is advantageous from a performance point of view; see [19] for a review.

In this case numerical integration requires using splitting techniques ([23]) that treat the potential and kinetic parts of H as separate Hamiltonians. Let $H^1 = -\log \pi(q)$ and $H^2 = \frac{1}{2} p^T \mathsf{P}(q)^{-1} p$. The Hamiltonian equations for H^1 are

$$\dot{q} = 0, \quad \dot{p} = -\nabla_q H^1 \tag{29}$$

and can be solved exactly:

$$q_\tau = q_0, \quad p_t = p_0 + \tau \, \nabla_q \log \pi(q)\big|_{q=q(0)} .$$

Then if there is a symplectic integrator or numerical solver for

$$\dot{q} = \nabla_p H^2, \quad \dot{p} = -\nabla_q H^2 \tag{30}$$

then one could use the two integrators together in a time-symmetric manner to generate MCMC proposals. Starting from (q_τ, p_τ) a single iteration of the composed integrator could be performed as follows:

- Compute $q_{\tau+\frac{\epsilon}{2}} = q_\tau, \quad p_{\tau+\frac{\epsilon}{2}} = p_\tau + \frac{\epsilon}{2} \, \nabla_q \log \pi(q)\big|_{q=q_\tau}$,
- Solve (30) for a time interval equal to ϵ starting with initial condition being $(q_{\tau+\frac{\epsilon}{2}}, p_{\tau+\frac{\epsilon}{2}})$ to get $\left(q_{\tau+\frac{3}{2}\epsilon}, p_{\tau+\frac{3}{2}\epsilon}\right)$,
- Compute $q_{\tau+2\epsilon} = q_{\tau+\frac{3}{2}\epsilon}, \quad p_{\tau+2\epsilon} = p_{\tau+\frac{3}{2}\epsilon} + \frac{\epsilon}{2} \, \nabla_q \log \pi(q)\big|_{q=q_{\tau+\frac{3}{2}\epsilon}}$.

This could be iterated L times and then one would need to apply a Metropolis accept-reject step to $(q_{\tau+2L\epsilon}, p_{\tau+2L\epsilon})$ compared to (q_τ, p_τ) as before with a similar acceptance ratio

$$\min(1, \exp(-H(q_{\tau+2L\epsilon}, p_{\tau+2L\epsilon}, q_{\tau+2L\epsilon}, p_{\tau+2L\epsilon}) + H(q_\tau, p_\tau)).$$

Remark 5. *The presentation so far does not make any reference to q being a point on a manifold except when defining π w.r.t the Hausdorff measure. [19] provide a detailed review for this case when $\mathsf{P}(q)^{-1}$ is a Riemannian metric tensor and discuss on how the Hamiltonian equations in* (30) *should be solved in detail. In this case, H^2 defines a metric and one can use co-geodesic flows (q_τ, p_τ) with $\dot{q} = \mathsf{P}(q)^{-1}p$ that follow a trajectory that leaves H^2 constant.*

Remark 6. *If $q = (\alpha, \mathcal{B})$ the variable of interest is not defined on a manifold, a few iterations of the above HMC procedure can be applied within Algorithm 1, each time for π being $p_\alpha(\cdot|\beta, y_{1:T})$*

and $p_{\mathcal{B}}(\cdot|\mathcal{A}, y_{1:T})$. This can be useful when direct Gibbs Sampling is not possible, e.g. when ϵ_t is not Gaussian.

3.2.2. *Geodesic Monte Carlo: Extending HMC to Connected Riemannian Manifolds*

In this section we work with the fact that a connected Riemannian manifold carries the structure of a metric space whose distance function is the arc length of a minimizing geodesic. For brevity we note that the arc length is defined for a simple illustration in the preliminaries section in Remark 7 and the geodesic path generally in Section 1.2.1. In particular, we will now concentrate on a class of Monte Carlo sampler, aptly named for the fact that it exploits the aforementioned property of traversing geodesics, and is known as GMC (see [3]). This is nothing more than a Monte Carlo sampler in which the path follows a geodesic flow, such as those defined in earlier optimization contexts in [51] and the references contained therein.

In the context of Monte Carlo methods the GMC framework extends earlier works by authors such as [51] and [19] and HMC for simulation from a distribution defined on a manifold. The idea of GMC is to develop a sampler for distributions defined on a manifold, that will traverse the mass of the distribution via paths through the support of the distribution on the manifold defined via shortest distance paths known as geodesics. In this way, the Markov chain should traverse efficiently by following geodesic paths on the surface of the manifold around the target mass of the distribution on the manifold. The intended purpose of such a construction is to efficiently explore the support of the target distribution defined on the manifold via a Markov chain constructed to move along geodesics on the manifold.

More concretely, let $\mathcal{M}$ be a manifold $q \in \mathcal{M}$ and suppose π is the density of interest w.r.t the to the Hausdorff measure. One can define the Hamiltonian as before $H(q,p) = -\log \pi(q) + \frac{1}{2} p^T \mathsf{P}(q)^{-1} p$, and then one needs to choose P and design Hamiltonian flows across $\mathcal{M}$ through its tangent space, as defined in the preliminary notations in Section 1.2. To achieve this [3] propose working with special cases that can be formed by embeddings of $\mathcal{M}$ in a Euclidean space.

In this case, they assume that one can obtain a bijective map $\xi : \mathcal{M} \to \mathbb{R}^n$ that maps every open set of $\mathcal{M}$ to an open set in $\mathbb{R}^n$ and let $x = \xi(q)$. Note these covering sets need not be open sets, but it suffices to restrict to this case.

The embedding proves very useful in characterizing the tangent space at a point q, namely T_q and choosing a convenient P, see the example and construct of such a tangent as explained with an illustration in Section 1.2.

Now, let $M_{q'} = \nabla_q \xi|_{q=q'}$, i.e. $M\,[ij] = \frac{\partial [\xi]_i}{\partial q_j}$, then $T_{q'}$ can be viewed as the column span of $M_{q'}$. In addition, $\mathscr{P}_q = M_q (M_q^T M_q)^{-1} M_q^T$ defines a projection on T_q.

[3] set $\mathsf{P}(q) = M_q^T M_q$ and then define a re-parameterisation of (q,p) to (x,v) where $x = \xi\,(q)$ and $v = \dot{x} = M_q \dot{q} = M_q (M_q^T M_q)^{-1} p$ given $\dot{q} = \mathsf{P}(q)^{-1} p$. The Hamiltonian (31) can be restated according to

$$H(x,v) = H^1(x) + H^2(v) = -\log \pi(x) + \frac{1}{2} v^T v, \tag{31}$$

and one can write Hamiltonian dynamics equations for H^1 and H^2 as before. Noting that $\nabla_x = M_q^T \nabla_q$, the solution flow in the embedded phase space corresponding to H^1 is given by

$$v_\tau = v_0 + \tau \mathscr{P}_q \, \nabla_q \log \pi(q)|_{\,q=q_0}. \tag{32}$$

As a result the evolution of the velocity term v_t can be described only using $q_t \in \mathcal{M}$. This is an important observation as we have effectively re-parameterised again the problem and work with $(q, v = \mathscr{P}_q p)$.

The solution of the Hamiltonian equations corresponding to H^2 is given by the geodesic flow on the manifold leaving H^2 constant when starting at q_0 with the initial velocity of v_0. In some cases, this can be written explicitly and more details on this be found in [3].

When this geodesic flow is available, GMC uses a proposal (within MCMC) that is constructed starting from a given q_0 by simulating a new value for $v_0 \sim N(0, \mathscr{P}_q)$ (due to $p \sim N(0, \mathsf{P}(q))$) and then iterating the following steps:

- Compute $q_{\frac{\epsilon}{2}} = q_0, \quad v_{\epsilon/2} = v_0 + \frac{\epsilon}{2}\mathscr{P}_{q_0}\, \nabla_q \log \pi(q)|_{\,q=q_0}$.
- Solve (q_τ, v_τ) according to an appropriate geodesic flow for H^2 for an interval ϵ starting with initial condition being $(q_{\frac{\epsilon}{2}}, v_{\frac{\epsilon}{2}})$ leading to $(q_{\frac{3\epsilon}{2}}, v_{\frac{3\epsilon}{2}})$.
- Compute $q_{2\epsilon} = q_{\frac{3}{2}\epsilon}$, $v_{2\epsilon} = v_{\frac{3}{2}\epsilon} + \frac{\epsilon}{2}\mathscr{P}_{q_{\frac{3}{2}\epsilon}}\, \nabla_q \log \pi(q)|_{\,q=q_{\frac{3}{2}\epsilon}}$

A Metropolis accept-reject step should then follow as usual.

3.2.3. *GMC for Stiefel Manifolds*

In order to implement GMC we need to be able to compute the following key quantities:

(1) the projection $\mathscr{P}$;
(2) the log densities $\log \pi$ together with its gradients; and
(3) the geodesic flows related to H^2.

Recall in our previous specification we assumed the existance of a bijective map $\xi : \mathcal{M} \to \mathbb{R}^n$ that maps every open set of $\mathcal{M}$ to an open set in $\mathbb{R}^n$ and let $x = \xi(q)$.

Remark 7. *A key observation in this section is that as far as $\mathscr{P}$ is concerned often one is able to compute it without requiring knowledge of ξ. Whilst embedding theorems guarantee existence of ξ for smooth or Riemannian manifolds, it is in general unknown and non-trivial to find such mappings. Fortunately in many cases, such as Stiefel manifolds, $\mathscr{P}$ can be described explicitly as $I - UU^T$ where U is an orthonormal basis of the* normal *to the tangent space, so knowledge of ξ is not required.*

To understand this remark more concretely, we must first recall our previous remark that the Stiefel manifold becomes a Riemannian manifold by introducing an inner product in its tangent spaces. In the context in which we work in this section, it is convenient to work with the choice of metric constructed from the "cannonical inner product", rather than the Euclidean inner product when considering the Riemannian structure of the Stiefel Manifold. Then one may obtain the results discussed in [3].

Precisely, the result can be developed by explicitly utilising the selection of the canonical inner product on the tangent space to define the Riemannian embedding of the Stiefel manifold. In this case, the canonical inner product will "weigh" the coordinates equally, where as explained in [32], the concept of the canonical inner product is to attempt to find a matrix A of some tangent vector say Z and weigh it by 1/2 in the inner product, something akin to what is achieved when the Euclidean inner product is adopted. This is performed in general for a Stiefel manifold by considering first representing the tangent vector by

$$Z = XA + X_\perp B \tag{33}$$

where the matrix $A = X^T Z$ and one may write $XA = XX^T Z$. Consequently this produces $(I - \frac{1}{2}XX^T)Z = Z - \frac{1}{2}XX^T Z = XA + X_\perp B - \frac{1}{2}XA = \frac{1}{2}XA + X_\perp B$ and one obtains after some substitutions of the above identities the form given by

$$\mathrm{tr}\left(Z^T(I - \frac{1}{2}XX^T)Z\right) = \frac{1}{2}\mathrm{tr}\left(A^T A\right) + \left(B^T B\right), \tag{34}$$

which allows one to then formally define the cannonical inner product according to the expression

$$\langle Z_1, Z_2 \rangle_c = \mathrm{tr}\left(Z_1^T (I - \frac{1}{2} X X^T) Z_2 \right), \tag{35}$$

and the resulting canonical metric $\langle Z, Z \rangle_c$ which is sometimes referred to as the Killing metric in the case of the orthogonal group (when $n = r$ for the Stiefel manifold). Clearly, this metric product will satisfy the conditions required in the preliminary material in Section 1.2 for properties of metrics required for the set-up.

Recall our goal is to be able to define flows on the Stiefel manifold. To achieve this we also need to consider the representation for differentials on this space. For instance, consider $X \in \mathbb{V}_{n,r}$ for a Stiefel manifold and consider a function F from $\mathbb{R}^{n \times r}$ to $\mathbb{R}$ and matrices $X, Z \in \mathbb{R}^{n \times r}$. Now, denote the differential of F by $\mathcal{D}F_X$ which is the derivative of F in the Z direction at X and is given by

$$\mathcal{D}F_X(Z) = \sum_{i,j} \frac{\partial F}{\partial X_{i,j}} Z_{i,j} = \mathrm{tr}\left(D^T Z \right), \tag{36}$$

where we denote by matrix $D = \left[\frac{\partial F}{\partial X_{i,j}} \right] \in \mathbb{R}^{n \times r}$. Hence, given a point X on the Stiefel manifold $X \in \mathbb{V}_{n,r}$ the resulting differential $\mathcal{D}F_X$ is a linear functional on the tangent space $T_X \in \mathbb{V}_{n,r}$. In this case, when working with the cannonical inner product, one can consider a vector AX and the resulting action of $\mathcal{D}F_X$ on the tangent space $T_X \in \mathbb{V}_{n,r}$ is represented by choice $A = (DX^T - XD^T)$, see a proof of this result in [32]. In general one can then refer to the vector $AX = (DX^T - XD^T)X$ by the more classical vector calculus type notation given by $\nabla_c F$, where this choice of notation simply provides an analogy to suggest that it is the gradient of F under the canonical metric. Note, in this construction the matrix A is a skew symmetric $n \times n$ matrix.

From this result, we may now return back to the application in our context to construct the GMC framework for the sampler of our Bayesian cointegration model. Hence, we can now be precise when we refer to the Stiefel manifold case. If we let $X \in \mathbb{V}_{n,r}$ be a matrix satisfying the required conditions, i.e $X^T X = I$. Then the above results tell as the following regarding the projection, for an arbitrary matrix $W \in \mathbb{R}^{n \times r}$, the projection onto $\mathbb{V}_{n,r}$ at X which can be represented by

$$\mathscr{P}_X(W) = W - \frac{1}{2} X (X^T W - W^T X), \tag{37}$$

see additional discussion in [3].

In addition, given this structure we can then also obtain in this case an explicit formulae for the geodesic flows on $\mathbb{V}_{n,r}$ which are given by the following expression

$$[X(\tau), v_X(\tau)] = [X(0), v_X(0)] \exp\left(\tau \begin{bmatrix} D & -S(0) \\ I & D \end{bmatrix} \right) \left[\begin{bmatrix} \exp(-\tau D) & 0 \\ 0 & \exp(-\tau D) \end{bmatrix} \right], \tag{38}$$

where $D = X(\tau)^T v_X(\tau)$ is constant over the geodesic and $S(0) = v_X(0)^T v_X(0)$; see [9] and [51] for explicit details. So for a given (differentiable w.r.t X) density $\pi(X)$ one can implement GMC as described in Algorithm 2.

Algorithm 2 One iteration of the GMC of [3] for sampling from a $\pi(X)$ with $X \in \mathbb{V}_{n,r}$.

(1) Sample $v_X \sim N(0, I_{n\cdot r})$ and apply projection at X, $v_X \leftarrow \mathscr{P}_X(v_X)$
(2) Compute $h \leftarrow \log \pi(X) - \frac{1}{2} v_X^T v_X$.
(3) for $\tau = 1, ..., T$ do:
  (a) Compute $v_X \leftarrow v_X + \frac{\epsilon}{2} \nabla_X \log \pi$ and apply projection $v_X \leftarrow \mathscr{P}_X(v_X)$.
  (b) Compute (X^*, v_{X^*}) by implementing (38) $X^*(0) = X$ and for a time interval ϵ.
  (c) Compute $v_{X^*} \leftarrow v_{X^*} + \frac{\epsilon_X}{2} \nabla_{X^*} \log \pi$ and apply projection $v_{X^*} \leftarrow \mathscr{P}_{X^*}(v_{X^*})$
(4) Compute $h^* = \log \pi(X^*) - \frac{1}{2} v_{X^*}^T v_{X^*}$ and sample $u \sim \mathcal{U}(0,1)$. If $u < \exp(h^* - h)$, then return X^*, otherwise return X.

3.3. GMC for Cointegration Parameters

To avoid identifiability issues we have restricted β to lie on a Stiefel manifold and follow $\beta^T \beta = I_r$. We have already presented how to implement GMC for β but have not looked at the densities and their gradients. The other variable of interest, $\alpha \in \mathbb{R}^{n \times r}$, is unrestricted, which can be interpreted as an element of Euclidean space or flat manifold. In this case, there is no need to perform any projections $\xi = I$ and the geodesic flows are just straight lines:

$$[a(\tau), v_\alpha(\tau)] = [\alpha(0), v_\alpha(0)] \begin{bmatrix} 1 & 0 \\ \tau & 1 \end{bmatrix}, \tag{39}$$

see Section 4.4 of [3] for more details on extending GMC for target distributions products of manifolds. Given expressions for the densities $\pi(\alpha, \beta)$ are available it is possible to proceed with GMC implementations targeting the cointegration parameters. We will look at two particular cases below.

3.3.1. *GMC Targetting Jointly α, β*

When one is interested in sampling directly from $p(\alpha, \beta | y_{1:T}, R)$ in (18), then the variable of interest (α, β) lies on the Cartesian product of a Euclidean space and a Stiefel manifold, so $(\alpha, \beta) \in \mathbb{R}^{n \times r} \times \mathbb{V}_{n,r}$, which is itself an embedded manifold. Geodesics (and tangent vectors) on $\mathbb{R}^{n \times r} \times \mathbb{V}_{n,r}$ are simply the Cartesian product of the geodesics (and tangent vectors resp.) on $\mathbb{R}^{n \times r}$ and $\mathbb{V}_{n,r}$ and similarly for the orthogonal projections $\mathscr{P}$. A detailed description of the GMC algorithm for $p(\alpha, \beta | y_{1:T}, R)$ is presented in supplementary material found in an earlier version of this paper, [46], together with derivations for the gradients for Gaussian ϵ_t.

3.3.2. *Using GMC within a Gibbs Sampler Approach*

To pursue greater efficiency than targetting jointly (α, β), we propose using GMC or HMC samplers within a Gibbs approach and perform an HMC sampler for $p(\alpha | \beta, y_{1:T})$ and GMC for $p(\beta | \alpha, y_{1:T})$. This is a fairly straightforward extension to what has been presented so far, which in Section 5 shows very good performance.

One could aim to extend using samplers within a Gibbs approach for different parameterizations of Π. The motivation comes from trying to extend the efficiency found in [41] for the cases where conjugate sampling from full conditionals is not possible. The latter can be potentially replaced with HMC or GMC samplers invariant to them. The key in [41] was the particular parameterization used and a partial collapsing step for κ. One can explore various parameterizations, but we emphasize that when a Metropolis sampler replaces a full conditional in a partially collapsed Gibbs sampler, care must be taken, otherwise the algorithm might not be valid anymore; see [68] for details.

For example, when plain HMC updates are used in Algorithm 1 instead of direct sampling from $p(\alpha|\beta, y_{1:T})$ and $p(\mathcal{B}|\mathcal{A}, y_{1:T})$ then the algorithm is not valid and in numerical results not shown here we confirmed that such an approach does not converge to the right stationary distribution.

We will consider using a singular value decomposition (SVD) for Π ([5, p. 20]). Let $\Pi = \mathcal{U}\mathcal{S}\mathcal{V}^T$, then β enters the model indirectly through $\mathcal{V}$. A nice feature of SVD is that the singular values can be viewed as mean reversion parameters and can be used to influence the rank when choosing priors. Such priors have been proposed in [37, 38], so when combined with priors on the spaces of matrices $\mathcal{U}, \mathcal{V}$, then GMC samplers can be particularly useful. In addition, to ensure that this singular value parameterization is unique ([5, p. 20]), we will restrict $\mathcal{U}$ to have all its first row elements positive (and denote the corresponding space as $\tilde{\mathbb{V}}_{n,n}$). The parameterization from the SVD should follow: $\mathcal{U} \in \tilde{\mathbb{V}}_{n,n}$, $\mathcal{S} = diag(s_1, \ldots, s_r, 0, \ldots, 0)$ with $s_1 > s_2 > \ldots > s_r > 0$, and $\mathcal{V} := [\beta, \beta_\perp] \in \mathbb{V}_{n,n}$.

A MCMC within Gibbs method for this parameterisation will alternate samplers targetting $p(\beta|\mathcal{S}, \mathcal{V}, y_{1:T})$, $p(\mathcal{U}|\mathcal{S}, \beta, y_{1:T})$ and $p(\mathcal{S}|\mathcal{U}, \beta, \mathcal{V}, y_{1:T})$. The first two conditionals can be sampled using GMC and for the latter any choice of MCMC sampler on $p(s_1, \ldots, s_r|\mathcal{U}, \mathcal{S}, \beta, y_{1:T})$ can be used. A single iteration of the algorithm is presented in Algorithm 3, which outputs $\tilde{\Pi} = \tilde{\mathcal{U}} diag(\tilde{s}_1, \ldots, \tilde{s}_r)\tilde{\beta}^T$. In step 2 (a) we present a sign-flipping method that ensures that the same unique SVD parameterization is preserved and $\mathcal{U}$ has only positive elements in the first row. For the particular implementation in Algorithm 3 this step is optional as the conditional $\mathcal{U}|\mathcal{S}, \beta, y_{1:T}$ does not require a unique parameterization: $\mathcal{U}$ is unique once $\mathcal{S}, \beta$ are fixed. We include this step as it can be useful in various extensions, such as a random scan Gibbs sampler. Similarly, the unique SVD decomposition requires $s_1 > s_2 > \ldots > s_r > 0$, which in Algorithm 3 is again optional, but if required can be imposed by either the MCMC proposals or by reordering of columns of $\mathcal{U}, \mathcal{S}, \mathcal{V}$.

Algorithm 3 One iteration of the GMC within Gibbs sampler for cointegration ECM.

(1) Iterate Algorithm 2 for $\pi(\beta) \propto p(\beta|\mathcal{U}, \mathcal{V}, y_{1:T})$ to get $\tilde{\beta}$
(2) Iterate Algorithm 2 for $\pi(\mathcal{U}) \propto p(\mathcal{U}|\tilde{\mathcal{V}}, \mathcal{S}, \tilde{\beta}, y_{1:T})$ to get $\tilde{\mathcal{U}}$.
  (a) to project from $\mathbb{V}_{n,n}$ to $\tilde{\mathbb{V}}_{n,n}$ sign flipping can be used:
  For $l = 1, \ldots, n$: if $\tilde{\mathcal{U}}[1, l] < 0$, set $\tilde{\mathcal{U}}[i, l] \leftarrow -\tilde{\mathcal{U}}[i, l]$ and $\tilde{\mathcal{V}}[i, l] \leftarrow -\tilde{\mathcal{V}}[i, l]$ for all $i = 1, \ldots, r$.
(3) Perform a few iterations of a MCMC sampler for $p(s_1, \ldots, s_r|\tilde{\mathcal{U}}, \tilde{\beta}, \tilde{\mathcal{V}}, y_{1:T})$ to get $\tilde{\mathcal{S}}$.

Remark 8. SVD can be employed also only for α instead of Π. Using then $\alpha = \mathcal{U}\mathcal{S}\mathcal{V}^T\beta^T$ requires using an extra (redundant) parameter, but this can be useful for comparing with other methods or parameterizations, when the priors are specified for α, β and one can derive appropriate prior distributions for $\mathcal{U}, \mathcal{S}, \mathcal{V}$. We will perform such comparisons later in Section 5.

Remark 9. Another approach could consider the polar decomposition for α, whereby $\Pi = \mathcal{A}\kappa^{1/2}\beta^T$. Here κ takes values in the space of positive semidefinite matrices, and $\mathcal{A}, \beta \in \mathbb{V}_{n,r}$. A GMC within Gibbs approach could alternate a small number of iterations between GMC samplers targetting $p(\beta|\mathcal{A}, \kappa, y_{1:T})$, $p(\kappa|\beta, \mathcal{A}, y_{1:T})$ and $p(\mathcal{A}|\beta, \kappa, y_{1:T})$. The GMC sampler for $p(\kappa|\beta, \mathcal{A}, y_{1:T})$ can be designed using a projection $\mathscr{P}$ based on eigen-decompositions or optimization (see [28, Section 4.2.3]) and geodesics found in [31] together with a similar GMC approach. In the interest of brevity we will not pursue this further.

4. Practical Estimation of Cointegration Space

In this section, we describe some additional details on estimation methods for the basis of cointegration space, such as how to assess similarity of estimated cointegration spaces and rank estimation. Whilst our main motivation comes from using MCMC in a Bayesian framework, the tools presented here can be used when comparing point estimates from different methods.

4.1. **Measure of Cointegration Similarity**

Suppose we have two estimates for the cointegration matrix, β_1 and β_2. As we will require a distance for comparison diagnostics, we will use the approach of [43]. Suppose we can decompose β_2 as follows:

$$\beta_2 = \beta_1\gamma_1 + \beta_{1\perp}\gamma_2,$$

where $\beta_{1\perp}$ is the orthogonal complement of β_1, $\gamma_1 \in \mathbb{R}^{r\times r}$ and $\gamma_2 \in \mathbb{R}^{(n-r)\times r}$. These matrices can be explicitly written as $\gamma_1 = \beta_1^T\beta_2$ and $\gamma_2 = \beta_{1\perp}^T\beta_2$. A distance measure between β_1 and β_2 is

$$d_s(\beta_1, \beta_2) = tr(\beta_2\beta_{1\perp}\beta_{1\perp}^T\beta_2)^{1/2}. \tag{40}$$

Note that a distance between β_1 and β_2 is equivalent to measuring a dissimilarity of $col(\beta_1)$ and $col(\beta_2)$. Also we have, $d_s(\beta_1, \beta_2) \leq min\,(r, (n-r))$, which is useful for interpretation.

4.2. **Bayesian Point Estimation**

Obtaining point estimates of the cointegration space needs attention when the posterior distribution is defined on a Stiefel manifold. Following [72], we use the Frobenius norm as the loss function,

$$l(\beta, \beta^*) = ||\beta\beta^T - \beta^*(\beta^*)^T||_F;$$

where β and β^* are semi-orthogonal. To provide a Bayesian point estimate of the posterior distribution for $col(\beta)$, we use the Posterior Mean Cointegration Space Estimator (PMCS) proposed in in [72]. The PMCS estimator is defined as

$$\hat{\beta} \stackrel{def}{=} \underset{\tilde{\beta}\in\mathbb{V}_{n,r}}{\arg\min}\mathbb{E}[\,l(\beta, \tilde{\beta})\,\big|\, y_{1:T}].$$

In [72], it was showed that the PMCS estimator can be computed as

$$\hat{\beta} = (\mathtt{v}_1, ..., \mathtt{v}_r), \tag{41}$$

where $\mathtt{v}_i$ is the eigenvector of $\mathbb{E}[\,\beta\beta^T\big|\, y_{1:T}]$ corresponding to the i-th largest eigenvalue. Given the Stiefel manifold is a compact space, all finite moments of the elements of β exist in the orthonormal normalization, which implies existence of $\mathbb{E}[\,\beta\beta^T\big|\, y_{1:T}]$. A closed form analytic expression for $\mathbb{E}[\,\beta\beta^T\big|\, y_{1:T}]$ is not available, so we resort to MCMC to estimate it. After N iterations of a MCMC procedure one can use $\frac{1}{N}\sum_{i=1}^N \beta_i\beta_i^T$, where β_i denotes the i-th sample of MCMC.

4.2.1. *Measure of Posterior Variation*

As a tool to assess the variation of posterior cointegration space distribution from the output of MCMC sampler, we will use the projective Frobenius span variation introduced in [72]. The projective Frobenius Span Variation (FSV) is defined as $\tau^2_{sp} = \frac{\mathbb{E}[\, d(\beta,\hat{\beta}) | y_{1:T}]}{r(p-r)/p}$, where $\hat{\beta}$ is the PMCS estimate of β. We can estimate this diagnostics with$\widehat{\tau^2_{sp}} = \frac{r - \sum_{i=1}^{r} \lambda_i}{r(p-r)/p}$, where λ_i is the i-th largest eigenvalue of $\frac{1}{N}\sum_{i=1}^{N} \beta_i \beta_i^T$.

4.3. **Rank Estimation**

While we have been considering r known so far, the cointegration rank has be to estimated. This can be done using a variety of approaches: Bayesian, frequentist or using a spectral methods. In a Bayesian setup, estimation of r is performed using Bayes factors, which in the context of ECMs are computed by means of Sevage-Dickey density ratio; see [39], [17] for an overview and [53] and [40] for financial applications. In frequentist settings, choosing r is routinely performed by means of hypothesis testing. Two most commonly used test statistics give rise to the "Trace Test" and "Maximum Eigenvalue Tests"; see [35] for details. In the work [55] the authors also study the effect of cointegration miss-specification of the rank and the role and influence of the identification constraints. They show that being conservative and overestimating an uncertain cointegration rank is often practically more robust in the model estimation and has less influence on Bayesian model parameter estimation compared to under-estimating the rank. Finally, one could use non-parametric or spectral approaches to estimate r (as well as the cointegration space), see [74] for a recent approach.

5. Numerical results

In the simulation results presented below, we use simulated data with $n = 4,\ r = 3$, (so $d_s(\beta, \hat{\beta}) < 1$.) In each case (Gaussian or Student-t), $T = 240$ the parameter values generating $y_{1:T}$ are the following:

$$\alpha^{true} = 0.2 \begin{pmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix};\ \beta^{true} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix};\ R = I_n$$

For the Student-t case we shall use $\omega = 20$. For the priors we use $\nu = 1$, $P_\tau = I_n$, and $G = I_r$. In order to quantify the and accuracy of MCMC samplers, we will compute the diagnostics based on their resulting chain output, e.g. $\{(\alpha_i, \beta_i)\}_{i=1}^{J}$, with $J = 10^4$ taken after a suitable burn-in period of 1000 iterations. We will use the Integrated Autocorrelation time (IACT) for each element of Π, $IACT = 1 + 2\sum_{i=1}^{\infty} a(i)$, where $a(i)$ denotes autocorrelation with lag i, to assess mixing efficiency of different MCMC samplers, the Average Distance $AD = \frac{1}{J}\sum_{i=2}^{J} d_s(\beta_i, \beta^{true})$ to compare the accuracy of the cointegration space estimation and similarly $||\hat{\Pi} - \Pi^{true}||_F$ to compare estimation of Π, where $\hat{\Pi}$ is constructed using the PMCS. We note that AD provides also insight on how much the sampler explores the support of the distribution of interest. In addition we will use FSV defined earlier as $\tau^2_{sp} = \frac{\mathbb{E}[\, d(\beta,\hat{\beta}) | y_{1:T}]}{r(p-r)/p}$ to assess the variation of the simulated posteria.

(A) Gibbs Sampler

1.17	1.30	1.24	1.22
1.22	1.26	1.13	1.22
1.29	1.22	1.25	1.29
1.25	1.26	1.19	1.19

(B) GMC for β and Gibbs for α

1.11	1.05	1.11	1.18
1.18	1.18	1.23	1.09
1.13	1.10	1.18	1.17
1.11	1.18	1.12	1.12

(C) GMC for β and HMC for α

7.78	8.88	9.80	13.43
8.62	8.67	10.31	13.47
9.59	9.53	9.76	13.82
8.65	9.68	9.62	13.57

(D) GMC targetting jointly (α, β)

75.70	93.60	94.06	141.02
93.20	72.47	71.54	107.08
49.98	59.12	105.03	143.28
34.70	79.78	87.27	147.57

TABLE 1. IACT for each entry of Π for Gaussian case

	Gibbs	GMC + Gibbs	GMC + HMC	GMC on (α, β)
AD	0.05	0.02	0.07	0.05
$\|\|\hat{\Pi} - \Pi^{true}\|\|_F$	0.09	0.47	0.10	0.12
τ_{sp}^2	0.003	0.0001	0.004	0.38

TABLE 2. Comparison of different MCMC samplers for Gaussian case: Gibbs Sampler, GMC for β with Gibbs for α, GMC for β with HMC for α, and GMC targetting jointly (α, β).

5.1. The Gaussian case

For the Gaussian case, in Tables 1, 2 we present the results comparing the Gibbs Sampler of [41], a Gibbs approach with GMC for β and the full conditional update for α, a Gibbs approach with GMC for β and HMC for α as well as GMC targetting jointly (α, β). The perfect Gibbs sampler and the Gibbs method with GMC for β and a perfect Gibbs update for α, show very good performance. In some cases the slight improvement when using GMC for β can be attributed to the additional computational effort. In addition, using GMC for α and HMC for β shows good performance that is in par with the Gibbs sampler in terms of the comparison in Table 2 but with a slower mixing as indicated by the IACT in Table 1. Finally, using GMC to target jointly (α, β) has very slow mixing and posterior exploration, but at least it manages to result to accurate point estimates.

5.2. The Student-t case

For the Student-t case we will consider two different cases. The first one would be to use the data augmentation (DA) approach outlined in 2.4.2. As the full conditionals of the auxiliary variables λ_t are tractable, a Gibbs sampler can be implemented (by alternating Algorithm 1 with $p(R|y_{1:T}, \alpha, \beta, \lambda_{1:T}, \omega)$ and $p(\lambda_t|y_{1:T}, \alpha, \beta, R, \omega)$ for each t). Also, when GMC is used in this case either as part of a Gibbs update or on its own, a Gaussian likelihood $p(y_{1:T}|\alpha, \beta, R, \omega, \lambda_{1:T})$ will be used in the implementation of 2. In the second case, we will apply the GMC samplers implemented

(A) Gibbs Sampler with data augmentation

1.23	1.46	1.19	1.43
1.19	1.19	1.11	1.09
1.20	1.24	1.09	1.00
1.16	1.43	1.27	1.06

(B) GMC for β and Gibbs for α

1.38	1.63	1.30	1.28
1.12	1.44	1.47	1.16
1.23	1.45	1.54	1.09
1.48	1.83	1.80	1.29

(C) GMC for β and HMC for α

10.05	14.63	14.91	22.25
16.12	14.58	16.63	26.21
12.42	12.06	13.12	24.18
13.93	15.44	15.10	22.96

(D) GMC for β and GMC with SVD for α

183.28	155.22	140.93	117.11
185.13	191.43	176.33	100.76
84.78	227.81	154.99	138.87
93.62	114.11	121.16	130.47

(E) GMC targetting jointly (α, β)

476.85054	703.5214	844.40542	873.98745
661.30924	684.36671	534.87151	696.74944
709.07688	427.01486	522.67915	667.17075
95.70291	409.73287	422.26648	704.8838

Table 3. IACT for each entry of Π for Student-t case with data augmentation

(A) GMC for β and HMC for α

9.019	11.40	14.39	22.54
12.09	11.41	13.99	22.51
11.75	9.14	14.61	22.59
8.52	12.71	12.63	20.98

(B) GMC for β and GMC with SVD for α

186.61	145.06	131.14	117.70
201.23	210.11	169.41	103.02
87.04	229.32	139.97	132.64
102.54	109.85	106.38	130.55

(C) GMC targetting jointly (α, β)

421.71	796.78	583.57	839.46
765.81	608.70	718.14	818.34
439.31	392.44	591.99	546.19
515.19	715.83	829.71	863.59

Table 4. IACT for each entry of Π for Student-t case with partial collapsing in GMC

with a Student-t distributed likelihoods $p(y_{1:T}|\alpha, \beta, R, \omega)$ instead. The rest of the updates for the remaining parameters will be as before, so one can view this approach as a partial collapsing scheme.

In Tables 3,4, 5 we present the analogous results as in the Gaussian case. When data augmentation is used we compare the same samplers as in the Gaussian case together with an implementation of GMC for the SVD decomposition of α. The latter is included for comparison purposes as for this choice of prior for α one can derive analytically the prior for the matrices and singular values in the SVD decomposition ([5, Theorem 1.5.4]). As mentioned in Remark 8, using SVD directly for Π is more parsimonious and hence one should expect better results (see below). Not surprisingly, the different methods compared behave similarly for the Student-t case with data augmentation to what has been seen in the Gaussian case. In Figure 1 we also present trace plots for the consecutive MCMC realizations of β and the components of polar decompositions of estimates of α, that is $\hat{\mathcal{A}}, \hat{\kappa}$.

(A) Data augmentation and Gaussian likelihood

	Gibbs	GMC + Gibbs	GMC + HMC	GMC with SVD for α	GMC on (α, β)
AD	0.01	0.01	0.01	0.01	0.01
$\|\|\hat{\Pi} - \Pi^{true}\|\|_F$	0.12	0.12	0.122	0.16	0.186
τ^2_{sp}	0.0001	0.0001	0.0003	0.0002	0.001

(B) Student-t likelihood with partial collapsing

	GMC + HMC	GMC with SVD for α	GMC on (α, β)
AD	0.009	0.01	0.40
$\|\|\hat{\Pi} - \Pi^{true}\|\|_F$	0.12	0.15	0.835
τ^2_{sp}	0.0002	0.002	0.04

Table 5. Comparison of different MCMC samplers for Student-t case:

For the case where the Student-t likelihood and partial collapsing is used, we only compare the GMC based methods. The results are similar to the data augmentation case, but they can be used as numerical evidence of validity of the GMC/HMC schemes in the non-Gaussian case. From all the results it is apparent that GMC for β with HMC for α performs best among the more generic samplers (that do not use the full conditionals). InTable 5, it shows similar performance as the Gibbs sampler and it is only inferior in terms of the IACTs, which admit higher values than the Gibbs sampler but are still reasonably good values. In addition, we note that the use of SVD for α results in an improvement in efficiency from targetting jointly (α, β) with GMC, but both samplers are much slower in the exploration of the posterior than the other methods.

5.2.1. *Using SVD for* Π

In this section, we examine the model reparameterisation based on SVD decomposition imposed directly on the long-run multiplier matrix Π. This case is treated separately as the priors used here are not equivalent with those used before. To our best knowledge, this has not been considered before in the context of Bayesian cointegration, however GMC enables us to pursue this direction. For convenience, we use $p(\Pi) \propto N_{n,n}(0, I_n, I_n)1_{rank(\Pi)=r}$. From [5, Theorem 1.5.4] we obtain the equivalent expression for the priors defined on the matrices and singular values in the SVD decomposition: $\mathcal{U}, \mathcal{V}$ are uniform distributions on $\tilde{\mathbb{V}}_{n,n}$, $\mathbb{V}_{n,r}$ resp. and

$$p(s_1, \ldots, s_r) \propto \exp\left(-\frac{1}{2}\sum_{l=1}^{r} s_l^2\right)\prod_{l=1}^{r} s_l^{n-r} \prod_{l<j,l=1}^{r}\left(s_l^2 - s_j^2\right)$$

For the MCMC sampler targetting $p(\mathcal{S}|\mathcal{U}, \beta, \mathcal{V}, y_{1:T})$ we use a Metropolis-Hastings approach, with the proposal being a random walk on the log space of each s_l with a step size of 0.1 and the noises being independent standard normals sorted in descending order. For $p(\beta|\mathcal{S}, \mathcal{V}, y_{1:T})$, $p(\mathcal{U}|\mathcal{S}, \beta, y_{1:T})$ GMC samplers are used. For samplers targetting $\mathcal{U}$ we use simpler to compute geodesic flows as presented in [3, 9]. The results are presented in Tables 6, 7 for a GMC within Gibbs approach in Algorithm 3 and a random scan implementation that updates β with probability 0.1 and $\mathcal{U}, \mathcal{S}$ with

(A) Data augmentation and Gaussian likelihood

42.79	65.59	45.50	38.51
84.48	85.44	50.98	48.45
50.17	38.771	90.38	58.15
68.85	95.60	98.25	57.76

(B) Student-t likelihood

77.13	61.71	92.96	80.65
100.46	105.14	104.02	51.99
110.36	118.72	105.39	62.49
66.02	47.73	70.46	61.40

(C) Student-t likelihood

65.65	75.13	89.68	49.95
57.12	99.29	77.99	41.70
66.46	98.35	132.39	40.17
64.78	112.83	97.63	61.12

(D) Student-t likelihood with Gaussian Data Augmentation

142.30	105.35	123.65	119.89
139.24	110.10	71.07	67.00
148.19	105.91	169.58	48.39
123.04	101.45	103.39	118.64

Table 6. IACT of using SVD for Π Top: Standard GMC within Gibbs; bottom: random scan Gibbs.

(A) Standard GMC within Gibbs

	Student-t with DA (Gaussian)	Student-t
AD	0.04	0.01
$\|\|\hat{\Pi} - \Pi^{true}\|\|_F$	0.16	0.15
τ_{sp}^2	0.001	0.002

(B) Random Scan Gibbs

Student-t with DA (Gaussian)	Student-t
0.01	0.01
0.14	0.16
0.001	0.0001

Table 7. Performance of GMC using SVD for Π for different likelihoods

0.45 each. From the IACTs we can observe an improvement in the mixing over the sampler that used SVD for α earlier.

The motivation behind using a random scan sampler is to reduce the computational cost. The results in Tables 6, 7 are similar with the the systematic scan. As the update for β is the most expensive step and hence is used less often. [4] show how geodesics for $\mathcal{U}$ can be computed more efficiently as opposed to the geodesics on $\mathbb{V}_{n,r}$. On the downside, one ends up with sampler mixing properties, but this can be improved by increasing the probability of update of β.

6. Discussion

Bayesian methods are favorable to use in situations when estimation of the parameters uncertainty is required, but requires advanced simulation techniques like MCMC. In this paper we presented the Gibbs Sampler of [41] together with the GMC algorithm of [3] that can be used to sample from a wide class of distributions defined on manifolds. We combined these two approaches and presented different approaches for Bayesian estimation of the cointegration space. The Gibbs sampling method is the current state of the art for this problem, but can be implemented only when full conditionals are available. On the other hand, GMC is generic and can be used in wider range of model setups, that also seem more realistic in finance, such as heavy tailed noises, time varying model parameters as well as different choices of prior distributions. In terms of performance in our numerical results we saw GMC performed very well and efficiently when used within a Gibbs update scheme (i.e. GMC for β and HMC for α), but when applied naively to target (α, β) jointly it turned out to be rather inefficient.

A question that may arise is whether Bayesian methods and sampling techniques will be accurate in practical contexts. We believe this is a topic demanding a detailed discussion, but we performed a few numerical experiments to motivate the work presented so far. These are contained in the supplementary material in [46], where we compared the Gibbs sampler of [41] with the spectral method of [74] under various model parameterization setups with Gaussian noise. Both methods seemed accurate overall with the Gibbs sampler performing better, but on the other hand the spectral method of [74] uses only a fraction of the computational cost. In both cases, accuracy seemed to be robust to the level of additive noise, e.g. the trace of R, but more sensitive to a mean reversion parameter of the implied spread process. Using only the spectral method of [74] we found that this sensitivity of the estimated cointegrated space on the mean reversion seemed to hold also when a non-Gaussian noise was used with a variety of features in terms of skewness, kurtosis, or heavy-ness of the tails.

In terms of Bayesian estimation of cointegration, when non-Gaussian (differentiable) densities with heavier tails are used the advanced GMC methods can be very valuable to assess the uncertainty around point estimates. In addition, GMC can be implemented also directly on SVD decompositions of Π. This lead to accurate estimates, but further work is required to improve the mixing of Algorithm 3. Here a simplification choice was made to apply Step 2 (a) of Algorithm 3 after a GMC iteration defined on $\mathbb{V}_{n,r}$, whereas Step 2 (a) should have been emdedded in the GMC procedure (and the implementation of Algorithm 2). More importantly from a modelling perspective SVD is appealing as the singular values act in a similar manner to mean reversion parameters mentioned earlier, so we believe that their accurate estimation will be critical for cointegration space estimation. In addition, given that computational tools like GMC are available and can perform well within a Gibbs scheme, this opens up the possibility of adopting a wider class of priors (and parameterizations such as SVD) than what described in Section 2.4.1. Priors could be defined on parameters resulting from SVD or polar decompositions directly and past works on defining priors in different contexts (e.g. [37, 38]) can be very relevant.

We conclude with some more possible aspects for further investigations from a computational perspective. In GMC and HMC we used a simple numerical integration scheme for the Hamiltonian dynamics numerical integrator, so there is plenty of room for improvements such as the recent work of [30]. Finally, interesting questions arise when n is very large. Is it possible to perform estimation efficiently and are the conclusions we reached here still valid? In numerical results not shown we have applied the above methods up to $n = 10$, and our conclusions seemed to hold. In order to use much higher values of n a significant amount of developments is needed.

References

[1] T. W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, pages 327–351, 1951.

[2] M. D. Branco and D. K. Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79(1):99–113, 2001.

[3] S. Byrne and M. Girolami. Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

[4] J. R. Cardoso and F. S. Leite. Exponentials of skew-symmetric matrices and logarithms of orthogonal matrices. *Journal of computational and applied mathematics*, 233(11):2867–2875, 2010.

[5] Y. Chikuse. *Statistics on special manifolds*, volume 174. Springer Science & Business Media, 2012.

[6] N. Dobigeon and J.-Y. Tourneret. Bayesian orthogonal component analysis for sparse representation. *IEEE Transactions on Signal Processing*, 58(5):2675–2685, 2010.

[7] J. J. Dolado, J. Gonzalo, and F. Marmol. *Cointegration*, pages 634–654. Blackwell Publishing Ltd, 2007.

[8] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195:216–222, 1987.

[9] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[10] R. F. Engle and C. W. Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276, 1987.

[11] R. F. Engle and B. S. Yoo. Forecasting and testing in co-integrated systems. *Journal of econometrics*, 35(1):143–159, 1987.

[12] F. J. Fabozzi, S. Rachev, and S. Focardi. Financial econometrics: From basics to advanced modeling techniques, 2007.

[13] H. Federer. *Geometric measure theory*. Springer, 2014.

[14] C. Fernández and M. F. Steel. Multivariate student-t regression models: Pitfalls and inference. *Biometrika*, pages 153–167, 1999.

[15] T. C. Fonseca, M. A. Ferreira, and H. S. Migon. Objective bayesian analysis for the student-t regression model. *Biometrika*, 95(2):325–333, 2008.

[16] L. Gatarek and S. Johansen. Optimal hedging with the cointegrated vector autoregressive model allowing for heteroscedastic errors. *Tinbergen Institute Discussion Papers*, 2016.

[17] L. Gatarek, L. F. Hoogerheide, R. Kleijn, and H. K. Van Dijk. Prior ignorance, normalization and reduced rank probabilities in cointegration models. Technical report, 2010.

[18] J. Geweke. Bayesian treatment of the independent student-t linear model. *Journal of applied econometrics*, 8(S1), 1993.

[19] M. Girolami and B. Calderhead. Riemann Manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214, 2011.

[20] C. Granger and A. Weiss. Time series analysis of error correction models. *Spectral analysis, seasonality, nonlinearity, methodology and forecasting: collected papers of Clive WJ Granger*, page 129, 2001.

[21] A. K. Gupta and D. K. Nagar. *Matrix variate distributions*, volume 104. CRC Press, 1999.

[22] E. Hairer, C. Lubich, and G. Wanner. Geometric numerical integration illustrated by the störmer–verlet method. *Acta numerica*, 12:399–450, 2003.

[23] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.

[24] J. D. Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

[25] R. I. Harris. *Using cointegration analysis in econometric modelling*, volume 82. Prentice Hall London, 1995.

[26] F. Hausdorff. Dimension und äußeres maß. *Mathematische Annalen*, 79(1):157–179, 1918.

[27] D. F. Hendry. Econometric modelling with cointegrated variables: an overview. *Oxford bulletin of economics and statistics*, 48(3):201–212, 1986.

[28] J.-B. Hiriart-Urruty and J. Malick. A fresh variational-analysis look at the positive semidefinite matrices world. *Journal of Optimization Theory and Applications*, 153(3):551–577, 2012.

[29] P. D. Hoff. Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2): 438–456, 2009.

[30] M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[31] A. Holbrook, A. Vandenberg-Rodes, and B. Shahbaba. Bayesian inference on matrix manifolds for linear dimensionality reduction. *arXiv preprint arXiv:1606.04478*, 2016.

[32] I. M. James. *The topology of Stiefel manifolds*, volume 24. Cambridge University Press, 1976.

[33] S. Johansen. Statistical analysis of cointegration vectors. *Journal of economic dynamics and control*, 12(2-3):231–254, 1988.

[34] S. Johansen. Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: Journal of the Econometric Society*, 59:1551–1580, 1991.

[35] S. Johansen and K. Juselius. Maximum likelihood estimation and inference on cointegration with applications to the demand for money. *Oxford Bulletin of Economics and statistics*, 52 (2):169–210, 1990.

[36] S. Johansen. *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford University Press on Demand, 1995.

[37] F. Kleibergen and R. Paap. Priors, posteriors and bayes factors for a bayesian analysis of cointegration. *Journal of Econometrics*, 111(2):223–249, 2002.

[38] F. Kleibergen and H. K. Van Dijk. Bayesian simultaneous equations analysis using reduced rank structures. *Econometric Theory*, 14(06):701–743, 1998.

[39] G. Koop, R. W. Strachan, H. K. v. Dijk, and M. Villani. Bayesian approaches to cointegratrion. Technical report, Econometric Institute Research Papers, 2005.

[40] G. Koop, R. Leon-Gonzalez, and R. Strachan. Bayesian inference in a cointegrating panel data model. In *Bayesian Econometrics*, pages 433–469. Emerald Group Publishing Limited, 2008.

[41] G. Koop, R. León-González, and R. W. Strachan. Efficient posterior simulation for cointegrated models with priors on the cointegration space. *Econometric Reviews*, 29(2):224–242, 2009.

[42] G. Koop, R. Leon-Gonzalez, and R. W. Strachan. Bayesian inference in the time varying cointegration model. *Tinbergen Institute Discussion Papers*, 2008.

[43] R. Larsson and M. Villani. A distance measure between cointegration spaces. *Economics Letters*, 70(1):21–27, 2001.

[44] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[45] K. Mardia and C. Khatri. Uniform distribution on a Stiefel manifold. *Journal of Multivariate Analysis*, 7(3):468–473, 1977.

[46] M. Marówka, G. W. Peters, N. Kantas, and G. Bagnarosa. Estimation of cointegrated spaces: A numerical case study on efficiency, accuracy and influence of the model noise. 2017. URL `https://ssrn.com/abstract=2918511`.

[47] P. Mattila. *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*, volume 44. Cambridge university press, 1999.

[48] R. M. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics, Springer, 1996.

[49] R. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press, 2010.

[50] N. M. Newmark. A method of computation for structural dynamics. *Journal of the engineering mechanics division*, 85(3):67–94, 1959.

[51] Y. Nishimori and S. Akaho. Learning algorithms utilizing quasi-geodesic flows on the stiefel manifold. *Neurocomputing*, 67:106–135, 2005.

[52] G. W. Peters, B. Kannan, B. Lasscock, and C. Mellen. Model Selection and Adaptive Markov chain Monte Carlo for Bayesian Cointegrated VAR model. *Bayesian Analysis*, 5(3):465–492, 2010.

[53] G. W. Peters, B. Kannan, B. Lasscock, C. Mellen, et al. Model selection and adaptive markov chain monte carlo for bayesian cointegrated VAR models. *Bayesian Analysis*, 5(3):465–491, 2010.

[54] G. W. Peters, B. Kannan, B. Lasscock, C. Mellen, S. Godsill, et al. Bayesian cointegrated vector autoregression models incorporating alpha-stable noise for inter-day price movements via approximate bayesian computation. *Bayesian Analysis*, 6(4):755–792, 2011.

[55] G. W. Peters, B. Lasscock, and K. Balakrishnan. Rank estimation in cointegrated vector autoregression models via automated trans-dimensional markov chain monte carlo. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop on*, pages 41–44. IEEE, 2011.

[56] P. C. B. Phillips. Optimal inference in cointegrated systems. *Econometrica*, 59(2):pp. 283–306, 1991.

[57] P. C. Phillips et al. Some exact distribution theory for maximum likelihood estimators of cointegrating coefficients in error correction models. *Econometrica*, 62:73–73, 1994.

[58] C. A. Rogers. *Hausdorff measures*. Cambridge University Press, 1998.

[59] J. H. Stock and M. W. Watson. Testing for common trends. *Journal of the American Statistical Association*, 83:pp. 1097–1107, 1988.

[60] R. Strachan and H. van Dijk. Valuing structure, model uncertainty and model averaging in vector autoregressive processes. *Tinbergen Institute Discussion Papers*, 2004.

[61] R. W. Strachan. Valid bayesian estimation of the cointegrating error correction model. *Journal of Business & Economic Statistics*, 2012.

[62] R. W. Strachan and H. K. Dijk. Bayesian model selection with an uninformative prior. *Oxford Bulletin of Economics and Statistics*, 65(s1):863–876, 2003.

[63] R. W. Strachan and B. Inder. Bayesian analysis of the error correction model. *Journal of Econometrics*, 123(2):307–325, 2004.

[64] K. Subbaraj and M. Dokainish. A survey of direct time-integration methods in computational structural dynamics. *Computers & Structures*, 32(6):1387–1401, 1989.

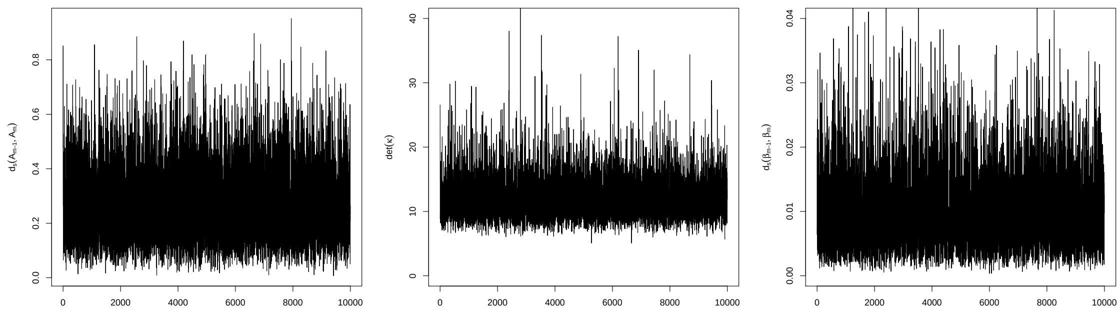
[65] K. Sugita. A Monte Carlo comparison of Bayesian testing for cointegration rank. *Economics Bulletin*, 29(3):2145–2151, 2009.

[66] K. Sugita. Testing for cointegration rank using bayes factors. Royal economic society annual conference 2002, Royal Economic Society, 2002.

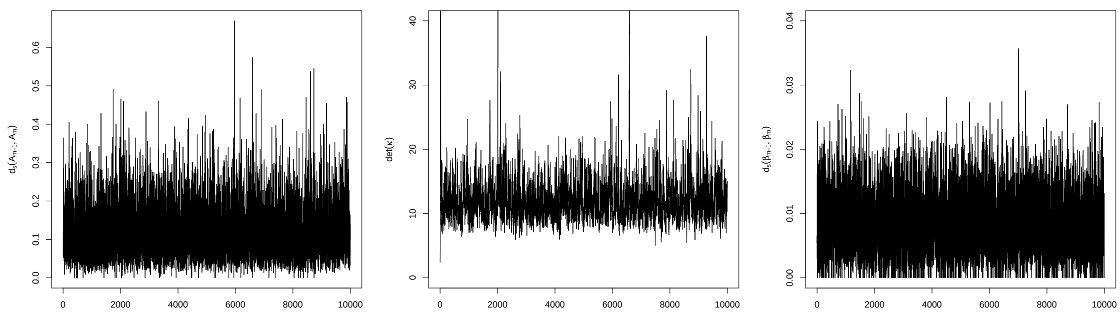
[67] D. Toczydlowska, G. W. Peters, M. C. Fung, and P. V. Shevchenko. Stochastic period and cohort effect state-space mortality models incorporating demographic factors via probabilistic robust principle components. *Risks*, 2017.

[68] D. A. Van Dyk and X. Jiao. Metropolis-hastings within partially collapsed gibbs samplers. *Journal of Computational and Graphical Statistics*, 24(2):301–327, 2015.

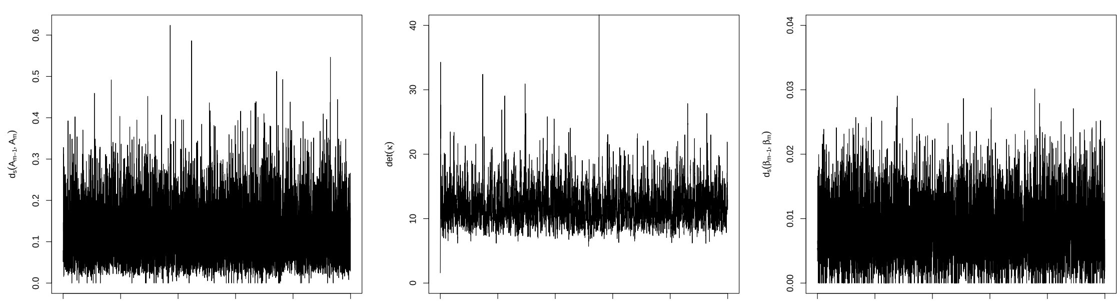
[69] L. Verlet. Computer "experiments" on classical fluids. thermodynamical properties of lennard-jones molecules. *Physical review*, 159(1):98, 1967.
[70] G. Vidyamurthy. *Pairs Trading: quantitative methods and analysis*, volume 217. John Wiley & Sons, 2004.
[71] M. Villani. Bayesian reference analysis of cointegration. *Econometric Theory*, 21(2):326–357, 2005.
[72] M. Villani. Bayesian point estimation of the cointegration space. *Journal of Econometrics*, 134(2):645–664, 2006.
[73] M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.
[74] R. Zhang, P. Robinson, and Q. Yao. Identifying cointegration by eigenanalysis. *arXiv preprint arXiv:1505.00821*, 2015.



(A) Gibbs sampler with data augmentation



(B) GMC for β and HMC for α with data augmentation (Gaussian likelihood in Algorithm 2)



(C) GMC for β and HMC for α for using Student-t likelihood in Algorithm 2

FIGURE 1. Trace plots from MCMC output for Student-t case: from left to right $d_s(\hat{\mathcal{A}}_m, \hat{\mathcal{A}}_{m-1})$,$|\hat{\kappa}_m|$, $d_s(\hat{\beta}_m, \hat{\beta}_{m-1})$ against MCMC iteration m.