econometrics

MDPI

Article

State-Space Models on the Stiefel Manifold with a New Approach to Nonlinear Filtering

Yukai Yang [1,2,*] and Luc Bauwens [3]

[1] Department of Statistics, Uppsala University, P.O. Box 513, SE-75120 Uppsala, Sweden
[2] Center for Data Analytics, Stockholm School of Economics, SE-11383 Stockholm, Sweden
[3] Center for Operations Research and Econometrics, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium; luc.bauwens@uclouvain.be
* Correspondence: yukai.yang@statistik.uu.se

Received: 30 July 2018; Accepted: 10 December 2018; Published: 12 December 2018

Abstract: We develop novel multivariate state-space models wherein the latent states evolve on the Stiefel manifold and follow a conditional matrix Langevin distribution. The latent states correspond to time-varying reduced rank parameter matrices, like the loadings in dynamic factor models and the parameters of cointegrating relations in vector error-correction models. The corresponding nonlinear filtering algorithms are developed and evaluated by means of simulation experiments.

Keywords: state-space models; Stiefel manifold; matrix Langevin distribution; filtering; smoothing; Laplace method; dynamic factor model; cointegration

JEL Classification: C32; C51

1. Introduction

The coefficient matrix of explanatory variables in multivariate time series models can be rank deficient due to some modelling assumptions, and the parameter constancy of the rank deficient matrix may be questionable. This may happen, for example, in the factor model, which construct very few factors by using a large number of macroeconomic and financial predictors, while the factor loadings are suspect to be time-varying. Stock and Watson (2002) state that it is reasonable to suspect temporal instability taking place in factor loadings, and later Stock and Watson (2009) and Breitung and Eickmeier (2011) find empirical evidence of instability. Another setting where instability may arise is in cointegrating relations (see e.g., Bierens and Martins (2010)), hence in the the reduced rank cointegrating parameter matrix of a vector error-correction model.

There are solutions in the literature to the modelling of the temporal instability of reduced rank parameter matrices. Such parameters are typically regarded as unobserved random components and in most cases are modelled as random walks on a Euclidean space; see, for example, Del Negro and Otrok (2008) and Eickmeier et al. (2014). In these works, the noise component of the latent processes (factor loading) is assumed to have a diagonal covariance matrix in order to alleviate the computational complexity and make the estimation feasible, especially when the dimension of the system is high. However, the random walk assumption on the Euclidean space cannot guarantee the orthonormality of the factor loading (or cointegration) matrix, while this type of assumption identifies the loading (or cointegration) space. Hence, other identification restrictions on the Euclidean space are needed. Moreover, the diagonality of the error covariance matrix of the latent processes contradicts itself when a permutation of the variables is performed.

In this work, we develop new state-space models on the Stiefel manifold, which do not suffer from the problems on the Euclidean space. It is noteworthy that Chikuse (2006) also develops state-space

models on the Stiefel manifold. The key difference between Chikuse (2006) and our work is that we keep the Euclidean space for the measurement evolution of the observable variables, while Chikuse (2006) puts them on the Stiefel manifold, which is not relevant for modelling economic time series. By specifying the time-varying reduced rank parameter matrices on the Stiefel manifold, their orthonormality is obtained by construction, and therefore their identification is guaranteed.

The corresponding recursive nonlinear filtering algorithms are developed to estimate the a posteriori distributions of the latent processes of the reduced rank matrices. By applying the matrix Langevin distribution on the a priori distributions of the latent processes, conjugate a posteriori distributions are achieved, which gives great convenience in the computational implementation of the filtering algorithms. The predictive step of the filtering requires solving an integral on the Stiefel manifold, which does not have a closed form. To compute this integral, we resort to a Laplace method.

The paper is organized as follows. Section 2 introduces the general framework of the vector models with time-varying reduced rank parameters. Two specific forms of the time-varying reduced rank parameters, which the paper is focused on, are given. Section 3 discusses some problems in the prevalent literature on modelling the time dependence of the time-varying reduced rank parameters, which underlie our modelling choices. Then, in Section 4, we present the novel state-space models on the Stiefel manifold. Section 5 presents the nonlinear filtering algorithms that we develop for the new state-space models. Section 6 presents several simulation based examples. Finally, Section 7 concludes and gives possible research extensions.

2. Vector Models with Time-Varying Reduced Rank Parameters

Consider the multivariate time series model with partly time-varying parameters

$$\boldsymbol{y}_t = \boldsymbol{A}_t \boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T, \tag{1}$$

where $\boldsymbol{y}_t$ is a (column) vector of dependent variables of dimension p, $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ are vectors of explanatory variables of dimensions q_1 and q_2, $\boldsymbol{A}_t$ and $\boldsymbol{B}$ are $p \times q_1$ and $p \times q_2$ matrices of parameters, and $\boldsymbol{\varepsilon}_t$ is a vector belonging to a white noise process of dimension p, with positive-definite covariance matrix $\boldsymbol{\Omega}$. For quasi-maximum likelihood estimation, we further assume that $\boldsymbol{\varepsilon}_t \sim N_p(\boldsymbol{0}, \boldsymbol{\Omega})$.

The distinction between $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ is introduced to separate the explanatory variables between those that have time-varying coefficients ($\boldsymbol{A}_t$) from those that have fixed coefficients ($\boldsymbol{B}$). In the sequel, we always consider that $\boldsymbol{x}_t$ is not void (i.e., $q_1 > 0$). The explanatory variables may contain lags of $\boldsymbol{y}_t$, and the remaining stochastic elements (if any) of these vectors are assumed to be weakly exogenous. Equation (1) provides a general linear framework for modelling time-series observations with time-varying parameters, embedding multivariate regressions and vector autoregressions. For an exposition of the treatment of such a model using the Kalman filter, we refer to Chapter 13 of Hamilton (1994).

We assume furthermore that the time-varying parameter matrix $\boldsymbol{A}_t$ has reduced rank $r < \min(p, q_1)$. This assumption can be formalized by decomposing $\boldsymbol{A}_t$ as $\boldsymbol{\alpha}_t \boldsymbol{\beta}_t'$, where $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are $p \times r$ and $q_1 \times r$ full rank matrices, respectively. If we allow both $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ to be time-varying, the model is not well focused and hard to explain, and its identification is very difficult. Hence, we focus on the cases where either $\boldsymbol{\alpha}_t$ or $\boldsymbol{\beta}_t$ is time-varying, that is, on the following two cases:

$$\text{Case 1:} \quad \boldsymbol{A}_t = \boldsymbol{\alpha}_t \boldsymbol{\beta}', \tag{2}$$

$$\text{Case 2:} \quad \boldsymbol{A}_t = \boldsymbol{\alpha} \boldsymbol{\beta}_t'. \tag{3}$$

Next, we explain how the two cases give interesting alternatives to modelling different kinds of temporal instability in parameters.

The case 1 model (Equations (1) and (2)) ensures that the subspace spanned by $\boldsymbol{\beta}$ is constant over time. This specification can be viewed as a cointegration model allowing for time-varying short-run adjustment coefficients (the entries of $\boldsymbol{\alpha}_t$) but with time-invariant long-run relations (cointegrating

subspace). To see this, consider that model (1) corresponds to a vector error-correction form of a cointegrated vector autoregressive model of order k with $\boldsymbol{X}_t$ as the dependent variables, if $\boldsymbol{y}_t = \Delta \boldsymbol{X}_t$, $\boldsymbol{x}_t = \boldsymbol{X}_{t-1}$, $\boldsymbol{z}_t$ contains $\Delta \boldsymbol{X}_{t-i}$ for $i = 1, \ldots, k-1$, as well as some predetermined variables. There are papers in the literature arguing that the temporal instability of the parameters in both stationary and non-stationary macroeconomic data does exist and cannot be overlooked. For example, Swanson (1998) and Rothman et al. (2001) give convincing examples in investigating the Granger causal relationship between money and output using a nonlinear vector error-correction model. They model the instability in $\boldsymbol{\alpha}$ by means of regime-switching mechanisms governed by some observable variable. An alternative to that modelling approach is to regard $\boldsymbol{\alpha}_t$ as a totally latent process.

The case 1 model also includes as a particular case the factor model with time-varying factor loadings. In the factor model context, the factors $\boldsymbol{f}_t$ are extracted from a number of observable predictors $\boldsymbol{x}_t$ by using the r linear combinations $\boldsymbol{f}_t = \boldsymbol{\beta}' \boldsymbol{x}_t$. Note that $\boldsymbol{f}_t$ is latent since $\boldsymbol{\beta}$ is unknown. Then, the corresponding factor model (neglecting the $\boldsymbol{B}\boldsymbol{z}_t$ term) takes the form

$$\boldsymbol{y}_t = \boldsymbol{\alpha}_t \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \tag{4}$$

where $\boldsymbol{\alpha}_t$ is a matrix of the time-varying factor loadings. The representation is quite flexible in the sense that $\boldsymbol{y}_t$ can be equal to $\boldsymbol{x}_t$ and then we reach exactly the same representation as Stock and Watson (2002), but we also allow them to be distinct. In Stock and Watson (2002), the factor loading matrix $\boldsymbol{\alpha}$ is time-invariant and the identification is obtained by imposing the constraints $q_1 \boldsymbol{\alpha} = \boldsymbol{\beta}$ and $\boldsymbol{\alpha}' \boldsymbol{\beta} = \boldsymbol{\beta}' \boldsymbol{\alpha} = \boldsymbol{\alpha}' \boldsymbol{\alpha} / q_1 = q_1 \boldsymbol{\beta}' \boldsymbol{\beta} = \boldsymbol{I}_r$. Notice that, if $\boldsymbol{\alpha}$ is time-varying but $\boldsymbol{\beta}$ time-invariant, these constraints cannot be imposed.

The case 2 model (Equations (1) and (3)) can be used to account for time-varying long-run relations in cointegrated time series, as $\boldsymbol{\beta}_t$ is changing. Bierens and Martins (2010) show that this may be the case for the long run purchasing power parity. In the case 2 model, there exist $p - r$ linearly independent vectors $\boldsymbol{\alpha}_\perp$ that span the left null space of $\boldsymbol{\alpha}$, such that $\boldsymbol{\alpha}_\perp' \boldsymbol{A}_t = \boldsymbol{0}$. Therefore, the case 2 model implies that the time-varying parameter matrix $\boldsymbol{\beta}_t$ vanishes in the structural vector model

$$\boldsymbol{\gamma}' \boldsymbol{y}_t = \boldsymbol{\gamma}' \boldsymbol{B} \boldsymbol{z}_t + \boldsymbol{\gamma}' \boldsymbol{\varepsilon}_t, \tag{5}$$

for any column vector $\boldsymbol{\gamma} \in \mathfrak{sp}(\boldsymbol{\alpha}_\perp)$, where $\mathfrak{sp}(\boldsymbol{\alpha}_\perp)$ denotes the space spanned by $\boldsymbol{\alpha}_\perp$, thus implying that the temporal instability can be removed in the above way. Moreover, $\boldsymbol{x}_t$ does not explain any variation of $\boldsymbol{\gamma}' \boldsymbol{y}_t$.

Another possible application for the case 2 model is the instability in the factor composition. Considering the factor model $\boldsymbol{y}_t = \boldsymbol{\alpha} \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t$, with time-invariant factor loading $\boldsymbol{\alpha}$, the factor composition may be slightly evolving through $\boldsymbol{\beta}_t$ in $\boldsymbol{f}_t = \boldsymbol{\beta}_t' \boldsymbol{x}_t$.

3. Issues about the Specification of the Time-Varying Reduced Rank Parameter

In the previous section, we have introduced two models with time-varying reduced rank parameters. In this section, in order to motivate our choices presented in Section 4, we discuss the specification in the literature of the dynamic process governing the evolution of the time-varying parameters.

Since the sequences $\boldsymbol{\alpha}_t$ or $\boldsymbol{\beta}_t$ in the two cases are unobservable in practice, it is quite natural to write the two models into the state-space form with a measurement equation like (1) for the observable variables and transition equations for $\boldsymbol{\alpha}_t$ or $\boldsymbol{\beta}_t$. To build the time dependence in the sequences of $\boldsymbol{\alpha}_t$ or $\boldsymbol{\beta}_t$ is of great practical interest as it enables one to use the historical time series data for conditional forecasting, especially by using the prevalent state-space model based approach. How to model the evolution of these time-varying parameters, nevertheless, is an open issue and needs careful investigation. Almost all the works in the literature of time series analysis hitherto only deal with state-space models on the Euclidean space. See, for example, the books by Hannan (1970); Anderson (1971); Koopman (1974); Durbin and Koopman (2012); and more recently Casals et al. (2016).

Consider, for example, the factor model (4) with time-varying factor loading $\boldsymbol{\alpha}_t$, but notice that the following discussion can be easily adapted to the cointegration model, where only $\boldsymbol{\beta}_t$ is time-varying. The traditional state-space framework on the Euclidean space assumes that the elements of the time-varying matrix $\boldsymbol{\alpha}_t$ evolve like random walks on the Euclidean space, see for example Del Negro and Otrok (2008) and Eickmeier et al. (2014). That is,

$$\text{vec}(\boldsymbol{\alpha}_{t+1}) = \text{vec}(\boldsymbol{\alpha}_t) + \boldsymbol{\eta}_t, \tag{6}$$

where vec denotes the vectorization operator, and the sequence of $\boldsymbol{\eta}_t$ is assumed to be a Gaussian strong white noise process with constant positive definite covariance matrix $\boldsymbol{\Sigma}_\eta$. Thus, Labels (1) and (6) form a vector state-space model, and the Kalman filter technique can be applied for estimating $\boldsymbol{\alpha}_t$.

A first problem of the model (6) is that the latent random walk evolution on the Euclidean space is strange. Consider the special case $p = 2$ and $r = 1$: in Figure 1, points 1–3 are possible locations of the latent variable $\text{vec}(\boldsymbol{\alpha}_t) = (\alpha_{1t}, \alpha_{2t})'$. Suppose that the next state $\boldsymbol{\alpha}_{t+1}$ evolves as in (6) with a diagonal covariance matrix $\boldsymbol{\Sigma}_\eta$. The circles centered around points 1–3 are contour lines such that, say, almost all the probability mass lies inside the circles. The straight lines OA and OB are tangent lines to circle 1 with A and B the tangent points; the straight lines OC and OD are tangent lines to circle 2; and the straight lines OE and OF are tangent lines to circle 3. The angles between the tangent lines depend on the location of the points 1-2-3: generally, the more distant a point from the origin, the smaller the corresponding angle despite some special ellipses. The plot shows that the distributions of the next subspace based on the current point differ for different subspaces (angles for 3 and 2 smaller than the angle for 1); even for the same subspace (points 2 and 3), the distribution of the subspace is different (angle for 3 smaller than angle for 2).

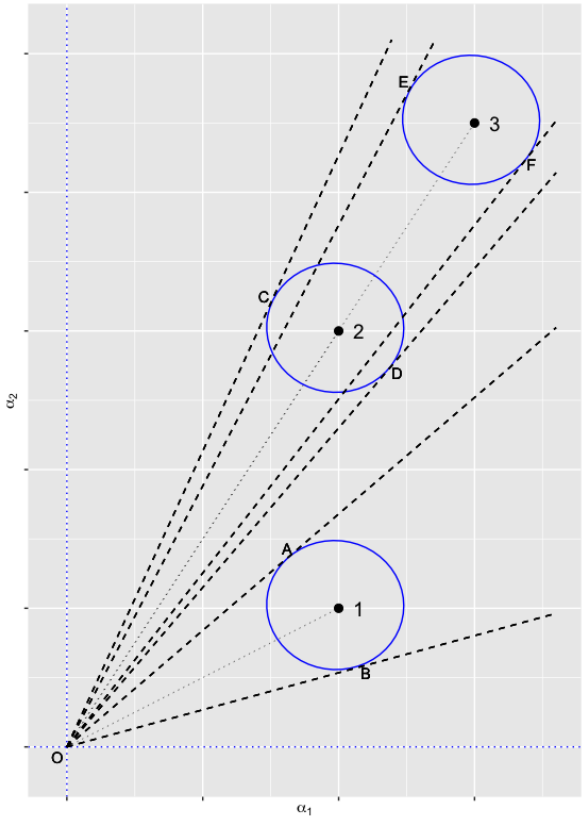

Figure 1. Euclidean state space for $p = 2$ and $r = 1$. Points 1–3 are possible locations of the latent variable $(\alpha_{1t}, \alpha_{2t})'$. Circles are isodensity contours assuming $(\alpha_{1,t+1}, \alpha_{2,t+1})'|(\alpha_{1t}, \alpha_{2t})' \sim N_2((\alpha_{1t}, \alpha_{2t})', I_2)$.

A second problem is the identification issue. The pair of $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}$ should be identified before we can proceed with the estimation of (1) and (6). If both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are time-invariant, it is common to assume the orthonormality (or asymptotic orthonormality) $\boldsymbol{\alpha}'\boldsymbol{\alpha}/q_1 = \boldsymbol{I}_r$ or $\boldsymbol{\alpha}'\boldsymbol{\alpha} = \boldsymbol{I}_r$ to identify the factors and then to estimate them by using the principle components method. However, when $\boldsymbol{\alpha}_t$ is evolving as (6), the orthonormality of $\boldsymbol{\alpha}_t$ can never be guaranteed for all t on the Euclidean space.

The alternative solution to the identification problem is to normalize the time-invariant part $\boldsymbol{\beta}$ as $(\boldsymbol{I}_r, \boldsymbol{b}')'$. The normalization is valid when the upper block of $\boldsymbol{\beta}$ is invertible, but if the upper block of $\boldsymbol{\beta}$ is not invertible, one can always permute the rows of $\boldsymbol{\beta}$ to find an invertible submatrix of order r rows for such a normalization. The permutation can be performed by left-multiplying $\boldsymbol{\beta}$ by a permutation matrix $\boldsymbol{P}$ to make its upper block invertible. In practice, it should be noted that the choice of the permutation matrix $\boldsymbol{P}$ is usually arbitrary and casual.

Even though the model defined by (1) and (6) is identified by some normalized $\boldsymbol{\beta}$, if one does not impose any constraint on the elements of the positive definite covariance matrix $\boldsymbol{\Sigma}_\eta$, the estimation can be very difficult due to computational complexity. A feasible solution is to assume that $\boldsymbol{\eta}_t$ is cross-sectionally uncorrelated. This restriction reduces the number of parameters, alleviates the complexity of the model, and makes the estimation much more efficient, but it may be too strong and imposes a priori information on the data. However, a third problem then arises. In the following two propositions, we show that any design like (1) and (6) with the restriction that $\boldsymbol{\Sigma}_\eta$ is diagonal is casual in the sense that it may lead to contradiction since the normalization of $\boldsymbol{\beta}$ is arbitrarily chosen.

Proposition 1. *Suppose that the reduced rank coefficient matrix $\boldsymbol{A}_t$ in (1) with rank r has the decomposition (2). By choosing some permutation matrix $\boldsymbol{P}_\beta$ $(p \times p)$, the time-invariant component $\boldsymbol{\beta}$ can be linearly normalized if the $r \times r$ upper block $\boldsymbol{b}_1$ in*

$$\boldsymbol{P}_\beta \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} \tag{7}$$

is invertible. Then, the corresponding linear normalization is

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{P}_\beta \boldsymbol{\beta} \boldsymbol{b}_1^{-1} = \begin{pmatrix} \boldsymbol{I}_r \\ \boldsymbol{b}_2 \boldsymbol{b}_1^{-1} \end{pmatrix}, \tag{8}$$

and the time-varying component is re-identified as $\tilde{\boldsymbol{\alpha}}_t = \boldsymbol{\alpha}_t \boldsymbol{b}_1'$.

Assuming that the time-varying component evolves by following

$$vec(\tilde{\boldsymbol{\alpha}}_{t+1}) = vec(\tilde{\boldsymbol{\alpha}}_t) + \boldsymbol{\eta}_t^\alpha. \tag{9}$$

Consider another permutation $\boldsymbol{P}_\beta^* \neq \boldsymbol{P}_\beta$ with the corresponding $\tilde{\boldsymbol{\alpha}}_t^*$, $\tilde{\boldsymbol{\beta}}^*$, $\boldsymbol{b}_1^*$ and $\boldsymbol{\eta}_t^{\alpha *}$. The variance–covariance matrices of $\boldsymbol{\eta}_t^\alpha$ and $\boldsymbol{\eta}_t^{\alpha *}$ are both diagonal if and only if $\boldsymbol{b}_1 = \boldsymbol{b}_1^*$.

Proof. See Appendix A. □

Proposition 2. *Suppose that the reduced rank coefficient matrix $\boldsymbol{A}_t$ in (1) with rank r has the decomposition (3). By choosing some permutation matrix $\boldsymbol{P}_\alpha$ $(p \times p)$, the constant component $\boldsymbol{\alpha}$ can be linearly normalized if the $r \times r$ upper block $\boldsymbol{a}_1$ in*

$$\boldsymbol{P}_\alpha \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{pmatrix} \tag{10}$$

is invertible. The corresponding linear normalization is

$$\tilde{\boldsymbol{\alpha}} = \boldsymbol{P}_\alpha \boldsymbol{\alpha} \boldsymbol{a}_1^{-1} = \begin{pmatrix} \boldsymbol{I}_r \\ \boldsymbol{a}_2 \boldsymbol{a}_1^{-1} \end{pmatrix}, \tag{11}$$

and the time-varying component is re-identified as $\tilde{\boldsymbol{\beta}}_t = \boldsymbol{\beta}_t \boldsymbol{a}_1'$. *Assuming that the time-varying component evolves by following*

$$vec(\tilde{\boldsymbol{\beta}}_{t+1}) = vec(\tilde{\boldsymbol{\beta}}_t) + \boldsymbol{\eta}_t^{\beta}. \tag{12}$$

Consider another permutation $\boldsymbol{P}_\alpha^* \neq \boldsymbol{P}_\alpha$ *with the corresponding* $\tilde{\boldsymbol{\alpha}}^*$, $\tilde{\boldsymbol{\beta}}_t^*$, $\boldsymbol{a}_1^*$ *and* $\boldsymbol{\eta}_t^{\beta *}$. *The variance–covariance matrices of* $\boldsymbol{\eta}_t^{\beta}$ *and* $\boldsymbol{\eta}_t^{\beta *}$ *are both diagonal if and only if* $\boldsymbol{a}_1 = \boldsymbol{a}_1^*$.

Proof. See Appendix B. □

The two corollaries below follow Propositions 1 and 2 immediately, showing that the assumption that the variance–covariance matrix $\boldsymbol{\Sigma}_\eta$ is always diagonal for any linear normalization is inappropriate.

Corollary 1. *Given the settings in Propostion 1, the variance–covariance matrices of the error vectors in forms like (9) based on different linear normalizations cannot be both diagonal if* $\boldsymbol{b}_1 \neq \boldsymbol{b}_1^*$ *where* $\boldsymbol{b}_1$ *and* $\boldsymbol{b}_1^*$ *are the upper block square matrices in forms like (7).*

Corollary 2. *Given the settings in Proposition 2, the variance–covariance matrices of the error vectors in forms like (12) based on different linear normalizations cannot be both diagonal if* $\boldsymbol{a}_1 \neq \boldsymbol{a}_1^*$ *where* $\boldsymbol{a}_1$ *and* $\boldsymbol{a}_1^*$ *are the upper block square matrices in forms like (10).*

One may argue that there is a chance for the two covariance matrices to be both diagonal, i.e., when $\boldsymbol{b}_1 = \boldsymbol{b}_1^*$. It should be noticed that the condition $\boldsymbol{b}_1 = \boldsymbol{b}_1^*$ does not imply that $\boldsymbol{P} = \boldsymbol{P}^*$. Instead, it implies that the permutation matrices move the same variables to the upper part of $\boldsymbol{\beta}$ with the same order. If this is the case, the two permutation matrices $\boldsymbol{P}$ and $\boldsymbol{P}^*$ are distinct but equivalent as the order of the variables in the lower part is trivial for linear normalization.

Since the choice of the permutation $\boldsymbol{P}$ and the corresponding linear normalization is arbitrary in practice, which is simply the order of $\boldsymbol{x}_t$ ($\boldsymbol{y}_t$ for case 2), the models with different $\boldsymbol{P}$ are telling different stories about the data. In fact, the model has been over-identified by the assumption that $\boldsymbol{\Sigma}_\eta$ must be diagonal. Consequently, the model becomes $\boldsymbol{\beta}$-normalization dependent, and the $\boldsymbol{\beta}$-normalization imposes some additional information on the data. This can be serious when the forecasts from the models with distinct normalizations of $\boldsymbol{\alpha}$ give totally different results. A solution to this "unexpected" problem may be to try all possible normalizations of $\boldsymbol{\alpha}$ and do model selection, that is, after estimating every possible model, pick the best model according to an information criterion. However, this solution is not always feasible because the number of possible permutations for $\boldsymbol{\alpha}$, which is equal to $q_1(q_1 - 1)\dots(q_1 - r + 1)$, can be huge. When the number of predictors is large, which is common in practice, the estimation of each possible model based on different normalization becomes a very demanding task.

Stock and Watson (2002) propose the assumption that the cross-sectional dependence between the elements in $\boldsymbol{\eta}_t$ is weak and the variances of the elements are shrinking with the increase of the sample size. Then, the aforementioned problem may not be so serious, as, intuitively, different normalizations with diagonal covariance matrix $\boldsymbol{\Sigma}_\eta$ may produce approximately or asymptotically the same results.

We have shown that the modelling of the time-varying parameter matrix in (2) as a process like (6) on the Euclidean space involves some problems. Firstly, the evolution of the subspace spanned by the latent process on the Euclidean space is strange. Secondly, the process does not comply with the orthonormality assumption to identify the pair of $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}$. Thus, a linear normalization is employed instead of the orthonormality. Thirdly, the state-space model on the Euclidean space suffers from the curse of dimensionality, and hence the diagonality of the covariance of the errors is often used with the linear normalization in order to alleviate the computational complexity when the dimension is high. This leads to two other problems: firstly, the diagonality assumption is inappropriate in the sense that different linear normalizations may lead to a contradiction; secondly, the model selection can be a tremendous task when there are many predictors.

In the following section, we propose that the time-varying parameter matrices $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ evolve on the Stiefel manifold, instead of the Euclidean space, and we show that the corresponding state-space models do not suffer from the aforementioned problems.

4. State-Space Models on the Stiefel Manifold

4.1. The Stiefel Manifold and the Matrix Langevin Distribution

Before presenting the state-space models on the Stiefel manifold, we introduce some concepts and terms. The *Stiefel manifold* $\mathbb{V}_{a,b}$, for dimensions a and b such that $a \geq b$, is a space whose points are *b-frames* in $\mathbb{R}^a$. A set of b orthonormal vectors in $\mathbb{R}^a$ is called a b-frame in $\mathbb{R}^a$. The Stiefel manifold is a collection of $a \times b$ full rank matrices $\boldsymbol{X}$ such that $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_b$; if $b = 1$, the Stiefel manifold is the unit circle if $a = 2$, sphere if $a = 3$, and hypersphere if $a > 3$. The link with the modelling presented in Section 2 and developed in the next subsection is that the time-varying matrix $\boldsymbol{\alpha}_t$ of (2) is assumed to be evolving in $\mathbb{V}_{p,r}$ (instead of a Euclidean space), and $\boldsymbol{\beta}_t$ of (3) in $\mathbb{V}_{q_1,r}$. Hence, each $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ is by definition orthonormal.

We also need to replace the assumption (6) that the distribution of $\text{vec}(\boldsymbol{\alpha}_{t+1})$ conditional on $\text{vec}(\boldsymbol{\alpha}_t)$ is $N_{p\times r}(\text{vec}(\boldsymbol{\alpha}_t), \boldsymbol{\Sigma}_\eta)$ by an appropriate distribution defined on $\mathbb{V}_{p,r}$, and likewise for $\text{vec}(\boldsymbol{\beta}_{t+1})$. A convenient distribution for this purpose is the *matrix Langevin* distribution (also known as *von Mises–Fisher distribution*) denoted by $ML(a, b, \boldsymbol{F})$. A random matrix $\boldsymbol{X} \in \mathbb{V}_{a,b}$ follows a matrix Langevin distribution if and only if it has the probability density function

$$f_{ML}(\boldsymbol{X}|a, b, \boldsymbol{F}) = \frac{\text{etr}\{\boldsymbol{F}'\boldsymbol{X}\}}{{}_0F_1(\frac{a}{2}; \frac{1}{4}\boldsymbol{F}'\boldsymbol{F})}, \tag{13}$$

where $\text{etr}\{\boldsymbol{Q}\}$ stands for $\exp\{\text{tr}\{\boldsymbol{Q}\}\}$ for any full rank square matrix $\boldsymbol{Q}$, $\boldsymbol{F}$ is a $a \times b$ matrix, and ${}_0F_1(a/2; \boldsymbol{F}'\boldsymbol{F}/4)$ is called $(0,1)$-type hypergeometric function with arguments $a/2$ and $\boldsymbol{F}'\boldsymbol{F}/4$. The hypergeometric function ${}_0F_1$ is unusual due to a matrix argument, see Herz (1955), and it is actually the normalizing constant of the density defined in (13), that is,

$${}_0F_1\left(\frac{a}{2}; \frac{1}{4}\boldsymbol{F}'\boldsymbol{F}\right) = \int \text{etr}\left\{\boldsymbol{F}'\boldsymbol{X}\right\} [\,\mathrm{d}\boldsymbol{X}], \tag{14}$$

where $[\,\mathrm{d}\boldsymbol{X}] = \wedge_{j=1}^{a-b} \wedge_{i=1}^{b} \boldsymbol{x}'_{b+j}\,\mathrm{d}\boldsymbol{x}_i \wedge_{i<j} \boldsymbol{x}'_j\,\mathrm{d}\boldsymbol{x}_i$, stands for the differential form of a Haar measure on the Stiefel manifold, $\boldsymbol{x}_i$ is a column vector of $\boldsymbol{X}$, and $\wedge$ is the exterior product of vectors.

The density function (13) is obtained from a normal density for a random matrix $\boldsymbol{Z}$ of dimension $a \times b$, defined as $vec(\boldsymbol{Z}) \sim N_{a\times b}(vec(\boldsymbol{M}), \boldsymbol{I}_a \otimes \boldsymbol{\Sigma})$ (where $\boldsymbol{M}$ is a matrix of dimension $a \times b$, and $\boldsymbol{\Sigma}$ is a positive definite matrix of dimension $b \times b$) by imposing that $\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{I}_b$. The parameter $\boldsymbol{F}$ of (13) is then equal to $\boldsymbol{M}\boldsymbol{\Sigma}^{-1}$.

The matrix $\boldsymbol{F}$ has a singular value decomposition $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$, where $\boldsymbol{U} \in \mathbb{V}_{a,b}$, $\boldsymbol{V}$ is a $b \times b$ orthogonal matrix, and $\boldsymbol{D} = \text{diag}\{d_1, d_2, \ldots, d_b\}$ is a diagonal matrix with singular values $d_1 \geq d_2 \ldots \geq d_b \geq 0$. Each pair of the column vectors in $\boldsymbol{U}$ and $\boldsymbol{V}$ corresponds to a singular value in $\boldsymbol{D}$. Notice that the hypergeometric function in (13) has the property that

$${}_0F_1\left(\frac{a}{2}; \frac{1}{4}\boldsymbol{F}'\boldsymbol{F}\right) = {}_0F_1\left(\frac{a}{2}; \frac{1}{4}\boldsymbol{D}^2\right), \tag{15}$$

see Khatri and Mardia (1977).

It can be shown that the density function (13) has maximum value $\exp(\sum_{i=1}^{b} d_i)$ at $\boldsymbol{X}_{\boldsymbol{m}} = \boldsymbol{U}\boldsymbol{V}'$, called the *modal orientation* of the matrix Langevin distribution. The mode is unique if $\min(d_i) > 0$. The diagonal matrix $\boldsymbol{D}$ is called *concentration* as it controls how tight the distribution is in the following sense: the larger d_i, the tighter the distribution is around the corresponding i-th column vector of the

modal orientation matrix. For more details about the matrix Langevin distribution, see, for example, Prentice (1982); Chikuse (2003); Khatri and Mardia (1977); and Mardia (1975).

The density function (13) is *rotationally symmetric* around $\boldsymbol{X_m}$, in the sense that the density at $\boldsymbol{H_1 X H_2'}$ is the same as that at $\boldsymbol{X}$ for all orthogonal matrices $\boldsymbol{H_1}$ (of dimension $a \times a$) and $\boldsymbol{H_2}$ (of dimension $b \times b$) such that $\boldsymbol{H_1 U} = \boldsymbol{U}$ and $\boldsymbol{H_2 V} = \boldsymbol{V}$ (hence $\boldsymbol{H_1 X}_m \boldsymbol{H_2'} = \boldsymbol{X}_m$).

Figure 2 illustrates the Stiefel manifold and Figure 3 three matrix Langevin (not normalized) densities $ML(2, 1, \boldsymbol{F})$ where $\boldsymbol{F} = \boldsymbol{U}DV' = (1/\sqrt{2}, 1/\sqrt{2})'D$, setting V (a scalar) equal to 1, for three values of D (a scalar); the smaller D, the flatter the density. In Figure 2, the modal orientation $\boldsymbol{U} = (1/\sqrt{2}, 1/\sqrt{2})'$ is shown for the densities of Figure 3, and the point at which the density values are minimal, this point being equal to $-\boldsymbol{U}$. The densities are shown on Figure 3 as functions of the angle θ shown on Figure 2, for θ between 0 and 2π, instead of being shown as lines above the unit circle. Rotational symmetry in this example means that, if we premultiply the random vector $\boldsymbol{X}$ by any orthogonal 2×2 matrix $\boldsymbol{H_1}$ that does not modify the modal orientation, the densities are not changed.

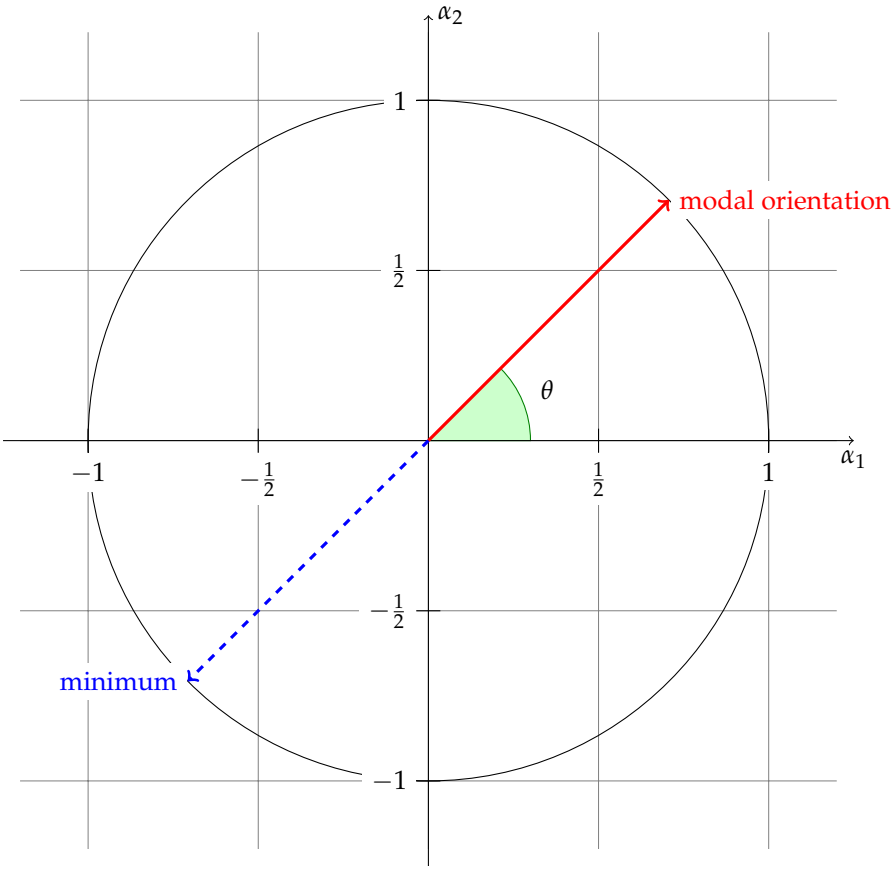

Figure 2. Stiefel manifold for $p = 2$ and $r = 1$.

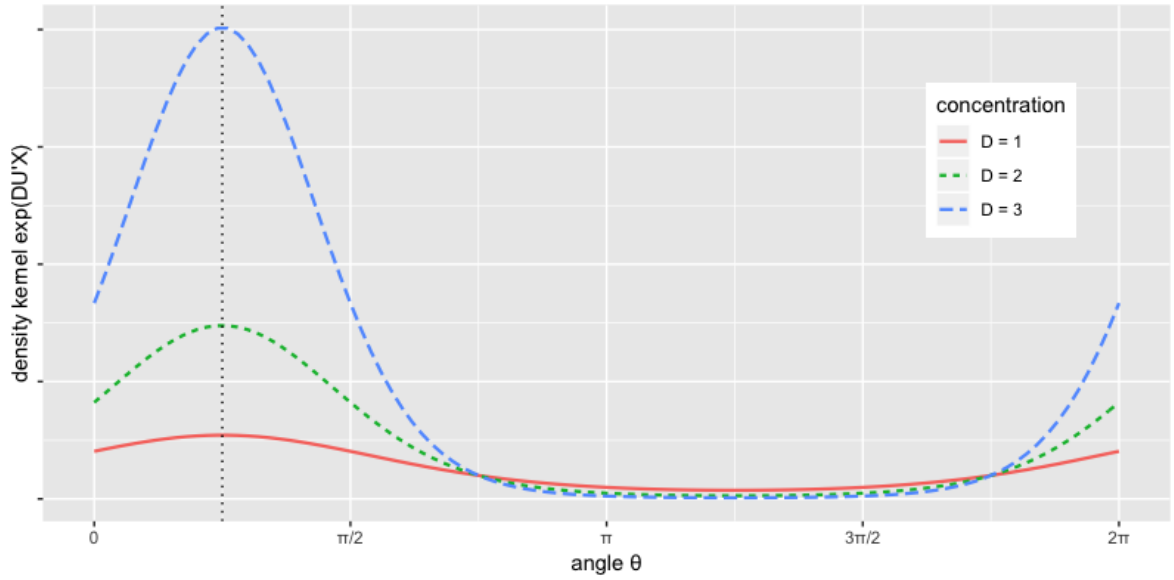

Figure 3. Matrix Langevin density kernels for $p = 2$ and $r = 1$.

4.2. Models

Chikuse (2006) develops a state-space model whose observable and latent variables are both evolving on Stiefel manifolds. For economic data, it is not appropriate to assume that the observable variables evolve on a Stiefel manifold, so that we keep the assumption that $\boldsymbol{y}_t$ evolves on a Euclidean space in the measurement Equation (1).

We define two state space models corresponding to the case 1 and case 2 models introduced in Section 2, with latent processes evolving over the Stiefel manifold and following conditional matrix Langevin distributions:

$$\begin{aligned} \text{Model 1:} \quad \boldsymbol{y}_t &= \boldsymbol{\alpha}_t\boldsymbol{\beta}'\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{\varepsilon}_t, \\ \boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t &\sim ML(p, r, \boldsymbol{a}_t\boldsymbol{D}\boldsymbol{V}'), \end{aligned} \tag{16}$$

$$\begin{aligned} \text{Model 2:} \quad \boldsymbol{y}_t &= \boldsymbol{\alpha}\boldsymbol{\beta}_t'\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{\varepsilon}_t, \\ \boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t &\sim ML(q_1, r, \boldsymbol{b}_t\boldsymbol{D}\boldsymbol{V}'), \end{aligned} \tag{17}$$

with the constraints that $\boldsymbol{a}_t\boldsymbol{V}' = \boldsymbol{\alpha}_t$ and $\boldsymbol{b}_t\boldsymbol{V}' = \boldsymbol{\beta}_t$, respectively. We assume in addition that the error $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\alpha}_{t+1}$ or $\boldsymbol{\beta}_{t+1}$ are mutually independent. The parameters of the ML distributions of the models are chosen so that the previous state of $\boldsymbol{\alpha}_t$ or $\boldsymbol{\beta}_t$ is the modal orientation of the next state. Thus, the transitions of the latent processes are random walks on the Stiefel manifold and evolve in the matrix Langevin way.

The models (16) and (17) are not yet identified due to the fact that the pairs between $\boldsymbol{a}_t$ or $\boldsymbol{b}_t$ and the nuisance parameter $\boldsymbol{V}$ can be arbitrarily chosen, and therefore the time-invariant $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are not identified as well. The identification problem can be solved by imposing $\boldsymbol{V} = \boldsymbol{I}_r$. Then, the identified version of the models is

$$\begin{aligned} \text{Model 1:} \quad \boldsymbol{y}_t &= \boldsymbol{\alpha}_t\boldsymbol{\beta}'\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{\varepsilon}_t, \\ \boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t &\sim ML(p, r, \boldsymbol{\alpha}_t\boldsymbol{D}), \end{aligned} \tag{18}$$

$$\begin{aligned} \text{Model 2:} \quad \boldsymbol{y}_t &= \boldsymbol{\alpha}\boldsymbol{\beta}_t'\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{\varepsilon}_t, \\ \boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t &\sim ML(q_1, r, \boldsymbol{\beta}_t\boldsymbol{D}). \end{aligned} \tag{19}$$

The new state-space models in (18) and (19) do not have the problems mentioned in Section 3, due to the fact that both $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are points in the Stiefel manifold. By construction, orthonormality is ensured, which is $\boldsymbol{\alpha}_t'\boldsymbol{\alpha}_t = I_r$ for Model 1, and similarly $\boldsymbol{\beta}_t'\boldsymbol{\beta}_t = I_r$ for Model 2. If the space spanned by the columns of $\boldsymbol{\alpha}_t$ (or the columns of $\boldsymbol{\beta}_t$) is subjected to a rotation, the model is fundamentally unchanged. Indeed, in the case of Model 1, let $\boldsymbol{H}$ be an orthogonal matrix ($p \times p$), and define the rotation $\tilde{\boldsymbol{\alpha}}_t = \boldsymbol{H}\boldsymbol{\alpha}_t$. Then, $\tilde{\boldsymbol{\alpha}}_t'\tilde{\boldsymbol{\alpha}}_t = \boldsymbol{\alpha}_t'H'H\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_t'\boldsymbol{\alpha}_t = \boldsymbol{I}_r$. A similar reasoning holds for Model 2.

More simple versions of the models in (18) and (19) are obtained by assuming that the evolutions of $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are independent of their previous states, with the same modal orientations $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ across time:

$$\begin{aligned} \text{Model } 1^*: \quad \boldsymbol{y}_t &= \boldsymbol{\alpha}_t\boldsymbol{\beta}'\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{\varepsilon}_t, \\ \boldsymbol{\alpha}_t &\sim ML(p, r, \boldsymbol{\alpha}_0\boldsymbol{D}), \end{aligned} \tag{20}$$

$$\begin{aligned} \text{Model } 2^*: \quad \boldsymbol{y}_t &= \boldsymbol{\alpha}\boldsymbol{\beta}_t'\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{z}_t + \boldsymbol{\varepsilon}_t, \\ \boldsymbol{\beta}_t &\sim ML(q_1, r, \boldsymbol{\beta}_0\boldsymbol{D}). \end{aligned} \tag{21}$$

If we assume that the random variation of $\boldsymbol{\alpha}_{t+1}$ in (18) or $\boldsymbol{\beta}_{t+1}$ in (19) are inside the subspace spanned by $\boldsymbol{\alpha}_t$ or $\boldsymbol{\beta}_t$ (hence $\boldsymbol{\alpha}_0$ or $\boldsymbol{\beta}_0$), then we have another two state space models. The corresponding conditional distributions of $\boldsymbol{\alpha}_{t+1}$ and $\boldsymbol{\beta}_{t+1}$ become truncated matrix Langevin distributions with the density functions:

$$f(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t) \quad \begin{cases} \propto \operatorname{etr}\left\{\boldsymbol{D}\boldsymbol{\alpha}_t'\boldsymbol{\alpha}_{t+1}\right\}, & \text{if} \quad \mathfrak{sp}(\boldsymbol{\alpha}_{t+1}) = \mathfrak{sp}(\boldsymbol{\alpha}_t) \quad \text{or} \quad \mathfrak{sp}(\boldsymbol{\alpha}_0) \\ = 0, & \text{otherwise.} \end{cases} \tag{22}$$

$$f(\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t) \quad \begin{cases} \propto \operatorname{etr}\left\{\boldsymbol{D}\boldsymbol{\beta}_t'\boldsymbol{\beta}_{t+1}\right\}, & \text{if} \quad \mathfrak{sp}(\boldsymbol{\beta}_{t+1}) = \mathfrak{sp}(\boldsymbol{\beta}_t) \quad \text{or} \quad \mathfrak{sp}(\boldsymbol{\beta}_0) \\ = 0, & \text{otherwise.} \end{cases} \tag{23}$$

These two models can be interesting if the spaces spanned by the time-varying $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are expected to be invariant over time.

Denote $\boldsymbol{\Delta} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T)$ in Model 1 or $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T)$ in Model 2; and let $\mathcal{F}_{t-1} = (\boldsymbol{x}_1, \boldsymbol{z}_1, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{x}_t, \boldsymbol{z}_t)$ represent all the observable information up to time $t-1$, such that $\mathrm{E}(\boldsymbol{y}_t|\mathcal{F}_{t-1}) = \boldsymbol{A}_t'\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{z}_t$; and let $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$.

The quasi-likelihood function for Model 1 based on Gaussian errors takes the form

$$f(\boldsymbol{Y}, \boldsymbol{\Delta}|\boldsymbol{\theta}) = \prod_{t=1}^{T}(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Omega}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\varepsilon}_t'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon}_t\right\} \frac{\operatorname{etr}\left\{\boldsymbol{D}\boldsymbol{\alpha}_{t-1}'\boldsymbol{\alpha}_t\right\}}{{}_0F_1(\frac{p}{2}; \frac{1}{4}\boldsymbol{D}^2)}, \tag{24}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{\Omega}, \boldsymbol{D}, \boldsymbol{\alpha}_0)$, $\boldsymbol{\varepsilon}_t = \boldsymbol{y}_t - \boldsymbol{\alpha}_t\boldsymbol{\beta}'\boldsymbol{x}_t - \boldsymbol{B}\boldsymbol{z}_t$.

The quasi-likelihood function for Model 2 based on Gaussian errors takes the form

$$f(\boldsymbol{Y}, \boldsymbol{\Delta}|\boldsymbol{\theta}) = \prod_{t=1}^{T}(2\pi)^{-\frac{p}{2}}|\boldsymbol{\Omega}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\varepsilon}_t'\boldsymbol{\Omega}^{-1}\boldsymbol{\varepsilon}_t\right\} \frac{\operatorname{etr}\left\{\boldsymbol{D}\boldsymbol{\beta}_{t-1}'\boldsymbol{\beta}_t\right\}}{{}_0F_1(\frac{p}{2}; \frac{1}{4}\boldsymbol{D}^2)}, \tag{25}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{B}, \boldsymbol{\Omega}, \boldsymbol{D}, \boldsymbol{\beta}_0)$, $\boldsymbol{\varepsilon}_t = \boldsymbol{y}_t - \boldsymbol{\alpha}\boldsymbol{\beta}_t'\boldsymbol{x}_t - \boldsymbol{B}\boldsymbol{z}_t$.

We treat the initial values $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ as the parameters to be estimated, but of course they can be regarded as given.

5. The Filtering Algorithms

In this section, for the models (18) and (19) defined in the previous section, we propose nonlinear filtering algorithms to estimate the a posteriori distributions of the latent processes based on the Gaussian error assumption in the measurement equations.

We start with Model 1 which has time-varying $\boldsymbol{\alpha}_t$. The filtering algorithm consists of two steps:

$$\text{Predict}: \quad f(\boldsymbol{\alpha}_t|\mathcal{F}_{t-1}) = \int f(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}) f(\boldsymbol{\alpha}_{t-1}|\mathcal{F}_{t-1})[\mathrm{d}\boldsymbol{\alpha}_{t-1}], \tag{26}$$

$$\text{Update}: \quad f(\boldsymbol{\alpha}_t|\mathcal{F}_t) \propto f(\boldsymbol{y}_t|\boldsymbol{\alpha}_t, \mathcal{F}_{t-1}) f(\boldsymbol{\alpha}_t|\mathcal{F}_{t-1}), \tag{27}$$

where the symbol $[\mathrm{d}\boldsymbol{\alpha}_{t-1}]$ stands for the differential form for a Haar measure on the Stiefel manifold. The predictive density in (26) represents the a priori distribution of the latent variable before observing the information at time t. The updating density, which is also called the filtering density, represents the a posteriori distribution of the latent variable after observing the information at time t.

The prediction step is quite tricky in the sense that, even if we can find the joint distribution of $\boldsymbol{\alpha}_t$ and $\boldsymbol{\alpha}_{t-1}$, which is the product $f(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}) f(\boldsymbol{\alpha}_{t-1}|\mathcal{F}_{t-1})$, we must integrate out $\boldsymbol{\alpha}_{t-1}$ over the Stiefel manifold. The density kernel $f(\boldsymbol{\alpha}_{t-1}|\mathcal{F}_{t-1})$ appearing in the integral in the first line of (27) comes from the previous updating step and is quite straightforward as it is proportional to the product of the density function of $\boldsymbol{y}_{t-1}$ and the predicted density of $\boldsymbol{\alpha}_{t-1}$ (see the updating step in (27)).

The initial condition for the filtering algorithm can be a Dirac delta function $f(\boldsymbol{\alpha}_0|\mathcal{F}_0)$ such that $f(\boldsymbol{\alpha}_0|\mathcal{F}_0) = \infty$ when $\boldsymbol{\alpha}_0 = \boldsymbol{U}_0$ where $\boldsymbol{U}_0$ is the modal orientation and zero otherwise, but the integral $\int f(\boldsymbol{\alpha}_0|\mathcal{F}_0)[\mathrm{d}\boldsymbol{\alpha}_0]$ is exactly equal to one.

The corresponding nonlinear filtering algorithm is recursive like the Kalman filter in linear dynamic systems. We start the algorithm with

$$f(\boldsymbol{\alpha}_1|\mathcal{F}_0) \propto \mathrm{etr}\{\boldsymbol{D}\boldsymbol{U}_0'\boldsymbol{\alpha}_1\}, \tag{28}$$

and proceed to the updating step for $\boldsymbol{\alpha}_1$ as follows:

$$f(\boldsymbol{\alpha}_1|\mathcal{F}_1) \propto \mathrm{etr}\{\boldsymbol{H}_1\boldsymbol{\alpha}_1'\boldsymbol{J}\boldsymbol{\alpha}_1 + \boldsymbol{C}_1'\boldsymbol{\alpha}_1\}, \tag{29}$$

where $\boldsymbol{H}_1 = -\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{x}_1\boldsymbol{x}_1'\boldsymbol{\beta}$, $\boldsymbol{J} = \boldsymbol{\Omega}^{-1}$, $\boldsymbol{C}_1 = \boldsymbol{U}_0\boldsymbol{D} + \boldsymbol{\Omega}^{-1}(\boldsymbol{y}_1 - \boldsymbol{B}\boldsymbol{z}_1)\boldsymbol{x}_1'\boldsymbol{\beta}$. Then, we move to the prediction step for $\boldsymbol{\alpha}_2$ and obtain the integral as follows:

$$f(\boldsymbol{\alpha}_2|\mathcal{F}_1) = \int f(\boldsymbol{\alpha}_2|\boldsymbol{\alpha}_1) f(\boldsymbol{\alpha}_1|\mathcal{F}_1)[\mathrm{d}\boldsymbol{\alpha}_1], \tag{30}$$

where

$$f(\boldsymbol{\alpha}_2|\boldsymbol{\alpha}_1) = \frac{\mathrm{etr}\{\boldsymbol{D}\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2\}}{{}_0F_1(\frac{a}{2}; \frac{1}{4}\boldsymbol{D}^2)}, \tag{31}$$

due to (13) and (15), and $f(\boldsymbol{\alpha}_1|\mathcal{F}_1)$ in (29). Hence, we have

$$f(\boldsymbol{\alpha}_2|\boldsymbol{\alpha}_1) f(\boldsymbol{\alpha}_1|\mathcal{F}_1) = \xi \cdot \mathrm{etr}\{\boldsymbol{D}\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2\} \cdot \mathrm{etr}\{\boldsymbol{H}\boldsymbol{\alpha}_1'\boldsymbol{J}\boldsymbol{\alpha}_1 + \boldsymbol{C}_1'\boldsymbol{\alpha}_1\}, \tag{32}$$

where ξ does not depend on $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. Unfortunately, there is no closed form solution to the integral (30) in the literature.

Another contribution of this paper is that we propose to approximate this integral by using the Laplace method. (see Wong (1989, chps. 2 and 9) for a detailed exposition). Rewrite the integral (30) as

$$f(\boldsymbol{\alpha}_2|\mathcal{F}_1) = \xi \int h(\boldsymbol{\alpha}_1) \exp\{p \cdot g(\boldsymbol{\alpha}_1)\}[\mathrm{d}\boldsymbol{\alpha}_1], \tag{33}$$

where p is the dimension of $\boldsymbol{y}_t$,

$$h(\boldsymbol{\alpha}_1) = \mathrm{etr}\{\boldsymbol{D}\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_2\} \leq \exp\left\{\sum_{i=1}^{r} d_i\right\}, \tag{34}$$

is bounded, and

$$g(\boldsymbol{\alpha}_1) = \text{tr}\{\boldsymbol{H}_1\boldsymbol{\alpha}_1'\boldsymbol{J}\boldsymbol{\alpha}_1 + \boldsymbol{C}_1'\boldsymbol{\alpha}_1\}/p, \tag{35}$$

which is twice differentiable with respect to $\boldsymbol{\alpha}_1$ and is assumed to be convergent to some nonzero value when $p \to \infty$.

Then, the Laplace method can be applied since the Taylor expansion on which it is based is valid in the neighbourhood for any point on the Stiefel manifold. It follows that, with $p \to \infty$, the integral (30) can be approximated by

$$\begin{aligned} f(\boldsymbol{\alpha}_2|\mathcal{F}_1) &\approx \xi\, h(\boldsymbol{U}_1)\exp\{p\, g(\boldsymbol{U}_1)\}, \\ &\propto \text{etr}\{\boldsymbol{D}\boldsymbol{U}_1'\boldsymbol{\alpha}_2\} \end{aligned} \tag{36}$$

where

$$\boldsymbol{U}_1 = \underset{\boldsymbol{\alpha}_1 \in \mathbb{V}_{p,r}}{\arg\max}\ \text{etr}\{\boldsymbol{H}_1\boldsymbol{\alpha}_1'\boldsymbol{J}\boldsymbol{\alpha}_1 + \boldsymbol{C}_1'\boldsymbol{\alpha}_1\}. \tag{37}$$

Given $f(\boldsymbol{\alpha}_2|\mathcal{F}_1) \propto \text{etr}\{\boldsymbol{D}\boldsymbol{U}_1'\boldsymbol{\alpha}_2\}$, then it can be shown that $f(\boldsymbol{\alpha}_2|\mathcal{F}_2)$ has the same form as (29) with $\boldsymbol{H}_2 = -\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{x}_2\boldsymbol{x}_2'\boldsymbol{\beta}$, $\boldsymbol{C}_2 = \boldsymbol{U}_1\boldsymbol{D} + \boldsymbol{\Omega}^{-1}(\boldsymbol{y}_2 - \boldsymbol{B}\boldsymbol{z}_2)\boldsymbol{x}_2'\boldsymbol{\beta}$.

Thus, by induction, we have the following proposition for the recursive filtering algorithm for state-space Model 1.

Proposition 3. *Given the state-space Model 1 in* (18) *with the quasi-likelihood function* (24) *based on Gaussian errors, the Laplace approximation based recursive filtering algorithm for* $\boldsymbol{\alpha}_t$ *is given by*

$$\text{Predict}: \quad f(\boldsymbol{\alpha}_t|\mathcal{F}_{t-1}) \propto \text{etr}\{\boldsymbol{D}\boldsymbol{U}_{t-1}'\boldsymbol{\alpha}_t\}, \tag{38}$$

$$\text{Update}: \quad f(\boldsymbol{\alpha}_t|\mathcal{F}_t) \propto \text{etr}\{\boldsymbol{H}_t\boldsymbol{\alpha}_t'\boldsymbol{J}\boldsymbol{\alpha}_t + \boldsymbol{C}_t'\boldsymbol{\alpha}_t\}, \tag{39}$$

where $\boldsymbol{H}_t = -\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{x}_t\boldsymbol{x}_t'\boldsymbol{\beta}$, $\boldsymbol{J} = \boldsymbol{\Omega}^{-1}$, $\boldsymbol{C}_t = \boldsymbol{U}_{t-1}\boldsymbol{D} + \boldsymbol{\Omega}^{-1}(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{z}_t)\boldsymbol{x}_t'\boldsymbol{\beta}$, *and*

$$\boldsymbol{U}_{t-1} = \underset{\boldsymbol{\alpha}_{t-1} \in \mathbb{V}_{p,r}}{\arg\max}\ \text{etr}\{\boldsymbol{H}_{t-1}\boldsymbol{\alpha}_{t-1}'\boldsymbol{J}\boldsymbol{\alpha}_{t-1} + \boldsymbol{C}_{t-1}'\boldsymbol{\alpha}_{t-1}\}. \tag{40}$$

Likewise, we have the recursive filtering algorithm for the state-space Model 2.

Proposition 4. *Given the state-space Model 2 in* (19) *with the quasi-likelihood function* (25) *based on Gaussian errors, the Laplace approximation based recursive filtering algorithm for* $\boldsymbol{\beta}_t$ *is given by*

$$\text{Predict}: \quad f(\boldsymbol{\beta}_t|\mathcal{F}_{t-1}) \propto \text{etr}\{\boldsymbol{D}\boldsymbol{U}_{t-1}'\boldsymbol{\beta}_t\}, \tag{41}$$

$$\text{Update}: \quad f(\boldsymbol{\beta}_t|\mathcal{F}_t) \propto \text{etr}\{\boldsymbol{H}\boldsymbol{\beta}_t'\boldsymbol{J}_t\boldsymbol{\beta}_t + \boldsymbol{C}_t'\boldsymbol{\beta}_t\}, \tag{42}$$

where $\boldsymbol{H} = -\frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}$, $\boldsymbol{J}_t = \boldsymbol{x}_t\boldsymbol{x}_t'$, $\boldsymbol{C}_t = \boldsymbol{U}_{t-1}\boldsymbol{D} + \boldsymbol{x}_t(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{z}_t)'\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}$, *and*

$$\boldsymbol{U}_{t-1} = \underset{\boldsymbol{\beta}_{t-1} \in \mathbb{V}_{q_1,r}}{\arg\max}\ \text{etr}\{\boldsymbol{H}\boldsymbol{\beta}_{t-1}'\boldsymbol{J}_{t-1}\boldsymbol{\beta}_{t-1} + \boldsymbol{C}_{t-1}'\boldsymbol{\beta}_{t-1}\}. \tag{43}$$

Several remarks related to the propositions follow.

Remark 1. *The distributions of predicted and updated* $\boldsymbol{\alpha}_t$ *and* $\boldsymbol{\beta}_t$ *in the recursive filtering algorithms are conjugate.*

The predictive distribution and the updating or filtering distribution are both known as the matrix Langevin–Bingham (or matrix Bingham–von Mises–Fisher) distribution; see, for example, Khatri and Mardia (1977). This feature is desirable as it gives great convenience in the computational implementation of the filtering algorithms.

Remark 2. *When estimating the predicted distribution of $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$, a numerical optimization for finding $\boldsymbol{U}_{t-1}$ is required.*

There are several efficient line-search based optimization algorithms available in the literature which can be easily implemented and applied. See Absil et al. (2008, chp. 4) for a detailed exposition.

Remark 3. *The predictive distributions in (38) and (41) are Laplace type approximations. Therefore, the dimensions of the data $\boldsymbol{y}_t$ in Model 1 and the predictors in Model 2 are expected to be high enough in order to achieve good approximations.*

For the high-dimensional factor models that use a large number of predictors, the filtering algorithms are natural choices to model the possible temporal instability, while a small value of the rank r implies the dimension reduction in forecasting. In the next section, our finding from simulation is that, even for small p and q_1, the approximations of the modal orientations can be very good.

Remark 4. *The recursive filtering algorithms make it possible to use both maximum likelihood estimation and the Bayesian analysis for the proposed state-space models.*

Next, we consider the models in (20) and (21). The corresponding filtering algorithms are similar to Propositions 3 and 4. The filtering algorithm for Model 1* is given by

$$\begin{aligned} \text{Predict}: &\quad f(\boldsymbol{\alpha}_t|\mathcal{F}_{t-1}) \propto \text{etr}\{\boldsymbol{D}\boldsymbol{\alpha}_0'\boldsymbol{\alpha}_t\}, &(44)\\ \text{Update}: &\quad f(\boldsymbol{\alpha}_t|\mathcal{F}_t) \propto \text{etr}\{\boldsymbol{H}_t\boldsymbol{\alpha}_t'\boldsymbol{J}\boldsymbol{\alpha}_t + \boldsymbol{C}_t'\boldsymbol{\alpha}_t\}, &(45)\end{aligned}$$

where $\boldsymbol{H}_t = -\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{x}_t\boldsymbol{x}_t'\boldsymbol{\beta}$, $\boldsymbol{J} = \boldsymbol{\Omega}^{-1}$, $\boldsymbol{C}_t = \boldsymbol{\alpha}_0\boldsymbol{D} + \boldsymbol{\Omega}^{-1}(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{z}_t)\boldsymbol{x}_t'\boldsymbol{\beta}$. In addition, for Model 2*, we have

$$\begin{aligned} \text{Predict}: &\quad f(\boldsymbol{\beta}_t|\mathcal{F}_{t-1}) \propto \text{etr}\{\boldsymbol{D}\boldsymbol{\beta}_0'\boldsymbol{\beta}_t\}, &(46)\\ \text{Update}: &\quad f(\boldsymbol{\beta}_t|\mathcal{F}_t) \propto \text{etr}\{\boldsymbol{H}\boldsymbol{\beta}_t'\boldsymbol{J}_t\boldsymbol{\beta}_t + \boldsymbol{C}_t'\boldsymbol{\beta}_t\}, &(47)\end{aligned}$$

where $\boldsymbol{H} = -\frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}$, $\boldsymbol{J}_t = \boldsymbol{x}_t\boldsymbol{x}_t'$, $\boldsymbol{C}_t = \boldsymbol{\beta}_0\boldsymbol{D} + \boldsymbol{x}_t(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{z}_t)'\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}$. We have the following remarks for both models.

Remark 5. *The predictive distributions do not depend on any previous information, which is due to the assumption of sequentially independent latent processes.*

Remark 6. *The predictive and filtering distributions for Model 1* and Model 2* are not approximations.*

We do not need to approximate integral like (30). Since $f(\boldsymbol{\alpha}_t|\mathcal{F}_{t-1})$ does not depend on $\boldsymbol{\alpha}_{t-1}$ in Model 1* and $f(\boldsymbol{\beta}_t|\mathcal{F}_{t-1})$ does not depend on $\boldsymbol{\beta}_{t-1}$ in Model 2*, $f(\boldsymbol{\alpha}_t|\mathcal{F}_{t-1})$ and $f(\boldsymbol{\beta}_t|\mathcal{F}_{t-1})$ can be directly moved outside the integral.

The smoothing distribution is defined to be the a posteriori distribution of the latent parameters given all the observations. We have the following two propositions for the smoothing distributions of the state-space models.

Proposition 5. *The smoothing distribution of Model 1 is given by*

$$f(\boldsymbol{\Delta}|\boldsymbol{\theta}, \boldsymbol{Y}) \quad \propto \quad \prod_{t=1}^{T} \text{etr}\{\boldsymbol{H}_t\boldsymbol{\alpha}_t{}'\boldsymbol{J}\boldsymbol{\alpha}_t + \boldsymbol{C}_t'\boldsymbol{\alpha}_t\}, \tag{48}$$

where $\boldsymbol{H}_t = -\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{x}_t\boldsymbol{x}_t'\boldsymbol{\beta}$, $\boldsymbol{J} = \boldsymbol{\Omega}^{-1}$, and $\boldsymbol{C}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{D} + \boldsymbol{\Omega}^{-1}(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{z}_t)\boldsymbol{x}_t'\boldsymbol{\beta}$.

Proposition 6. *The smoothing distribution of Model 2 is given by*

$$f(\boldsymbol{\Delta}|\boldsymbol{\theta},\boldsymbol{Y}) \propto \prod_{t=1}^{T} \text{etr}\{\boldsymbol{H}\boldsymbol{\beta_t}'\boldsymbol{J}_t\boldsymbol{\beta_t} + \boldsymbol{C}_t'\boldsymbol{\beta_t}\}, \tag{49}$$

$\boldsymbol{H} = -\frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}$, $\boldsymbol{J}_t = \boldsymbol{x}_t\boldsymbol{x}_t'$, $\boldsymbol{C}_t = \boldsymbol{\beta}_{t-1}\boldsymbol{D} + \boldsymbol{x}_t(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{z}_t)'\boldsymbol{\Omega}^{-1}\boldsymbol{\alpha}$.

There is no closed form for the smoothing distributions as the corresponding normalizing constants are unknown. Hoff (2009) develops a Gibbs sampling algorithm that can be used to sample from these smoothing distributions.

6. Evaluation of the Filtering Algorithms by Simulation Experiments

To investigate the performance of the filtering algorithm in Proposition 3, we consider several settings based on data generated from Model 1 in (18) for different values of its parameters.

Recall that at each iteration of the recursive algorithm, the predictive density kernel in (38) is a Laplace type approximation of the true predictive density which takes an integral form as (30), and hence the resulting filtering density is an approximation as well. It is of great interest to check the performance of the approximation under different settings. Since the exact filtering distributions of the latent process are not available, we resort to comparing the true (i.e., generated) value $\boldsymbol{\alpha}_t$ and the filtered modal orientation at time t from the filtering distribution $f(\boldsymbol{\alpha}_t|\mathcal{F}_t)$, which is $\boldsymbol{U}_t$ as defined in (40). The modal orientations are expected to be distributed around the true values across time if the algorithm performs well.

Then, a measure of distance between two points in the Stiefel manifold is needed for the comparison. We consider the squared Frobenius norm of the difference between two matrices or column vectors:

$$\begin{aligned} F^2(X,Y) &= ||X-Y||^2 = \text{tr}\{(X-Y)'(X-Y)\} \\ &= \text{tr}\{X'X + Y'Y - X'Y - Y'X\}. \end{aligned} \tag{50}$$

If the two matrices or column vectors X and Y are points in the Stiefel manifold, then it holds that $F^2(X,Y) = 2r - 2\text{tr}\{X'Y\} \in [0, 4r]$, and $F^2(X,Y)$ takes the minimum 0 when $X = Y$ (closest) and the maximum $4r$ when $X = -Y$ (furthest). Thus, we employ the normalized distance

$$\delta(X,Y) = F^2(X,Y)/4r \in [0,1], \tag{51}$$

which is matrix dimension free.

Note that the modal orientation of the filtering distribution is not supposed to be consistent to the true value of the latent process with the increase of the sample size T. As a matter of fact, the sample size is irrelevant to the consistency which can be seen from the filtering density (39). We should note that the filtering distribution in (39) also has concentration or dispersion which is determined by $\boldsymbol{H}_t$, $\boldsymbol{J}$ (the inverse of $\boldsymbol{\Omega}$) and $\boldsymbol{C}_t$ (the current information, i.e. $\boldsymbol{y}_t$, $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$), together with the parameters, while the previous information has limited influence only through the orthonormal matrix $\boldsymbol{U}_{t-1}$. Since the concentration of the filtering distribution does not shrink with the increase of the sample size, we use $T = 100$ in all the experiments. If the filtering distribution has big concentration, the filtered modal orientations are expected to be close to the true values and hence the normalized distances close to zero and less dispersed.

The data generating process follows Model 1 in (18). Since we input the true parameters in the filtering algorithm, the difference $\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{z}_t$ is perfectly known and then there is no need to consider the effect of $\boldsymbol{B}\boldsymbol{z}_t$. Thus, it is natural to exclude $\boldsymbol{B}\boldsymbol{z}_t$ from the data generating process.

We consider the settings with different combinations of

- $T = 100$, the sample size,

- $p \in \{2, 3, 10, 20\}$, the dimension of the dependent variable $\boldsymbol{y}_t$,
- $r \in \{1, 2\}$, the rank of the matrix $\boldsymbol{A}_t$,
- $\boldsymbol{x}_t$, the explanatory variable vector has dimension $q_1 = 3$ ensuring that $q_1 > r$ always holds, and each $\boldsymbol{x}_t$ is sampled independently (over time) from a $N_3(\boldsymbol{0}, \boldsymbol{I}_3)$,
- $\boldsymbol{\beta} = (1, -1, 1)'/\sqrt{3}$,
- $\boldsymbol{\alpha}_0 = (1, -1, 1, \ldots)'/\sqrt{p}$, the initial value of $\boldsymbol{\alpha}_t$ sequence for the data generating process,
- $\boldsymbol{\Omega} = \rho \boldsymbol{I}_p$, the covariance matrix of the errors is diagonal with $\rho \in \{0.1, 0.5, 1\}$,
- $\boldsymbol{D} = d\boldsymbol{I}_r$, and $d \in \{5, 50, 500, 800\}$.

The simulation based experiment of each setting consists of the following three steps:

1. We sample from Model 1 by using the identified version in (18). First, simulate $\boldsymbol{\alpha}_t$ given $\boldsymbol{\alpha}_{t-1}$, and then $\boldsymbol{y}_t$ given $\boldsymbol{\alpha}_t$. We save the sequence of the latent process $\boldsymbol{\alpha}_t$, $t = 1, \ldots, T$.
2. Then, we apply the filtering algorithm on the sampled data to obtain the filtered modal orientation $\boldsymbol{U}_t$, $t = 1, \ldots, T$.
3. We compute the normalized distances $\delta_t(\boldsymbol{\alpha}_t, \boldsymbol{U}_t)$ and report by plotting them against the time t.

We use the same seed, which is one, for the underlying random number generator throughout the experiments so that all the results can be replicated. Sampling values form the matrix Langevin distribution can be done by the rejection method described in Section 2.5.2 of Chikuse (2003).

Figure 4 depicts the results from the setting $p \in \{2, 10, 20\}$, $r = 1$, $\rho = 0.1$ and $d = 50$. We see that the sequences of the normalized distances δ_t are persistent. This is a common phenomenon throughout the experiments, and, intuitively, it can be attributed to the fact that the current δ_t depends on the previous one through the pair of $\boldsymbol{U}_t$ and $\boldsymbol{\alpha}_t$. For the low dimensional case $p = 2$, almost all the distances are very close to 0, which means that the filtered modal orientations are very close to the true ones, despite few exceptions. However, for the higher dimensional cases $p = 10$ and 20, the distances are at higher levels and are more dispersed. This is consistent with the fact that, given the same concentration $d = 50$, an increase of the dimension the orthonormal matrix or vector goes along with an increase of the dispersion of the corresponding distributions on the Stiefel manifold, as the volume of the manifold explodes with the increase of the dimensions (both p and r).

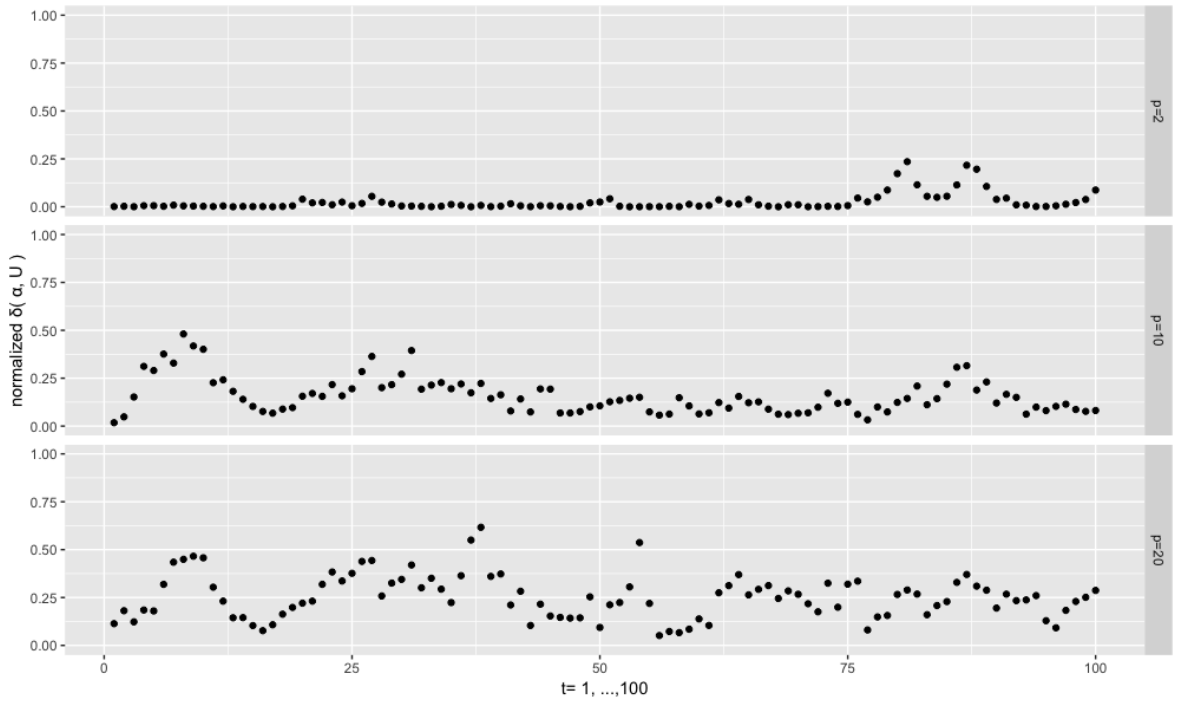

Figure 4. Normalized distances δ_t for the settings $p \in \{2, 10, 20\}$, $r = 1$, $\rho = 0.1$ and $d = 50$.

Figure 5 displays the results for the same setting $p \in \{2, 10, 20\}$, $r = 1$, $\rho = 0.1$ but with a much higher concentration $d = 500$. We see that the curse of dimensionality can be remedied through a higher concentration as the distances for the high dimensional cases are much closer to zero than when $d = 50$.

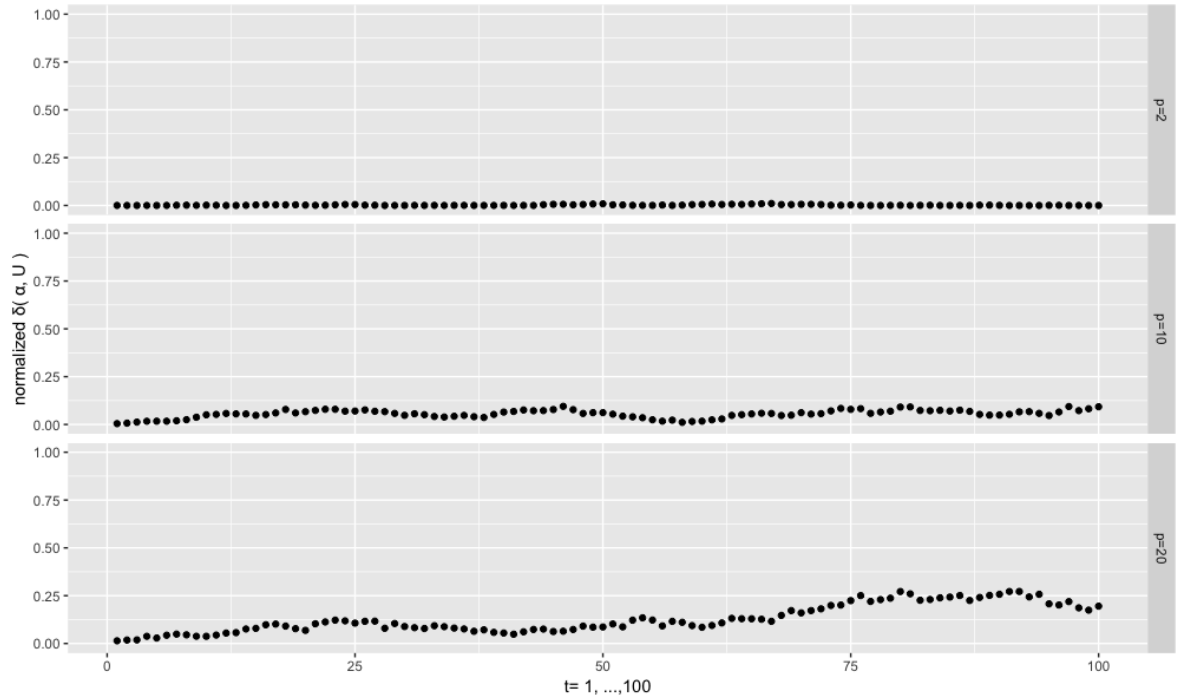

Figure 5. Normalized distances δ_t for the settings $p \in \{2, 10, 20\}$, $r = 1$, $\rho = 0.1$ and $d = 500$.

The magnitude ρ of the variance of the errors affects the results of the filtering algorithm as well, as it determines the concentration of the filtering distribution, which can be seen from (39) through $\boldsymbol{J}$ and $\boldsymbol{C}_t$ (both depend on the inverse of $\boldsymbol{\Omega}$). The following experiments apply the settings $p = 2$, $r = 1$ and $d \in \{5, 50, 500\}$ showing the impact of different ρ on the filtering results. Figure 6 depicts the results with $\rho = 1$, and Figure 7 with $\rho = 0.1$. We see that the normalized distances become closer to zero when a lower ρ is applied. Their variability also decreases for the lowest value of $d = 5$ and for the intermediate value $d = 50$. It is worth mentioning that, in the two cases corresponding to the two bottom plots of the figures, the matrix $\boldsymbol{C}_t$ dominates the density function, which implies that the filtering distribution resembles a highly concentrated matrix Langevin.

In the following experiments, our focus is on the investigation of the filtering algorithm when r approaches p. We consider the setting $p = 3$ with the rank number $r \in \{1, 2\}$, with $\rho = 0.1$ and $d = 500$. Figure 8 depicts the results. The normalized distances are stable at a low level for the case $p = 3$ with $r = 1$, but a high level (around 0.5) in the case $p = 3$ with $r = 2$. A higher concentration ($d = 800$) reduces the latter level to about 0.12, as can be seen on the lower plot of Figure 8. We conclude that the approximation of the true filtering distribution tends to fail when the matrix $\boldsymbol{\alpha}_t$ tends to a square matrix, that is, $p \approx r$, and therefore the filtering algorithms proposed in this paper seems to be appropriate when p is sufficiently larger than r.

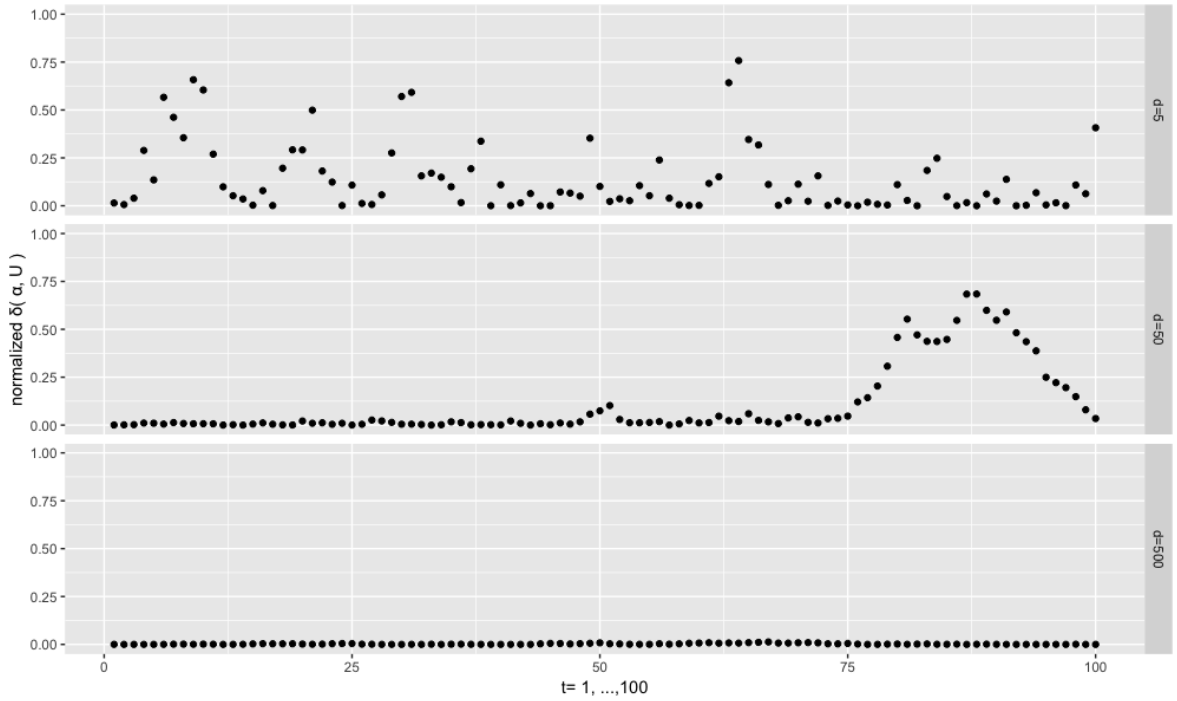

Figure 6. Normalized distances δ_t for the settings $p = 2$, $r = 1$, $\rho = 1$ and $d \in \{5, 50, 500\}$.

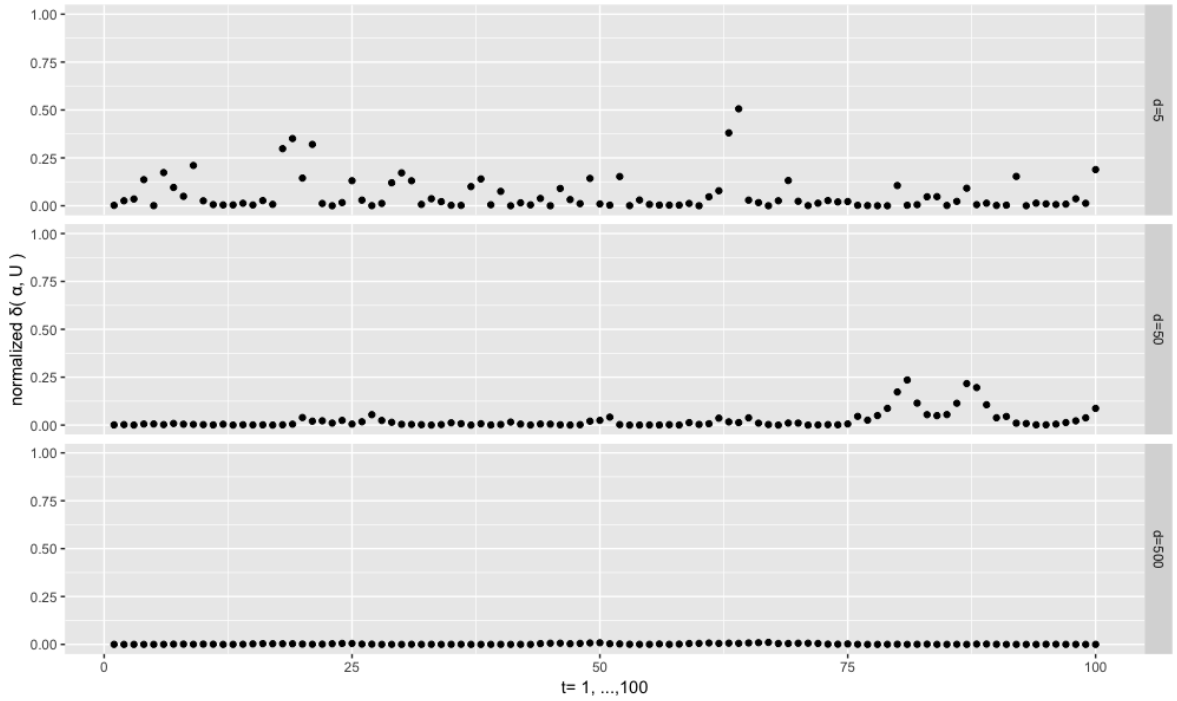

Figure 7. Normalized distances δ_t for the settings $p = 2$, $r = 1$, $\rho = 0.1$ and $d \in \{5, 50, 500\}$.

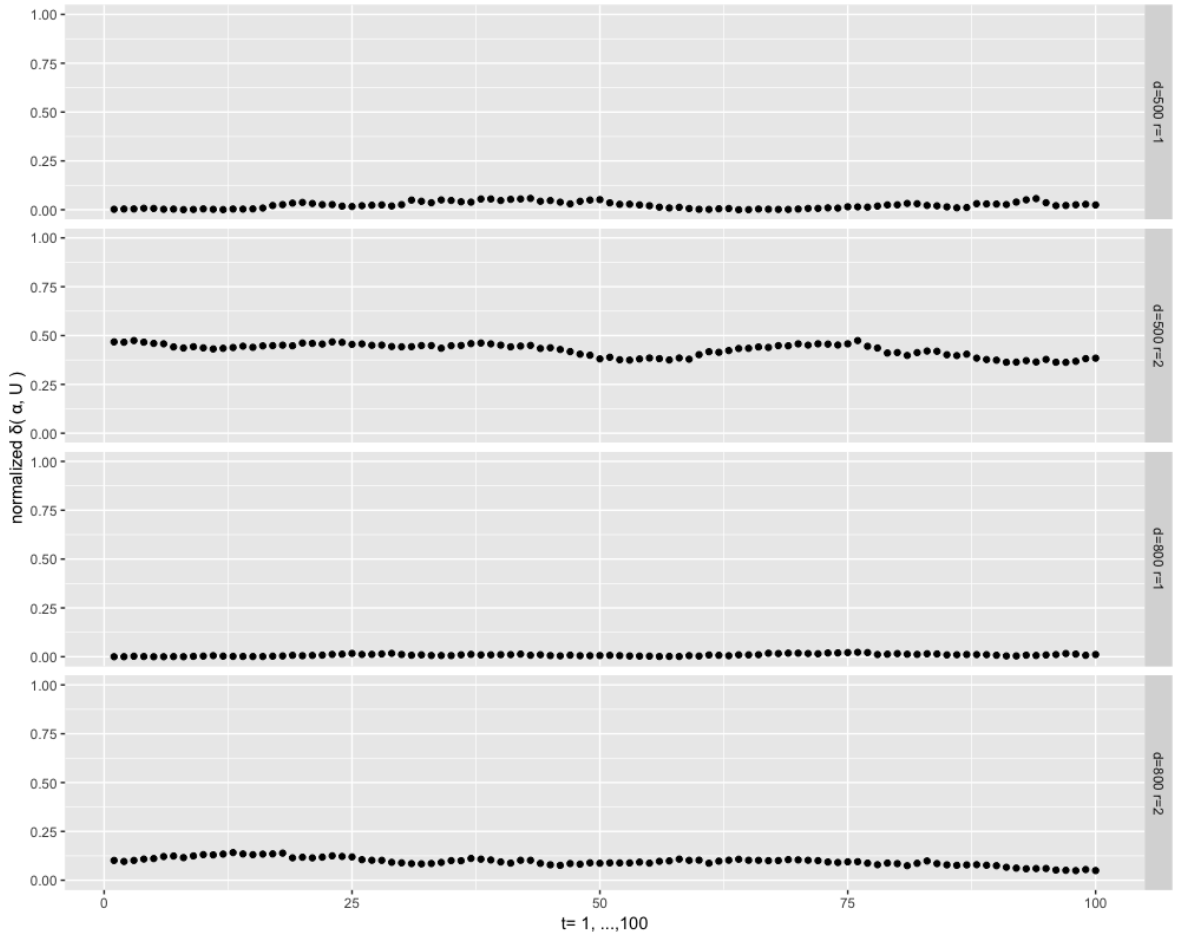

Figure 8. Normalized distances δ_t for the settings $p = 3$, $r \in \{1, 2\}$, $\rho = 0.1$ and $d = \{500, 800\}$.

All the previous experiments are based on the true initial value $\boldsymbol{\alpha}_0$, but, in practice, this is unknown. The filtering algorithm may be sensitive to the choice of the initial value. In the following experiments, we look into the effect of a wrong initial value. The setting is $p \in \{2, 10, 20\}$, $r = 1$, $\rho = 0.1$ and $d = 50$, and we use as initial value $-\boldsymbol{\alpha}_0$, which is the furthest point in the Stiefel manifold away from the true one. Figure 9 depicts the results. We see that in all the experiments the normalized distances move towards zero, hence the filtered values approach the true values in no more than 20 steps. After that, the level and dispersion of the distance series are similar to what they are in Figure 4 where the true initial value is used. Thus, we can conclude that the effect of a wrongly chosen initial value is temporary.

We have conducted similar simulation experiments for Model 2 in (19) to investigate the performance of the algorithm proposed in Proposition 4. We find similar results to those for Model 1. All the experiments that we have conducted are replicable using the R code available at https://github.com/yukai-yang/SMFilter_Experiments, and the corresponding R package SMFilter implementing the filtering algorithms of this paper is available at the Comprehensive R Archive Network (CRAN).

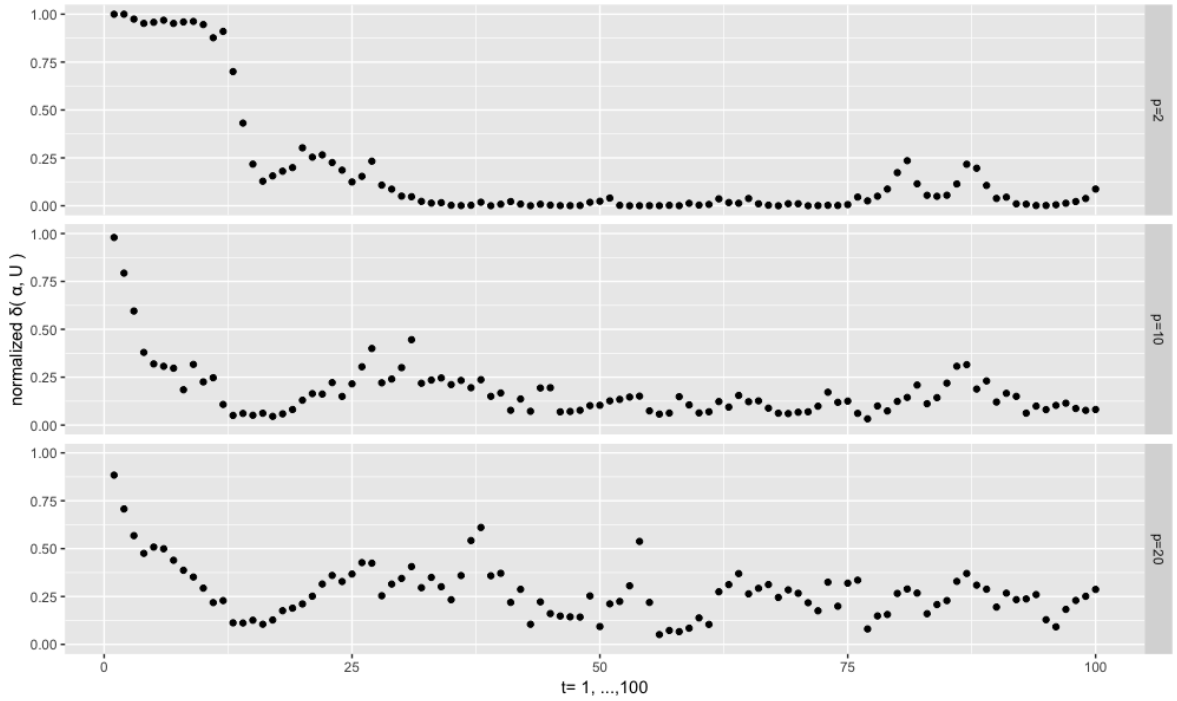

Figure 9. Normalized distances δ_t for the settings $p \in \{2, 10, 20\}$, $r = 1$, $\rho = 0.1$ and $d = 50$. The initial value of the filtering algorithm is $-\boldsymbol{\alpha}_0$.

7. Conclusions

In this paper, we discuss the modelling of the time dependence of the time-varying reduced rank parameters in multivariate time series models and develop novel state-space models whose latent states evolve on the Stiefel manifold. Almost all the existing models in the past literature only deal with the case where the evolution of the latent processes takes places on the Euclidean space, and we point out that this approach can be problematic. These problems motivate the development of the novel state-space models. The matrix Langevin distribution is proposed to specify the sequential evolution of the corresponding latent processes over the Stiefel manifold. Nonlinear filtering algorithms for the new models are designed, wherein the integral for computing the predictive step is approximated by applying the Laplace method. An advantage of the matrix Langevin distribution is that the a priori and a posteriori distributions of the latent variables are conjugate. The new models can be useful when the temporal instability of some parameters of multvariate models is suspected, for example, cointegration models with time-varying short-run adjustment or time-varying long-run relations, and factor models with time-varying factor loading.

Further research is needed in several directions. The most obvious one is the implementation of estimation methods, which can be maximum likelihood or Bayesian inference, and the investigation of their properties. This will enable us to apply the models to data. In this paper, we only consider the case where the latent variables evolve on the Stiefel manifold in a 'random walk' way. It will be interesting to consider the case where the latent variables evolve on the Stiefel manifold but in a mean-reverting way.

Author Contributions: Conceptualization, Y.Y.; Methodology, Y.Y.; Software, Y.Y.; Formal analysis, Y.Y. and L.B.; Investigation, Y.Y. and L.B.; Validation, Y.Y. and L.B.; Funding acquisition, Y.Y.; Writing—original draft, Y.Y.; Writing—review and editing, Y.Y. and L.B.

Funding: This research was funded by Jan Wallander's and Tom Hedelius's Foundation grant number P2016-0293:1.

Acknowledgments: The authors would like to thank the two referees for helpful comments. Yukai Yang acknowledges support from Jan Wallander's and Tom Hedelius's Foundation. Responsibility for any errors or shortcomings in this work remains ours.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proof of Proposition 1

In the model (1) with the decomposition (2), both the rows of $\boldsymbol{\beta}$ and the order of the variables $\boldsymbol{x}_t$ are permuted by $\boldsymbol{P}_\beta$ as follows:

$$\boldsymbol{A}_t\boldsymbol{x}_t = \boldsymbol{\alpha}_t\boldsymbol{\beta}'\boldsymbol{x}_t = \boldsymbol{\alpha}_t\boldsymbol{\beta}'\boldsymbol{P}_\beta'\boldsymbol{P}_\beta\boldsymbol{x}_t. \tag{A1}$$

The time-invariant component $\boldsymbol{\beta}$ can be linearly normalized if the $r \times r$ upper block $\boldsymbol{b}_1$ in (7) is invertible. It follows that the corresponding linear normalization defined in (8) is due to

$$\boldsymbol{\alpha}_t\boldsymbol{\beta}'\boldsymbol{P}_\beta'\boldsymbol{P}_\beta\boldsymbol{x}_t = \boldsymbol{\alpha}_t(\boldsymbol{b}_1', \boldsymbol{b}_2')\boldsymbol{P}_\beta\boldsymbol{x}_t = \boldsymbol{\alpha}_t\boldsymbol{b}_1'(\boldsymbol{b}_1')^{-1}(\boldsymbol{b}_1', \boldsymbol{b}_2')\boldsymbol{P}_\beta\boldsymbol{x}_t = \tilde{\boldsymbol{\alpha}}_t\tilde{\boldsymbol{\beta}}'\boldsymbol{P}_\beta\boldsymbol{x}_t, \tag{A2}$$

where $\tilde{\boldsymbol{\alpha}}_t = \boldsymbol{\alpha}_t\boldsymbol{b}_1'$ is the new time-varying component following the evolution (9).

Consider another permutation $\boldsymbol{P}_\beta^* \neq \boldsymbol{P}_\beta$. Similarly, we have

$$\boldsymbol{A}_t\boldsymbol{x}_t = \boldsymbol{\alpha}_t\boldsymbol{\beta}'\boldsymbol{P}_\beta^{*\prime}\boldsymbol{P}_\beta^*\boldsymbol{x}_t, \tag{A3}$$

together with

$$\boldsymbol{P}_\beta^*\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{b}_1^* \\ \boldsymbol{b}_2^* \end{pmatrix}, \quad \tilde{\boldsymbol{\beta}}^* = \boldsymbol{P}_\beta^*\boldsymbol{\beta}\boldsymbol{b}_1^{*-1} = \begin{pmatrix} \boldsymbol{I}_r \\ \boldsymbol{b}_2^*\boldsymbol{b}_1^{*-1} \end{pmatrix}, \quad \tilde{\boldsymbol{\alpha}}_t^* = \boldsymbol{\alpha}_t\boldsymbol{b}_1^{*\prime}, \tag{A4}$$

where $\boldsymbol{b}_1^*$ is also invertible. Then, we can have the evolution

$$\mathrm{vec}(\tilde{\boldsymbol{\alpha}}_{t+1}^*) = \mathrm{vec}(\tilde{\boldsymbol{\alpha}}_t^*) + \boldsymbol{\eta}_t^{\alpha *}. \tag{A5}$$

Assume that the error vector $\boldsymbol{\eta}_t^\alpha$ in (9) has zero mean and a diagonal variance–covariance matrix. From (A1)–(A3), we have

$$\boldsymbol{A}_t = \tilde{\boldsymbol{\alpha}}_t\tilde{\boldsymbol{\beta}}'\boldsymbol{P}_\beta = \tilde{\boldsymbol{\alpha}}_t^*\tilde{\boldsymbol{\beta}}^{*\prime}\boldsymbol{P}_\beta^*, \tag{A6}$$

and hence it follows that

$$\tilde{\boldsymbol{\alpha}}_t^* = \tilde{\boldsymbol{\alpha}}_t\tilde{\boldsymbol{\beta}}'\boldsymbol{P}_\beta\boldsymbol{P}_\beta^{*\prime}\boldsymbol{\kappa}, \tag{A7}$$

where the $q_1 \times r$ matrix $\boldsymbol{\kappa}$ satisfies $\tilde{\boldsymbol{\beta}}^{*\prime}\boldsymbol{\kappa} = \boldsymbol{I}_r$. The existence of $\boldsymbol{\kappa}$ is guaranteed by the fact that $\boldsymbol{\beta}$ has full rank and so does $\tilde{\boldsymbol{\beta}}^*$.

Thus, the vectorized $\tilde{\boldsymbol{\alpha}}_{t+1}^*$ can be written as

$$\begin{aligned} \mathrm{vec}(\tilde{\boldsymbol{\alpha}}_{t+1}^*) &= \mathrm{vec}(\tilde{\boldsymbol{\alpha}}_{t+1}\tilde{\boldsymbol{\beta}}'\boldsymbol{P}_\beta\boldsymbol{P}_\beta^{*\prime}\boldsymbol{\kappa}) = ((\boldsymbol{\kappa}'\boldsymbol{P}_\beta^*\boldsymbol{P}_\beta'\tilde{\boldsymbol{\beta}}) \otimes \boldsymbol{I}_p)\,\mathrm{vec}(\tilde{\boldsymbol{\alpha}}_{t+1}) \\ &= ((\boldsymbol{\kappa}'\boldsymbol{P}_\beta^*\boldsymbol{P}_\beta'\tilde{\boldsymbol{\beta}}) \otimes \boldsymbol{I}_p)\,\mathrm{vec}(\tilde{\boldsymbol{\alpha}}_t) + ((\boldsymbol{\kappa}'\boldsymbol{P}_\beta^*\boldsymbol{P}_\beta'\tilde{\boldsymbol{\beta}}) \otimes \boldsymbol{I}_p)\,\boldsymbol{\eta}_t^\alpha \\ &= \mathrm{vec}(\tilde{\boldsymbol{\alpha}}_t^*) + \boldsymbol{\eta}_t^{\alpha *}, \end{aligned} \tag{A8}$$

due to (9) and (A5). Hence, it can be seen that $\boldsymbol{\eta}_t^{\alpha *} = ((\boldsymbol{\kappa}'\boldsymbol{P}_\beta^*\boldsymbol{P}_\beta'\tilde{\boldsymbol{\beta}}) \otimes \boldsymbol{I}_p)\,\boldsymbol{\eta}_t^\alpha$, and that $\boldsymbol{\eta}_t^{\alpha *}$ has diagonal variance–covariance matrix if and only if $\boldsymbol{\kappa}'\boldsymbol{P}_\beta^*\boldsymbol{P}_\beta'\tilde{\boldsymbol{\beta}}$ is diagonal given that $\boldsymbol{\eta}_t^\alpha$ has diagonal variance–covariance matrix.

Next, we need to verify whether $\boldsymbol{\kappa}'\boldsymbol{P}_\beta^*\boldsymbol{P}_\beta'\tilde{\boldsymbol{\beta}}$ is diagonal and investigate under what condition it will be diagonal. By substituting $\tilde{\boldsymbol{\beta}}$ with (8), we obtain

$$\boldsymbol{\kappa}'\boldsymbol{P}_\beta^*\boldsymbol{P}_\beta'\tilde{\boldsymbol{\beta}} = \boldsymbol{\kappa}'\boldsymbol{P}_\beta^*\boldsymbol{P}_\beta'\boldsymbol{P}_\beta\boldsymbol{\beta}\boldsymbol{b}_1^{-1} = \boldsymbol{\kappa}'\boldsymbol{P}_\beta^*\boldsymbol{\beta}\boldsymbol{b}_1^{-1}. \tag{A9}$$

In addition, we know that, by substituting $\tilde{\boldsymbol{\beta}}^*$ with (A4),

$$\boldsymbol{\kappa}'\tilde{\boldsymbol{\beta}}^* = \boldsymbol{\kappa}'\boldsymbol{P}_\beta^*\boldsymbol{\beta}\boldsymbol{b}_1^{*-1} = \boldsymbol{I}_r. \tag{A10}$$

Since the $r \times r$ square matrix $\boldsymbol{\kappa}' \boldsymbol{P}_\beta^* \boldsymbol{\beta}$ has full rank, it can be seen that $\boldsymbol{\eta}_t^{\alpha *}$ has diagonal variance–covariance matrix if and only if $\boldsymbol{b}_1 = \boldsymbol{b}_1^*$.

Appendix B. Proof of Proposition 2

In the model (1) with the decomposition (3), the rows of $\boldsymbol{\alpha}$ are permuted by $\boldsymbol{P}_\alpha$ as follows:

$$\boldsymbol{A}_t \boldsymbol{x}_t = \boldsymbol{\alpha} \boldsymbol{\beta}_t' \boldsymbol{x}_t = \boldsymbol{P}_\alpha' \boldsymbol{P}_\alpha \boldsymbol{\alpha} \boldsymbol{\beta}_t' \boldsymbol{x}_t. \tag{A11}$$

Notice that we can remove $\boldsymbol{P}_\alpha'$ in the equation, which means that we choose not to permute back to the original order of the dependent variables $\boldsymbol{y}_t$. The linear normalization (11) is obtained by

$$\boldsymbol{P}_\alpha' \boldsymbol{P}_\alpha \boldsymbol{\alpha} \boldsymbol{\beta}_t' \boldsymbol{x}_t = \boldsymbol{P}_\alpha' \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{pmatrix} \boldsymbol{\beta}_t' \boldsymbol{x}_t = \boldsymbol{P}_\alpha' \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{pmatrix} \boldsymbol{a}_1^{-1} \boldsymbol{a}_1 \boldsymbol{\beta}_t' \boldsymbol{x}_t = \boldsymbol{P}_\alpha' \tilde{\boldsymbol{\alpha}} \tilde{\boldsymbol{\beta}}_t' \boldsymbol{x}_t, \tag{A12}$$

where $\tilde{\boldsymbol{\beta}}_t = \boldsymbol{\beta}_t \boldsymbol{a}_1'$ is the new time-varying component following evolution (12).

Consider another permutation $\boldsymbol{P}_\alpha^* \neq \boldsymbol{P}_\alpha$, such that

$$\boldsymbol{A}_t \boldsymbol{x}_t = \boldsymbol{P}_\alpha^{*\prime} \boldsymbol{P}_\alpha^* \boldsymbol{\alpha} \boldsymbol{\beta}_t' \boldsymbol{x}_t, \tag{A13}$$

together with

$$\boldsymbol{P}_\alpha^* \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{a}_1^* \\ \boldsymbol{a}_2^* \end{pmatrix}, \quad \tilde{\boldsymbol{\alpha}}^* = \boldsymbol{P}_\alpha^* \boldsymbol{\alpha} \boldsymbol{a}_1^{*-1} = \begin{pmatrix} \boldsymbol{I}_r \\ \boldsymbol{a}_2^* \boldsymbol{a}_1^{*-1} \end{pmatrix}, \quad \tilde{\boldsymbol{\beta}}_t^* = \boldsymbol{\beta}_t \boldsymbol{a}_1^{*\prime}, \tag{A14}$$

where $\boldsymbol{a}_1^*$ is invertible. Then, we can have the evolution

$$\operatorname{vec}(\tilde{\boldsymbol{\beta}}_{t+1}^*) = \operatorname{vec}(\tilde{\boldsymbol{\beta}}_t^*) + \boldsymbol{\eta}_t^{\beta *}. \tag{A15}$$

We assume that the error vector $\boldsymbol{\eta}_t^\beta$ in (12) has zero mean and diagonal variance–covariance matrix. From (A11)–(A13), we have

$$\boldsymbol{A}_t = \boldsymbol{P}_\alpha' \tilde{\boldsymbol{\alpha}} \tilde{\boldsymbol{\beta}}_t' = \boldsymbol{P}_\alpha^{*\prime} \tilde{\boldsymbol{\alpha}}^* \tilde{\boldsymbol{\beta}}_t^{*\prime}, \tag{A16}$$

and hence it follows that

$$\tilde{\boldsymbol{\beta}}_t^* = \tilde{\boldsymbol{\beta}}_t \tilde{\boldsymbol{\alpha}}' \boldsymbol{P}_\alpha \boldsymbol{P}_\alpha^{*\prime} \delta, \tag{A17}$$

where the $p \times r$ matrix δ satisfies $\tilde{\boldsymbol{\alpha}}^{*\prime} \delta = \boldsymbol{I}_r$. The existence of δ is guaranteed by the fact that $\boldsymbol{\alpha}$ has full rank and so does $\tilde{\boldsymbol{\alpha}}^*$.

Then, we get the vectorized $\tilde{\boldsymbol{\beta}}_{t+1}^*$:

$$\begin{aligned} \operatorname{vec}(\tilde{\boldsymbol{\beta}}_{t+1}^*) &= \operatorname{vec}(\tilde{\boldsymbol{\beta}}_{t+1} \tilde{\boldsymbol{\alpha}}' \boldsymbol{P}_\alpha \boldsymbol{P}_\alpha^{*\prime} \delta) = ((\delta' \boldsymbol{P}_\alpha^* \boldsymbol{P}_\alpha' \tilde{\boldsymbol{\alpha}}) \otimes \boldsymbol{I}_{q_1}) \operatorname{vec}(\tilde{\boldsymbol{\beta}}_{t+1}) \\ &= ((\delta' \boldsymbol{P}_\alpha^* \boldsymbol{P}_\alpha' \tilde{\boldsymbol{\alpha}}) \otimes \boldsymbol{I}_{q_1}) \operatorname{vec}(\tilde{\boldsymbol{\beta}}_t) + ((\delta' \boldsymbol{P}_\alpha^* \boldsymbol{P}_\alpha' \tilde{\boldsymbol{\alpha}}) \otimes \boldsymbol{I}_{q_1}) \boldsymbol{\eta}_t^\beta \\ &= \operatorname{vec}(\tilde{\boldsymbol{\beta}}_t^*) + \boldsymbol{\eta}_t^{\beta *}, \end{aligned} \tag{A18}$$

due to (12) and (A15). Hence, it can be seen that $\boldsymbol{\eta}_t^{\beta *} = ((\delta' \boldsymbol{P}_\alpha^* \boldsymbol{P}_\alpha' \tilde{\boldsymbol{\alpha}}) \otimes \boldsymbol{I}_{q_1}) \boldsymbol{\eta}_t^\beta$, and that $\boldsymbol{\eta}_t^{\beta *}$ has a diagonal variance–covariance matrix if and only if $\delta' \boldsymbol{P}_\alpha^* \boldsymbol{P}_\alpha' \tilde{\boldsymbol{\alpha}}$ is diagonal given that $\boldsymbol{\eta}_t^\beta$ has a diagonal variance–covariance matrix.

The investigation of under what condition $\delta' \boldsymbol{P}_\alpha^* \boldsymbol{P}_\alpha' \tilde{\boldsymbol{\alpha}}$ is diagonal is similar to the previous proof. By substituting $\tilde{\boldsymbol{\alpha}}$ with (11), we obtain

$$\delta' \boldsymbol{P}_\alpha^* \boldsymbol{P}_\alpha' \tilde{\boldsymbol{\alpha}} = \delta' \boldsymbol{P}_\alpha^* \boldsymbol{P}_\alpha' \boldsymbol{P}_\alpha \boldsymbol{\alpha} \boldsymbol{a}_1^{-1} = \delta' \boldsymbol{P}_\alpha^* \boldsymbol{\alpha} \boldsymbol{a}_1^{-1}. \tag{A19}$$

By substituting $\tilde{\boldsymbol{\beta}}^*$ with (A14), we obtain that

$$\boldsymbol{\delta}'\tilde{\boldsymbol{\alpha}}^* = \boldsymbol{\delta}'\boldsymbol{P}_{\alpha}^*\boldsymbol{\alpha}\boldsymbol{a}_1^{*-1} = \boldsymbol{I}_r. \tag{A20}$$

Since the $r \times r$ square matrix $\boldsymbol{\delta}'\boldsymbol{P}_{\alpha}^*\boldsymbol{\alpha}$ has full rank, it can be seen that $\boldsymbol{\eta}_t^{\beta*}$ has diagonal variance–covariance matrix if and only if $\boldsymbol{a}_1 = \boldsymbol{a}_1^*$.

References

Absil, Pierre-Antoine, Robert Mahony, and Rodolphe Sepulchre. 2008. *Optimization Algorithms on Matrix Manifolds*. Princeton: Princeton University Press.

Anderson, Theodore Wilbur. 1971. *The Statistical Analysis of Time Series*. New York: Wiley.

Bierens, Herman J., and Luis F. Martins. 2010. Time-varying cointegration. *Econometric Theory* 26: 1453–90. [CrossRef]

Breitung, Jörg, and Sandra Eickmeier. 2011. Testing for structural breaks in dynamic factor models. *Journal of Econometrics* 163: 71–84. [CrossRef]

Casals, Jose, Alfredo Garcia-Hiernaux, Miguel Jerez, Sonia Sotoca, and A. Alexandre Trindade. 2016. *State-Space Methods for Time Series Analysis: Theory, Applications and Software*. Chapman & Hall/CRC, Monographs on Statistics & Applied Probability. New York: CRC Press.

Chikuse, Yasuko. 2003. *Statistics on Special Manifolds*. New York: Springer.

Chikuse, Yasuko. 2006. State space models on special manifolds. *Journal of Multivariate Analysis* 97: 1284–94. [CrossRef]

Del Negro, Marco, and Christopher Otrok. 2008. *Dynamic Factor Models with Time-Varying Parameters: Measuring Changes in International Business Cycles*. Staff Report No. 326. New York: Federal Reserve Bank of New York.

Durbin, James, and Siem Jan Koopman. 2012. *Time Series Analysis by State Space Methods*, 2nd ed. Oxford: Oxford University Press.

Eickmeier, Sandra, Wolfgang Lemke, and Massimiliano Marcellino. 2014. Classical time varying factor-augmented vector auto-regressive models—Estimation, forecasting and structural analysis. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 178: 493–533. [CrossRef]

Hamilton, James Douglas. 1994. *Time Series Analysis*. Princeton: Princeton University Press.

Hannan, Edward J. 1970. *Multiple Time Series*. New York: Wiley.

Herz, Carl S. 1955. Bessel functions of matrix argument. *Annals of Mathematics* 61: 474–523. [CrossRef]

Hoff, Peter D. 2009. Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics* 18: 438–56. [CrossRef]

Khatri, C. G., and Kanti V. Mardia. 1977. The von Mises-Fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society, Series B* 39: 95–106. [CrossRef]

Koopman, Lambert Herman. 1974. *The Spectral Analysis of Time Series*. New York: Academic Press.

Mardia, Kanti V. 1975. Statistics of directional data (with discussion). *Journal of the Royal Statistical Society, Series B* 37: 349–93.

Prentice, Michael J. 1982. Antipodally symmetric distributions for orientation statistics. *Journal of Statistical Planning and Inference* 6: 205–14. [CrossRef]

Rothman, Philip, Dick van Dijk, and Philip Hans Franses. 2001. A Multivariate STAR analysis of the relationship between money and output. *Macroeconomic Dynamics* 5: 506–32.

Stock, James, and Mark Watson. 2009. Forecasting in dynamic factor models subject to structural instability. In *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*. Edited by David F. Hendry, Jennifer Castle and Neil Shephard. Oxford: Oxford University Press, pp. 173–205.

Stock, James H., and Mark W. Watson. 2002. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97: 1167–79. [CrossRef]

Swanson, Norman Rasmus. 1998. Finite sample properties of a simple LM test for neglected nonlinearity in error correcting regression equations. *Statistica Neerlandica* 53: 76–95. [CrossRef]

Wong, Roderick S. C. 2001. *Asymptotic Approximations of Integrals*. In *Classics in Applied Mathematics*. Philadelphia: Society for Industrial and Applied Mathematics.