A state-space approach to time-varying reduced-rank regression

Barbara Brunea, Wolfgang Scherrer, and Efstatia Bura

ABSTRACT
We propose a new approach to reduced-rank regression that allows for time-variation in the regression coefficients. The Kalman filter based estimation allows for usage of standard methods and easy implementation of our procedure. The EM-algorithm ensures convergence to a local maximum of the likelihood. Our estimation approach in time-varying reduced-rank regression performs well in simulations, with amplified competitive advantage in time series that experience large structural changes. We illustrate the performance of our approach with a simulation study and two applications to stock index and Covid-19 case data.

KEYWORDS
EM-algorithm; Kalman filter; time-varying parameters; vector error correction model

1. Introduction
The simultaneous modeling and fitting of \( p \) time series when \( p \) is large is a challenging task as the number of parameters grows with \( p^2 \) and can often exceed the number of observations. Five main approaches to reduce the number of parameters in multivariate time series models have been proposed: (1) reduced-rank vector autoregressive (VAR) and multivariate autoregressive index models (MAI) models (Ahn and Reinsel, 1988; Carriero, Kapetanios, and Marcellino, 2016; Velu, Reinsel, and Wichern, 1986), where the parameter matrix is assumed to have a lower rank structure; (2) shrinkage Bayesian VAR with shrinking prior distributions for controlling the number of parameters (Ghosh, Khare, and Michailidis, 2019; Litterman, 1986; Sims and Zha, 1998); (3) sparse VAR estimation, where a low-dimensional analysis of groups of series is first carried out to identify rows and columns of parameter matrices that can be assumed to have zero values, as in Davis, Zang, and Zheng (2016); (4) regularized estimation, where a penalty term is included in the estimation objective function (see, e.g., Nicholson, Matteson, and Bien, 2017; Song and Bickel, 2011); and (5) the popular dynamic factor model (DFM), introduced by Geweke (1977) and Sargent and Sims (1977). Using the idea of risk diversification discussed in Chamberlain (1983), the so called generalized dynamic factor models (GDFM) have been introduced, extensively studied and developed by many researchers (see, e.g., Anderson and Deistler, 2008; Bai and Ng, 2002; Forni et al. 2000, 2005, 2015; Peña and Yohai, 2016; Stock and Watson, 2002, 2005). In GDFMs, the number of parameters typically grows linearly with the number of time series. A recent review of these five approaches can be found in Peña and Tsay (2020).
Classical linear time series models assume that the (auto-) covariance structure and hence the serial (linear) dependency structure of the data is constant. This may not be a realistic assumption in many applications. For example, many economic data sets show significant structural changes.

Vector-autoregressive (VAR) models with time-varying parameters (TVP) form an important line of research for TVP time series models. They have been successfully applied to model time-varying relationships between multiple economic time series (e.g., in Lubik and Matthes, 2016). Stock and Watson (2009) and Breitung and Eickmeier (2011) found empirical evidence of temporal instability of coefficients in dynamic factor models for large panels of economic data. DFM that are able to deal with time-dependent parameters have been suggested in Del Negro and Otrok (2008) and Su and Wang (2017). Time-variation may also arise in the cointegrating parameter matrix of a vector error-correction model (see, e.g., Bierens and Martins, 2010).

Time-varying parameter is a special case of stochastic parameter models. An indication of the extent of interest in such models is the fact that reviews of related literature appeared as early as 1973 in Rosenberg (1973) followed by Johnson (1977). The estimation of TVP models for time series has been approached from different perspectives. Frequentist methodology includes state-space modeling (see e.g. Chapter 18 in Lütkepohl, 2005; Eickmeier, Lemke, and Marcellino, 2015; and Chapter 13 of Hamilton, 1994) and non-parametric approaches (see e.g. Bierens and Martins, 2010; Su and Wang, 2017; or Coulombe, 2021). Another research direction follows Bayesian estimation, often also in combination with state-space modeling (see, among many others, Baştürk, Hoogerheide, and Dijk, 2017; Del Negro and Otrok, 2008; Doan, Litterman, and Sims, 1984; Lubik and Matthes, 2016; Primiceri, 2005).

Since its introduction by Izenman (1975), reduced-rank regression (RRR) has been successfully applied in multivariate regression modeling, mostly to independent and identically distributed (i.i.d.) data. In this paper, we model a vector of series jointly using reduced-rank regression with dynamics in the parameter matrix and apply classical tools from state-space models. The time-varying parameters are considered as latent states in a Gaussian state-space system. This modeling approach has been used successfully to fit multivariate time series models with time-varying coefficients (see, e.g., Hamilton, 1994, Chapter 13; Durbin and Koopman, 2012, Section 6.3.1; Guo, 1990; Kilian and Lütkepohl, 2017, Chapter 18; Lütkepohl, 2005, Chapter 18).

We model different types of temporal instability by decomposing the reduced-rank coefficient matrix in the multivariate time series regression model as the product of two full-rank time-varying matrices. To the best of our knowledge, the only paper presenting a similar methodology is Yang and Bauwens (2018), who assume the time-varying coefficient matrices live on the Stiefel manifold, the space of semi-orthogonal matrices (see e.g. Chikuse, 2003). Yang and Bauwens (2018)’s approach does not lead to an estimate of all model parameters. In a different context, state-space models with states on the Stiefel manifold were considered by Tompkins and Wolfe (2007) and Bordin and Bruno (2019), where estimation was again not resolved. This restriction appears to impede the development of a feasible estimation algorithm for the parameters, and in consequence, prediction of the responses.

In our Gaussian state-space formulation, the time-varying coefficient matrices live in a Euclidean space instead, which allows for evaluation of the likelihood via the Kalman filter (Shumway and Stoffer, 2017, Ch. 6). Maximum likelihood estimation of all parameters is carried out using an EM-algorithm with closed form updates. We refer to this model, together with the corresponding estimation algorithm, as time-varying reduced-rank regression (TVRRR).

The rest of this paper is organized as follows. We review time-constant and time-varying reduced-rank regression in Sec. 2. We specify our models in Sec. 2.1 and present our estimation and model fitting algorithm in Sec. 2.2. In Sec. 2.3, we define the BIC criterion we use to estimate the rank of the parameter matrix. Section 3 contains our simulation studies and results, which confirm the usefulness of TVRRR in accurately modeling time series data whose relationship changes over time. There, we also show it is particularly robust to structural breaks in the
coefficients. In Sec. 4, we showcase the adaptability of our model by analyzing two data sets. The first consists of seven financial indices, which are considered during two time periods, when international markets experienced extreme stress: the Great Recession in 2008 and the Covid-19 crisis of 2020. Here, we fit vector error correction models (VECM) with time-varying cointegrating relations. As a second example, we consider the number of weekly Covid-19 cases in 12 European countries and fit a TVP-VAR(1) with a reduced-rank coefficient matrix. Section 5 contains our concluding remarks.

2. Time-varying reduced-rank regression

The general reduced-rank regression (RRR) model is given by

\[ y_t = Cx_t + \Gamma u_t + \varepsilon_t, \quad t = 1, \ldots, T, \ldots \]  

where \( y_t \in \mathbb{R}^p \) is a \( p \)-dimensional target vector, \( x_t \in \mathbb{R}^q \) is a \( q \)-dimensional predictor, and \( C \in \mathbb{R}^{p \times q} \) is a matrix with \( \text{rank}(C) = d \leq \min(p, q) \). The latter implies that \( C \) can be decomposed as the product of two rank \( d \) matrices \( z \in \mathbb{R}^{p \times d} \) and \( \beta \in \mathbb{R}^{q \times d} \) such that \( C = z \beta \). The errors \( \varepsilon_t \) are normal white noise with covariance \( \Omega \). The vector \( u_t \in \mathbb{R}^s \) contains additional explanatory variables with unrestricted coefficient matrix \( \Gamma \in \mathbb{R}^{p \times s} \). Both \( x_t \) and \( u_t \) are observable. Throughout, we use boldfaced letters and symbols to denote vectors or matrices and non-boldfaced for scalars. \( A' \) denotes the transpose of a matrix \( A \) and \( I_s \) stands for the \((s \times s)\) identity matrix.

Velu, Reinsel, and Wichern (1986) were the first to use RRR in vector autoregressive (VAR) models. In a VAR(\( m \)) model, the explanatory variables \( x_t \) are lags of the target variable \( y_t \), \( x_t = (y_{t-1}', \ldots, y_{t-m}')' \in \mathbb{R}^{mp} \). While fitting full rank VAR models requires estimation of a large number of parameters if the number of time series \( p \) is large, RRR introduces sparsity in the relationship between the lagged variables and the response by assuming that the coefficient matrix \( C \in \mathbb{R}^{p \times mp} \) has rank \( d \ll p \). Cointegration models are another popular application of RRR in the time series context. In the vector error correction model (VECM), the cointegrating relations of multiple non-stationary time series are represented by a reduced-rank matrix (see e.g. Lütkepohl, 2005, Part II).

RRR may also be simply used as a tool for constructing approximate yet parsimonious regression models. Reinsel and Velu (1998) provide a comprehensive review of reduced-rank regression models and their applications in both i.i.d. and time-series settings. The principle of RRR underlies several methods in sufficient dimension reduction based on inverse regression (Bura and Cook, 2003). For example, parametric inverse regression (Bura and Cook, 2001) and principal fitted components (Cook and Forzani, 2008) rely on RRR methodology.

The most relevant interpretation to our development is of RRR as a factor model. Specifically, \( y_t = \mathbf{z}' \beta \), where \( \mathbf{z} \in \mathbb{R}^{p \times d} \) and \( \beta \in \mathbb{R}^{q \times d} \) with the distinguishing feature that the factors, for given \( \beta \), are observable, as opposed to latent in “standard” factor models. The factors summarize all the information in \( x_t \) that is necessary for approximating \( y_t \). If \( x_t = y_{t-1} \), the model can be interpreted as a simple dynamic factor model, where the factors \( \mathbf{f}_t \) have a VAR(1) dynamic (Peña and Poncela, 2006).

We incorporate time variation in the coefficients of model (1) by writing,

\[ y_t = C_t x_t + \Gamma u_t + \varepsilon_t. \]  

We assume that the rank \( d < \min(p, q) \) of \( C_t \) is fixed and we model the time dependence of \( C_t \) as

(A) \( C_t = \mathbf{z}_t \beta_t \), where \( \mathbf{z}_t \in \mathbb{R}^{p \times d}, \beta_t \in \mathbb{R}^{q \times d} \)

(B) \( C_t = \mathbf{z} \beta_t \), where \( \mathbf{z} \in \mathbb{R}^{p \times d}, \beta_t \in \mathbb{R}^{q \times d} \)
In all three, the component matrices in the decomposition are of full rank $d$. In (A) and (B), one of the two factors in the decomposition of $C_t$ is assumed to be constant. Hence, the time-variation is attributed to the other. This simplifies the interpretation and estimation of the resulting model.

In model (A), it is assumed that the relevant information contained in $x_t$ may be extracted by non time-varying linear combinations of the regressors $x_t$. However, the coefficients of the regression of $y_t$ on $b_0 x_t$ may vary with time. In this case the row space of $C_t$, i.e. the row space of $b_0$, is independent of time. The case of the column space of $C_t$ being constant is modeled by (B) with an analogous interpretation. In model (C), both the column and the row space of $C_t$ are allowed to be time dependent.

Both matrices in the decompositions (A), (B) and (C) can only be identified up to invertible transformations. For model (C), the relation $a_t b_0 t = a_t H_t H_t^{-1} b_0 t$ holds for any invertible $H_t \in \mathbb{R}^{d \times d}$, so that each pair $a_t$, $b_0$ can only be identified up to right multiplication with a nonsingular matrix.

In models (A) and (B), the transformation $H_t = H$ has to be independent of time. For example, in model (A),

\[ C_t = x_t \beta^t = \tilde{x}_t \tilde{\beta}^t \iff \tilde{x}_t = x_t H \text{ and } \tilde{\beta}_t = \beta_t (H^{-1})^t. \]

(6)

Ensuring identifiability requires additional restrictions, such as fixing a $d \times d$ dimensional sub-matrix of $\beta$ to the identity matrix. We demonstrate at the end of this section that the estimates of the proposed algorithm are invariant to transformations as in (6).

In addition to the ambiguity of the component matrices introduced by the time-specific transformations $H_t$, the observations $y_t$ in model (C) are a non-linear function of $x_t$ and $\beta_t$, rendering model fitting and parameter estimation non-trivial. This model cannot be represented as a linear state-space system and one would need to resort to non-linear filtering methods, such as the extended Kalman filter or particle filters, for estimation of the time-varying parameters $x_t$ and $\beta_t$ (see e.g. Särkkä, 2013). On the other hand, models (A) and (B) are sufficiently flexible to capture various types of time-variation. A possible interpretation of model (A) could be a factor model with time-varying factor loadings: The factor composition remains constant, even though the factors’ influence on different response variables can vary over time. With model (B), we can, among other applications, model time-varying cointegration spaces in vector error correction models. Moreover, time-varying factor compositions are covered by this formulation. We focus on models (A) and (B) in the rest of the paper.

### 2.1. Model specification

The reduced-rank regression model (2) with time-varying parameters $C_t$ as defined in (3) and (4), can be written as

\[
\begin{align*}
(A) \quad y_t &= x_t \beta^t x_t + \Gamma u_t + \epsilon_t \\
&= (x_t^t \beta \otimes I_p) \text{vec}(x_t) + \Gamma u_t + \epsilon_t, \\
\end{align*}
\]

(7)

\[
\begin{align*}
(B) \quad y_t &= x_t \beta^t x_t + \Gamma u_t + \epsilon_t \\
&= (x_t^t \otimes x) \text{vec}(\beta_t) + \Gamma u_t + \epsilon_t, \\
\end{align*}
\]

(8)

where the errors are i.i.d. with $\mathbb{E}(\epsilon_t) = 0$ and $\mathbb{Cov}(\epsilon_t) = \Omega$. Both are regression models, where some of the coefficients are time-dependent. These time-varying parameters are modeled as (latent) states in a state space system, as in, e.g. Hamilton (1994).
Both cases (A) and (B) can be treated similarly. We therefore outline the estimation procedure in detail only for model (A). The derivation for model (B) is analogous. In order to simplify the exposition, we let \( \Gamma = 0 \). However, the inclusion of this term is straightforward.

In absence of specific prior knowledge of the dynamics of the time-varying parameters, we assume that they follow a random walk. The time-varying parameters, i.e. states, are assumed to evolve according to some density \( f \) that is centered at the previous state,

\[
\text{vec}(\mathbf{x}_{t+1}) \sim f(\text{vec}(\mathbf{x}_t)),
\]

as in, for example, Del Negro and Otrok (2008), Eickmeier, Lemke, and Marcellino (2015) and Yang and Bauwens (2018).

A common approach is to let \( f \) be a multivariate normal density,

\[
\text{vec}(\mathbf{x}_{t+1}) \sim \mathcal{N}(\text{vec}(\mathbf{x}_t), \Sigma).
\]

The number of parameters \( pd(pd+1)/2 \) in the state covariance matrix \( \Sigma \) is prohibitive for estimation, even for moderate sized models. This fact necessitates the imposition of further structure on \( \Sigma \). As has been noted by Yang and Bauwens (2018), this structure should be compatible with the inherent non-identifiability of the factorization (6). For example, the widely used simplification of a diagonal \( \Sigma \) is not preserved under invertible transformations, since the transformed states \( \mathbf{x}_t \mathbf{H} \) evolve as \( \text{vec}(\mathbf{x}_{t+1}\mathbf{H}) \sim \mathcal{N}(\text{vec}(\mathbf{x}_t \mathbf{H}), (\mathbf{H}' \otimes I)\Sigma(\mathbf{H} \otimes I)) \). In this case, the covariance matrix is non-diagonal in general.

To bypass this limitation and facilitate identifiability of the parameter matrices, Yang and Bauwens (2018) propose a non-linear state space model on the set of all semi-orthogonal matrices, the Stiefel manifold

\[
\mathbb{V}_{p,d} = \{ \mathbf{A} \in \mathbb{R}^{p \times d} : \mathbf{A}' \mathbf{A} = \mathbf{I}_d \}.
\]

In their specification, the state matrices \( \mathbf{x}_t \) perform a random walk on the space (11) by requiring \( f \) in (9) to be the density of a matrix Langevin (or von Mises-Fisher) distribution. This distribution derives from restricting a specific matrix normal distribution to orthogonal matrices (see e.g. Chikuse, 2003, Ch. 2). Although Yang and Bauwens (2018) derive closed form solutions for the predicted and updated states, the resulting posterior likelihoods involve intractable normalizing constants that depend on the model’s time-constant parameters. This renders parameter estimation hard, as reflected in the absence of an estimation procedure. The model cannot be fitted to observed data and filtering the states is only possible if the unknown parameters are given a priori.

We do not require the time-varying parameter matrices \( \mathbf{x}_t \) be semi-orthogonal, but allow them to vary freely in \( \mathbb{R}^{p \times d} \), which results in more flexible parameter estimates. We maintain the multivariate normality assumption (10) and further assume that the covariance matrix \( \Sigma \) is separable; i.e., \( \Sigma = \Sigma_c \otimes \Sigma_r \) with \( \Sigma_c \in \mathbb{R}^{d \times d} \), \( \Sigma_r \in \mathbb{R}^{p \times p} \). This Kronecker product structure is invariant to transformations \( \mathbf{x}_t \rightarrow \mathbf{x}_t \mathbf{H} \), since

\[
\text{Cov}(\text{vec}(\mathbf{x}_t \mathbf{H})) = \text{Cov}((\mathbf{H}' \otimes I_p)\text{vec}(\mathbf{x}_t)) = (\mathbf{H}'\Sigma_c \mathbf{H}) \otimes \Sigma_r.
\]

Furthermore, this structure reduces the number of parameters from \( pd(pd+1)/2 \) to at most \( (p(p+1) + d(d+1))/2 - 1 \). The decomposition \( \Sigma = \Sigma_c \otimes \Sigma_r \) corresponds to the assumption that the states follow a \( (p \times d) \)-dimensional matrix normal distribution (Gupta and Nagar, 2000),

\[
\mathbf{x}_t \sim \mathcal{N}_{p \times d}(\mathbf{x}_{t-1}; \Sigma_r, \Sigma_c),
\]

which agrees with the interpretation of the time-varying coefficients as rank \( d \)-matrices.

A further reduction of the number of parameters that need to be estimated is achieved by setting \( \Sigma_r = I_p \), which further relieves the computational burden for large datasets. Due to this assumption the (changes of the) rows of \( \mathbf{x}_t \) are independent and identically distributed, so that
we only need to estimate \(d(d+1)/2\) parameters, a computationally non-prohibitive number when the rank of the matrix is small. The same applies to model (B). The limitation to this form of matrix normal distribution complies with the matrix Langevin distribution used in Yang and Bauwens (2018), as the latter derives from a matrix normal distribution with the above covariance structure (Chikuse, 2003).

The resulting state-space model for model (A) is thus given by

\[
\begin{align*}
\text{vec}(z_{t+1}) &= \text{vec}(z_t) + \eta_t \\
y_t &= (x_t' \beta \otimes I_p) \text{vec}(z_t) + \Gamma u_t + \varepsilon_t,
\end{align*}
\]

where \(\varepsilon_t \sim \mathcal{N}(0, \Omega)\) and \(\eta_t \sim \mathcal{N}(0, \Sigma_c \otimes I_p)\) are independent i.i.d. sequences. Analogously, model (B) is of the form

\[
\begin{align*}
\text{vec}(\beta'_{t+1}) &= \text{vec}(\beta'_t) + \xi_t \\
y_t &= (x_t' \otimes \beta') \text{vec}(\beta'_t) + \Gamma u_t + \varepsilon_t
\end{align*}
\]

with \(\xi_t \sim \mathcal{N}(0, I_q \otimes \Sigma_c)\).

### 2.2. State estimation and model fitting

The regressors \(x_t, u_t, t = 1, \ldots, T\), are understood as functions of explanatory variables \(z_t\) and lagged dependent variables. That is,

\[
(x_t', u_t')' = g(z_t, y_{t-1}, \ldots, y_{t-m}).
\]

We let \(\mathcal{F}_s\) denote the sigma-algebra generated by all the exogenous variables \(z_t\) and the observed target variables up to time \(s\), \(\mathcal{F}_s = \sigma(z_1, \ldots, z_t, y_{t-m+1}, \ldots, y_0, y_1, \ldots, y_s)\). This setting includes as special cases VAR\(m\), VARX\(m\) and VECM\(m\) models, in which the predictors are either functions of the lagged variables, or of the exogenous variables, or both.

The time-varying parameters \(\beta_t\) and \(z_t\) in models (14), resp. (15), can be inferred using the Kalman filter. Again, we outline the procedure in detail only for model (A). Following Shumway and Stoffer (2017, Ch. 6), we use the following notation for the filtered and smoothed coefficient matrices and the corresponding covariances.

\[
\begin{align*}
\mathbf{x}_t^i &= \mathbb{E}(x_t \mid \mathcal{F}_s) \\
\mathbf{P}_t^i &= \mathbb{C}ov(\text{vec}(z_t) \mid \mathcal{F}_s) \\
\mathbf{P}_{t, t-1}^i &= \mathbb{C}ov(\text{vec}(x_t), \text{vec}(x_{t-1}) \mid \mathcal{F}_s).
\end{align*}
\]

In addition, we assume the noise vectors are jointly normally distributed with

\[
\begin{pmatrix}
\varepsilon_t \\
\eta_t
\end{pmatrix}
\text{i.i.d.} \sim \mathcal{N}\left(\begin{pmatrix}
0 \\
\Omega
\end{pmatrix}, \begin{pmatrix}
\Omega & 0 \\
0 & \Sigma_c \otimes I_p
\end{pmatrix}\right)
\]

and independent of \(\mathcal{F}_0\) and the initial state \(x_0\). Then, the conditional expectations and covariances in (17) can be computed by the usual Kalman filtering and smoothing recursions (Shumway and Stoffer, 2017, Ch. 6). For the initial state, we assume \(\text{vec}(x_0) \mid \mathcal{F}_0 \sim \mathcal{N}(\text{vec}(a_0), \Psi_0)\), providing the starting values \(x_0 = a_0\) and \(P_0 = \Psi_0\) for the Kalman filter and smoother.

The estimated states and the corresponding covariance matrices in (17) depend on the initial values \(a_0, \Psi_0\) and the unknown parameters in system (14). We denote this set by\(^2\)

\(^1\)Analogously, we let \(\text{vec}(\beta'_t) \mid \mathcal{F}_0 \sim \mathcal{N}(\text{vec}(b'_0), \Psi_0)\) for model (B).

\(^2\)For model (B), \(\Theta_0 = \{b_0, \Psi_0, \beta, \Omega, \Sigma, \Gamma\} \).
Table 1. Parameter updates for models (A) and (B), given all other parameters and the smoothed states and covariances.

\begin{align*}
(A) & \quad \alpha_0 \leftarrow \mathbf{x}_0' \\
\Psi_0 & \leftarrow \mathbf{P}_0 \\
\Sigma_t & \leftarrow \frac{1}{T_p} \sum_{i=1}^{\text{rank}(C)} \sigma_i u_i u_i' \\
\text{where } C & = \sum_{t=1}^{T} \left( \text{vec}(\mathbf{x}_t') \text{vec}(\mathbf{x}_t')' + \mathbf{P}_t + \text{vec}(\mathbf{x}_{t-1}') \text{vec}(\mathbf{x}_{t-1}')' + \mathbf{P}_{t-1}' - (\text{vec}(\mathbf{x}_t') \text{vec}(\mathbf{x}_{t-1}')' + \mathbf{P}_{t-1}') \right) \\
& = \sum_{i=1}^{\text{rank}(C)} \sigma_i u_i u_i' \text{ and vec}(U) = u_i u_i' \in \mathbb{R}^{d \times d} \\
\beta' & \leftarrow \text{unvec} \left( \left[ \sum_{t=1}^{T} (x_t x_t' \otimes \Omega^{-1} x_t') + (x_t x_t' \otimes \mathbf{P}_t) \right]^{-1} \left[ \sum_{t=1}^{T} \text{vec}(\mathbf{x}_t' \Omega^{-1} (y_t - \mathbf{u}_t) x_t') \right] \right) \\
\Omega & \leftarrow T^{-1} \sum_{t=1}^{T} (y_t - x_t' \beta' x_t - \mathbf{u}_t') (y_t - x_t' \beta' x_t - \mathbf{u}_t)'
\end{align*}

\begin{align*}
(B) & \quad \beta_0' \leftarrow \mathbf{P}_0' \\
\Psi_0 & \leftarrow \mathbf{P}_0' \\
\Sigma_t & \leftarrow \frac{1}{T_p} \\
\sum_{i=1}^{\text{rank}(C)} \sigma_i u_i u_i' \\
\text{where } C & = \sum_{t=1}^{T} \left( \text{vec}((\beta_t')') \text{vec}((\beta_t')')' + \mathbf{P}_t' + \text{vec}((\beta_{t-1}')') \text{vec}((\beta_{t-1}')')' + \mathbf{P}_{t-1}' - (\text{vec}((\beta_t')') \text{vec}((\beta_{t-1}')')' + \mathbf{P}_{t-1}') \right) \\
& = \sum_{i=1}^{\text{rank}(C)} \sigma_i u_i u_i' \text{ and vec}(U) = u_i u_i' \in \mathbb{R}^{d_q \times q} \\
\alpha & \leftarrow \left[ \sum_{t=1}^{T} (y_t - \mathbf{u}_t) x_t \beta_t' \right] \left[ \sum_{t=1}^{T} (\beta_t' x_t x_t' \beta_t' + (x_t \otimes \mathbf{I}) \mathbf{P}_t' (x_t \otimes \mathbf{I}_d)) \right]^{-1} \\
\Omega & \leftarrow T^{-1} \sum_{t=1}^{T} \left( y_t - \mathbf{u}_t \right) \left( y_t - \mathbf{u}_t \right)' + (x_t \otimes \mathbf{I}) \mathbf{P}_t' (x_t \otimes \mathbf{I}_d) \\
\Gamma & \leftarrow \left[ \sum_{t=1}^{T} (y_t - \mathbf{u}_t) x_t \right] \left[ \sum_{t=1}^{T} u_t u_t' \right]^{-1}
\end{align*}

Note: For better readability the notation does not show dependence on the current parameter estimates $\Theta^{(0)}$. In the update for $\beta'$, the subscript $i:j,k:l$ is used to refer to a submatrix with rows $i$ to $j$ and columns $k$ to $l$. unvec denotes the inverse vec operation, i.e. $a = \text{vec}(A) \iff A = \text{unvec}(a)$.

$$\Theta = \Theta_A = \{ \alpha_0, \Psi_0, \beta, \Omega, \Sigma, \Gamma \}.$$  \hfill (19)

The negative log-likelihood of $Y = \{y_1, ..., y_T\}$ conditional on $\mathcal{F}_0$ is (up to constants) given by

$$-2 \log (L_Y(\Theta)) = \sum_{t=1}^{T} \log |\Sigma_t| + \sum_{t=1}^{T} (y_t - x_t^{-1} \beta' x_t - \mathbf{u}_t) \Sigma_t^{-1} (y_t - x_t^{-1} \beta' x_t - \mathbf{u}_t).$$  \hfill (20)

where $\Sigma_t = (x_t' \beta \otimes \mathbf{I}_p) \mathbf{P}_t^{-1} (x_t' \beta \otimes \mathbf{I}_p) + \Omega$ (Gibson and Ninness, 2005; Shumway and Stoffer, 2017). The log-likelihood (20) is a highly non-linear function of the parameters (19). We bypass its direct minimization using the expectation-maximization (EM) algorithm. Optimization is based on the (conditional) joint likelihood of the data $Y$ and the states $A = \{a_0, ..., a_T\}$, denoted by $L_{Y,A}(\Theta)$.

Given a set of initial values $\Theta^{(0)}$ for the model parameters (19), the EM-algorithm consists of two steps that are iterated in turn until certain stopping criteria are met.

**E-Step:**
Evaluate the conditional expectation of the joint log-likelihood given \( F_T \) as a function of the parameters \( \Theta \), while assuming \( \Theta^{(i)} \) are the “true” parameters:

\[
Q(\Theta|\Theta^{(i)}) = \mathbb{E}_{\Theta^{(i)}}(-2 \ln (L_{Y,A}(\Theta)) \ | \ F_T).
\] (21)

**M-Step:**

Minimize the expected negative log-likelihood in \( \Theta \), i.e. find

\[
\Theta^{(i+1)} = \arg\min_{\Theta} Q(\Theta|\Theta^{(i)}).
\] (22)

The E-step includes smoothing of the latent states given the current set of parameters and starting values. The expected log-likelihood (21) for model (A) is given by

\[
Q(\Theta|\Theta^{(i)}) = \ln |\Psi_0| + \text{tr}(\Psi_0^{-1}[P_0^T + \text{vec}(x_0^T - a_0]\text{vec}(x_0^T - a_0)]) + Tp \ln |\Sigma_c| + \text{tr}\left((\Sigma_c \otimes I_p)^{-1}\sum_{t=1}^{T}(\text{vec}(x^T_t)\text{vec}(x^T_t)' + P^T_t + \text{vec}(x^T_{t-1})\text{vec}(x^T_{t-1}') + P^T_{t-1} - \text{vec}(x^T_t)\text{vec}(x^T_{t-1}'))) - T \ln |\Omega| + \text{tr}\left(\Omega^{-1}\sum_{t=1}^{T}(y_t'y_t' - y_t'x_t'\beta x_t' - x_t'\beta'y_t' + (x_t'\beta \otimes I_p)(\text{vec}(x^T_t)\text{vec}(x^T_t)' + P^T_t)(x_t'\beta \otimes I_p)'))) \right.
\] (23)

The smoothed states \( x^T_t \) and the covariance matrices \( P^T_t \) and \( P^T_{t,t-1} \) in (23) depend on the current set of parameters \( \Theta^{(i)} \). However, for better readability, this dependence is not reflected in the notation.

For the M-step we use a simple iterative scheme, where we, in turn, compute the minimizer of the expected log-likelihood with respect to one of the parameters from (19) while keeping the others fixed. The minimizers for \( a_0 \), \( \Psi_0 \), and \( \Sigma_c \) are independent of the other parameters. Inspection of (23) reveals that, for model (A), the updates for \( \beta \), \( \Omega \) and, if included, \( \Gamma \), depend on each other. Nevertheless, it is straightforward to compute the minimizer for one of these parameters while keeping the other two fixed. The explicit formulas for the above minimizers are given in Table 1. The updates within each M-step are iterated until the changes of the parameter matrices are negligible, which is achieved fast in our simulations. The EM-algorithm terminates if either the change in the parameter estimates, or the relative change in the data log-likelihood (20) is negligible.

By construction, the EM-algorithm is guaranteed to increase the data likelihood in every step of the estimation procedure (see e.g. Casella and Berger, 2002, Thm. 7.2.20.). If it converges, it reaches a stationary point; i.e., a local maximum of the data likelihood.

The non-identifiability of the decompositions (A) and (B) in (6) does not interfere with the estimation algorithm in the sense that the estimates for the matrices \( C_t \) and the predictions for \( y_{t+1} \) do not depend on the chosen factorization. For example, consider model (A) and an arbitrary nonsingular matrix \( H \) and define \( \tilde{x}_t = x_tH \) and \( \tilde{\beta} = \beta H^{-1}' \). Then, the corresponding state-space system

\[
\text{vec}(\tilde{x}_{t+1}) = \text{vec}(\tilde{x}_t) + \tilde{\eta}_t, \\
y_t = (x_t'\tilde{\beta} \otimes I_p)\text{vec}(\tilde{x}_t) + \varepsilon_t
\] (24)

is equivalent to the (original) state space system (14) if we transform the noise covariances and the starting values for the Kalman filter correspondingly. This means that the filtered and
smoother and the corresponding covariances of these two systems are related to each other by the same transformation. That is, \( \tilde{\mathbf{x}}_t = \mathbf{x}_t \mathbf{H} \) and \( \tilde{\mathbf{P}}_t = (\mathbf{H}' \otimes \mathbf{I}_{p}) \mathbf{P}_t (\mathbf{H} \otimes \mathbf{I}_{p}) \) for \( s \in \{ t - 1, t, T \} \) and all \( t = 1, \ldots, T \). This observation also carries over to the EM-updates in Table 1. For example, the relationships \( \tilde{\beta}^{(j)}_t = \beta^{(j)}_t \mathbf{H}^{-1} \), \( \tilde{\Sigma}_t = \mathbf{H}' \Sigma_t \mathbf{H} \), and \( \tilde{\Omega}^{(j)} = \Omega^{(j)} \) hold for the estimates of models (24) and (14).

### 2.2.1. Choice of starting values

The choice of starting values \( \Theta^{(0)} \) used for numerical optimization of the likelihood can have a strong influence on the resulting model fit due to local optima. However, our EM-algorithm is robust to the choice of starting values in our experiments.

We start with the initialization of the coefficient matrices \( \mathbf{a}_0 \) and \( \beta \) for model (A), and respectively \( \mathbf{x} \) and \( \mathbf{b}_0 \) for model (B). We denote the starting values for both cases by \( \mathbf{x}^{(0)} \) and \( \beta^{(0)} \), and determine them by time-constant RRR, as follows. We let \( \mathbf{x}^{(0)} = \mathbf{x}_{RR} \), and \( \beta^{(0)} = \beta_{RR} \) where \( \mathbf{x}_{RR} \) and \( \beta_{RR} \) are minimizers of the criterion

\[
\min_{\mathbf{a}, \beta} \sum_{t=1}^{T} (y_t - \mathbf{a}\beta' \mathbf{x}_t)' \mathbf{W}(y_t - \mathbf{a}\beta' \mathbf{x}_t)
\]  

(25)

for any positive definite, symmetric matrix \( \mathbf{W} \). The solutions to (25) give a rank \( d \)-approximation of the ordinary least squares (OLS) regression matrix \( \mathbf{C}_{OLS} \) (Reinsel and Velu, 1998, Ch. 2). A close look at the OLS estimate reveals why the above starting values are reasonable. Assuming that \( \mathbb{E}(\mathbf{e}_t \mid \mathbf{x}_1, \ldots, \mathbf{x}_T) = 0 \), the conditional expectation of the OLS estimate given the predictors is

\[
\tilde{\mathbf{C}} = \mathbb{E}(\mathbf{C}_{OLS} \mid \mathbf{x}_1, \ldots, \mathbf{x}_T) = \sum_{t=1}^{T} \mathbf{C}_t \mathbf{x}_t \mathbf{x}_t' \left( \sum_{i=1}^{T} \mathbf{x}_i \mathbf{x}_i' \right)^{-1}.
\]  

(26)

This conditional expectation is a weighted average of the time-varying matrices \( \mathbf{C}_t \). Thus, its reduced-rank decomposition yields “averages” of the time-varying coefficients \( \mathbf{x}_t \) (resp. \( \beta_t \)). In particular, for model (B), \( \text{span}(\mathbf{C}_t) = \text{span}(\mathbf{x}) \) and thus also \( \text{span}(\tilde{\mathbf{C}}) = \text{span}(\mathbf{x}) \approx \text{span}(\mathbf{x}_{RR}) \), hinting that \( \mathbf{x}_{RR} \) is a reasonable starting value for \( \mathbf{x} \). If pronounced time-variation is suspected, it may be advisable to obtain starting values from a reduced-rank regression that only uses the first \( \tau < T \) observations of the time series.

The initial value for the error covariance \( \mathbf{\Omega} \) is set to the residual covariance of the time-constant RRR,

\[
\mathbf{\Omega}^{(0)} = T^{-1} \sum_{t=1}^{T} (y_t - \mathbf{x}^{(0)} \beta^{(0)}' \mathbf{x}_t)(y_t - \mathbf{x}^{(0)} \beta^{(0)}' \mathbf{x}_t)'.
\]

If the number of parameters is large compared to the sample size, it can be restricted to be diagonal or to be a multiple of the identity \( \mathbf{\Omega} = \sigma^2 \mathbf{I}_p \). The updates for the EM-algorithm can be adjusted accordingly. Simulations imply that assuming a diagonal covariance matrix does not necessarily decrease the accuracy of the estimation of the states and the prediction performance of the model. If we wish to include additional predictors \( \mathbf{u}_t \), the starting values for \( \mathbf{\Gamma} \) and correspondingly adjusted starting values for \( \mathbf{x} \), \( \beta \) and \( \mathbf{\Omega} \) can be obtained from reduced-rank regression with an additional full rank coefficient matrix as derived in (Reinsel and Velu, 1998, Ch. 3). In our simulations, the choice of \( \mathbf{W} \) in (25) did not have a strong influence on the parameter estimates. We used either the identity \( \mathbf{I}_p \) or the least squares residual covariance.

The uncertainty about the initial states \( \mathbf{x}_0 \), resp. \( \beta_0 \), is represented in the choice of the initial covariance matrix \( \mathbf{\Psi}_0 \) : The larger the covariance, the more flexibility and uncertainty about the initial state is allowed for.
Finally, there is no obvious choice for starting values of the state covariance matrix $\Sigma_c$. Generally, the magnitude of the entries of $\Sigma_c$ governs the flexibility of the resulting state estimates—the basic paths of the states will be similar regardless of the magnitude of $\Sigma_c$. Our simulation studies indicate that different starting values for $\Sigma_c$ only marginally influence the value of the likelihood and the residual variance at the time of convergence. Nevertheless, it is advisable to try different values and check for stability of the resulting estimates.

2.3. Rank selection

Rank selection for the $\text{tv} \text{RRR}$ model can be achieved by information criteria. We use the Bayesian information criterion (BIC), first introduced by Schwarz (1978), and given by

$$\text{BIC}(d) = -2 \log (L_Y(\Theta)) + \log (T) \, K_d,$$

Figure 1. The top panel(s) show a trajectory of the five time series $\{y_i\}, i = 1, \ldots, 5$ generated by the VAR(1) model (28) with a structural break of the coefficient matrix at $t = 50$. The bottom panel shows the corresponding estimation error $\text{err}_t = \|C_t - C_t\|_F/\|C_t\|_F$ for the three estimates considered: oracle RRR, time-constant RRR and $\text{tv} \text{RRR}$. 

$\text{B. BRUNE ET AL.}$
Table 2. Coefficient matrices in the simulation studies for models (A) and (B).

| Name           | Model | Specification of matrices                                                                 | Parameters |
|----------------|-------|-------------------------------------------------------------------------------------------|------------|
| Fixed          | A     | $\tilde{x}_t \equiv x \forall t = 1, \ldots, T \sim \text{USTiefel}(p, d)$                | –          |
|                | B     | $\tilde{b}_t \equiv \beta \forall t = 1, \ldots, T \sim \text{USTiefel}(q, d)$          | –          |
| Deterministic  | A     | Deterministic linear transition between two matrices                                        | Shift $s \in \{1, 1.5, 2\}$ |
|                |       | $x_1, x_2 \sim \text{USTiefel}(p, d)$ such that $x_t = s^{-1}x_1 + s^{-1}x_2$             |            |
|                | B     | Deterministic linear transition between two matrices                                        | Shift $s \in \{1, 1.5, 2\}$ |
|                |       | $\tilde{b}_1, \tilde{b}_1 \sim \text{USTiefel}(q, d)$ such that $\tilde{b}_t = s^{-1}\tilde{b}_1 + s^{-1}\tilde{b}_2$ |            |
| Random walk    | A     | Matrices follow a random walk, i.e.                                                        | $\Sigma_c = s \cdot \Sigma$ |
|                |       | $x_t \sim N_{a,d}(\tilde{x}_t, I_p, \Sigma_c)$                                            | where $s \in \{0.1, 1, 2\}$ |
|                | B     | Matrices follow a random walk, i.e.                                                        | $\Sigma_c = s \cdot \Sigma$ |
|                |       | $b_t \sim N_{a,d}(\tilde{b}_t-1; I_q, \Sigma_c)$                                          | where $s \in \{0.1, 1, 2\}$ |
|                |       | where $\Sigma_1 = 0.005$, $\Sigma_2 = \begin{pmatrix} 0.005 & 0.0025 \\ 0.0025 & 0.005 \end{pmatrix}$, $\Sigma_3 = \begin{pmatrix} 0.0025 & 0.0025 \\ 0.0025 & 0.0025 \\ 0.00125 & 0.0025 & 0.005 \end{pmatrix}$ |

Note: The time-constant parts are drawn as $\tilde{\beta} \sim \text{USTiefel}(q, d)$, resp. $\tilde{x} \sim \text{USTiefel}(p, d)$.

where $\hat{\Theta}$ denotes the maximum likelihood estimates (computed with the EM algorithm) and $K_d$ is the number of free parameters of the model. The estimate for the rank is the minimizer of this BIC criterion (27). For model (A), $\Sigma_c$ has $d(d + 1)/2$ free parameters, and $\tilde{\beta}$ has $q \cdot d$ free parameters, so that we set $K_d = d(d + 1)/2 + q \cdot d$. Similarly, $K_d = d(d + 1)/2 + p \cdot d$ for model (B). We do not count the parameters in the matrices $\Omega$ and $\Gamma$, since the dimension of these matrices does not depend on $d$. We also do not count the parameters for the distribution of the initial state $a_0$ (resp. $b_0$) and $\Psi_0$.

Another popular criterion for model selection is Akaike’s information criterion (AIC; Akaike, 1973), where $\log(T)$ in (27) is replaced by 2. However, it has been shown in similar, time-constant settings [e.g. Aznar and Salvador (2002) for the choice of the cointegrating rank in vector error correction models; Bai and Ng (2002) for the number of factors in dynamic factor models] that AIC results in inconsistent rank estimates as the penalty term needs to tend to infinity at a certain rate to achieve consistency. Our simulation studies indicate that AIC does not lead to consistent estimates in our model. Specifically, its performance deteriorates with increasing sample size.

3. Simulation studies

In this section, we illustrate the filter performance via a toy example, and demonstrate the flexibility and usefulness of TTVRRR in different settings in a simulation study.

3.1. Illustration of filter performance

We generate data $y_t$ from a five-dimensional VAR(1)-model with rank $d = 2$ coefficient matrix,

$$y_t = \tilde{x}_t \tilde{\beta}'y_{t-1} + \epsilon_t, \quad \text{where } \tilde{x}_t = \begin{cases} \tilde{x}_1, & \text{for } t \leq 50; \\ \tilde{x}_2, & \text{for } t > 50 \end{cases}; \quad \epsilon_t \sim N(0, I_5),$$

where $\tilde{x}_1$, $\tilde{x}_2$ and $\tilde{\beta}$ are random orthonormal matrices. Model (28) represents a structural break at time $t = 50$. The realizations of the five time series are displayed in the top panel of Fig. 1. The structural break at time $t = 50$ induces a considerable change in the autocovariance structure of the time series.
We fit a TVRRR model using the estimation procedure in Sec. 2.2. The estimated coefficient matrices are compared to those obtained from time-constant RRR and “oracle” RRR. For the latter, we assume the location of the structural break is known and fit two separate RRR models. We assess the accuracy of the estimate \( \beta_{C_t} = a_t \beta_0 \) of the coefficient matrix \( C_t = a_t \beta_0 \) with the normalized Frobenius norm of the estimation error, 
\[
\text{err}_{C_t} := \frac{||C_t - \hat{C}_t||_F}{||C_t||_F}.
\]

(29)

Its values are displayed in the bottom panel of Fig. 1. The TVRRR coefficient estimates (---) are very close to those of oracle RRR (___). At the structural break, the error jumps up rapidly, yet the coefficient estimates quickly recover and roughly line up with those of oracle RRR. As expected, time-constant RRR errors (—) are the largest as the method simply averages over the two matrices \( a_1 \) and \( a_2 \) (as can be seen from Eq. (26)). This toy example also indicates the EM-algorithm we use to fit TVRRR is robust to the choice of the starting values, which were obtained from RRR, thus far from the truth.

3.2. Simulation

The data for our simulation study are drawn from the model equations

\[
(A) \quad y_t = \alpha_t \beta x_t + \epsilon_t, \quad \text{and} \quad (B) \quad y_t = \alpha_t \beta x_t + \epsilon_t.
\]

(30)

The predictors \( x_t \in \mathbb{R}^q \) are specified as exogenous predictors, drawn from \( q \) independent MA(2) processes. The errors \( \epsilon_t \) are normally distributed with mean zero and diagonal covariance matrix \( \Omega \). The diagonal entries of \( \Omega \) are drawn from the uniform distribution on the interval \([0.5, 1.5]\). Starting values for the EM-algorithm are obtained from time-constant reduced-rank regression as described in Sec. 2.2.1. The weight matrix \( W \) in (25) is selected as the residual covariance from the ordinary least squares regression of \( y_t \) on \( x_t \). The initial state covariance \( W_0 \) is set to \( W_0 = 1000 \cdot I_{pd} \) for model (A), and \( W_0 = 1000 \cdot I_{qd} \) for model (B). The large variances reflect the uncertainty about the initial values used for the Kalman filter. We also conducted experiments where the data follow a VAR(1) process with time-varying coefficients, i.e. we set \( x_t = y_{t-1} \) in (30). Due to the random walk generation of the coefficient matrices, the time series often diverge. For this reason, we only report results for exogenous \( x_t \). However, when the VAR time series remain bounded, our method performs well.

To study the ability of the proposed method to capture different types of time variation, we generate datasets under three different types of time-varying coefficients. They are displayed in Table 2. The notation \( \alpha \sim \text{UStiefel}(p, d) \) indicates that \( \alpha \) is drawn from the uniform distribution on the Stiefel manifold \( \mathbb{V}_{p,d} \) (see (11)) as defined in Chikuse (2003).
The fixed scenario was included in order to evaluate the robustness of TVRRR in the case of a true RRR model with time-invariant parameters. The deterministic transition produces a gradual change of the parameters from the starting to the terminal coefficient matrix. The random walk agrees with the modeling assumption that underlies our procedure and represents random fluctuations in the coefficients. The matrices \( R_d \) depend on the state dimension \( d \) and are selected as multiples of the matrices at the bottom of Table 2.

The dimension of the target time series vector takes values \( p \in \{5, 10, 20\} \), the predictors are generated with dimension \( q \in \{5, 10, 20\} \), and the rank takes values \( d \in \{1, 2, 3\} \). Each time series has length \( T = 200 \). The validation time series of length \( s = 200 \) is generated by extrapolating the data generating processes in Table 2. For each combination of parameters, we draw 10 sets of coefficient matrices, and repeat the model fitting procedure 10 times, resulting in a total of \( 10 \times 10 = 100 \) simulation runs per setting.

The in-sample MSE for model (A) in (14) is

\[
MSE_{is} = \frac{1}{T p} \sum_{t=1}^{T} (y_t - \tilde{x}_t \tilde{\beta}_t^{'} x_t)^{'} (y_t - \tilde{x}_t \tilde{\beta}_t^{'} x_t). \tag{31}
\]

The corresponding out-of-sample MSE is obtained from one-step-ahead predictions for the validation data using the Kalman filter and the fitted parameter matrices.

| \( s \) | \( d \) | \( MSE \) (RRR) | \( MSE \) (TVRRR) | \( err_{C_t} \) (RRR) | \( err_{C_t} \) (TVRRR) |
|-----|-----|------------|------------|----------------|----------------|
| A   | 1   | In sample | 1.0096 (0.1119) | 0.9151 (0.1028) | 0.3793 (0.1383) | 0.2894 (0.0879) |
|     |     | Out of sample | 1.6575 (0.2608) | 1.0450 (0.1115) | – | – |
|     | 2   | In sample | 1.0801 (0.0607) | 0.8848 (0.0680) | 0.3770 (0.1202) | 0.2831 (0.0741) |
|     |     | Out of sample | 2.3940 (0.3604) | 1.1295 (0.0783) | – | – |
|     | 3   | In sample | 1.0660 (0.0779) | 0.8049 (0.0634) | 0.3351 (0.0531) | 0.2548 (0.0329) |
|     |     | Out of sample | 2.9252 (0.4102) | 1.1117 (0.0980) | – | – |
| 1.5 | 1   | In sample | 1.1022 (0.1054) | 0.9334 (0.0590) | 0.2608 (0.1483) | 0.1502 (0.0646) |
|     |     | Out of sample | 2.1126 (0.4480) | 1.0875 (0.0734) | – | – |
|     | 2   | In sample | 1.0666 (0.1011) | 0.7717 (0.0777) | 0.2096 (0.0474) | 0.1141 (0.0163) |
|     |     | Out of sample | 3.4841 (0.6698) | 1.0365 (0.0978) | – | – |
|     | 3   | In sample | 1.2539 (0.0764) | 0.7955 (0.0802) | 0.2670 (0.0665) | 0.1405 (0.0246) |
|     |     | Out of sample | 4.6271 (0.7045) | 1.1954 (0.1063) | – | – |
| 2   | In sample | 1.0805 (0.0891) | 0.8659 (0.0791) | 0.1671 (0.0493) | 0.0785 (0.0180) |
|     |     | Out of sample | 2.4777 (0.4529) | 1.0386 (0.0855) | – | – |
|     | In sample | 1.3422 (0.1467) | 0.8173 (0.0909) | 0.2651 (0.0767) | 0.0975 (0.0170) |
|     |     | Out of sample | 6.3858 (1.0800) | 1.1556 (0.1121) | – | – |
|     | 3   | In sample | 1.4983 (0.1434) | 0.7842 (0.0646) | 0.2706 (0.0411) | 0.1079 (0.0150) |
|     |     | Out of sample | 8.3580 (1.3044) | 1.2588 (0.0982) | – | – |
| B   | 1   | In sample | 1.0672 (0.0998) | 0.9662 (0.0904) | 0.4490 (0.1771) | 0.3457 (0.1020) |
|     |     | Out of sample | 1.6924 (0.3180) | 1.0948 (0.0925) | – | – |
|     | 2   | In sample | 1.0602 (0.0793) | 0.9215 (0.0789) | 0.3635 (0.0786) | 0.2677 (0.0284) |
|     |     | Out of sample | 2.1380 (0.2809) | 1.1122 (0.0917) | – | – |
|     | 3   | In sample | 1.0636 (0.1344) | 0.8826 (0.1219) | 0.3315 (0.0644) | 0.2677 (0.0467) |
|     |     | Out of sample | 2.8633 (0.2484) | 1.1558 (0.1437) | – | – |
| 1.5 | 1   | In sample | 1.0876 (0.0913) | 0.9661 (0.0662) | 0.3587 (0.2036) | 0.2154 (0.1228) |
|     |     | Out of sample | 2.4989 (0.8415) | 1.0957 (0.0808) | – | – |
|     | 2   | In sample | 1.1106 (0.1144) | 0.8594 (0.0984) | 0.3262 (0.0976) | 0.1807 (0.0381) |
|     |     | Out of sample | 3.9585 (0.7443) | 1.0933 (0.1091) | – | – |
|     | 3   | In sample | 1.2790 (0.0800) | 0.8965 (0.0693) | 0.3382 (0.0807) | 0.1842 (0.0315) |
|     |     | Out of sample | 5.3404 (0.7817) | 1.2561 (0.0886) | – | – |
| 2   | 1   | In sample | 1.0893 (0.0906) | 0.9098 (0.0762) | 0.3070 (0.1289) | 0.1582 (0.0455) |
|     |     | Out of sample | 3.6538 (0.9878) | 1.0592 (0.0930) | – | – |
|     | 2   | In sample | 1.2554 (0.0794) | 0.8661 (0.0715) | 0.3119 (0.0762) | 0.1303 (0.0258) |
|     |     | Out of sample | 6.5400 (1.0845) | 1.1381 (0.0772) | – | – |
|     | 3   | In sample | 1.3520 (0.0754) | 0.8032 (0.0271) | 0.2873 (0.0534) | 0.1235 (0.0191) |
|     |     | Out of sample | 8.3886 (1.5862) | 1.2044 (0.0295) | – | – |
\[ \text{MSE}_{\text{soa}} = \frac{1}{tp} \sum_{t=1}^{T+t} (y_i - \alpha_t^{-1} \beta' x_t)'(y_i - \alpha_t^{-1} \beta' x_t). \] (32)

We assess the accuracy of the estimate \( \hat{C}_t = \alpha_t \tilde{\beta} \) with the normalized Frobenius distances of the estimated coefficient matrices in (29). For model (B), these criteria are defined in an analogous manner. Throughout the simulation study, the performance is juxtaposed to that of the regular reduced-rank regression reference model.

### 3.2.1. Results

We report results for \( p = q = 10 \). The respective tables for other pairs of \( p \) and \( q \) are comparable and can be found in the supplementary material.

First, we present results for the fixed data generating process. Table 3 reports the in- and out-of-sample MSE in (31) and (32), respectively, and relative estimation errors (29) for RRR and TVRRR in models (A) and (B). Under this setting, simple RRR is more appropriate than TVRRR. This results in larger relative estimation errors \( \text{err}_{C_t} \) even though it seems that forecasting precision is only mildly affected. The in-sample MSE of TVRRR is smaller whereas the out-of-sample MSE is slightly larger than the respective MSEs of RRR in both models. TVRRR is more flexible than the time-constant RRR. As a consequence, it tends to slightly overfit the training data, resulting in a lower in-sample MSE and a higher out-of-sample MSE in the validation data.

Next, we report the simulation results for the deterministic transition of the coefficients as specified in Table 2. The shift (\( s \)) is set to 1 (“small”), 1.5 (“medium”) and 2 (“large”). As can be seen in Table 4, this setting incurs a significant loss in estimation precision for RRR with increasing variation of the parameter matrices. While the in-sample fit is not much affected, out-of-sample prediction fails. In contrast, TVRRR is uniformly better than RRR. It has slightly worse out-of-sample than in-sample MSE, while both remain roughly stable across all setting combinations. Moreover, the coefficient estimation accuracy of TVRRR improves as the shift between the starting and final coefficient matrices increases. This effect is probably partly due to the relative error criterion. The estimation accuracy of RRR stays roughly the same across all settings.

![Figure 2](image-url)

Figure 2: \( \text{err}_{C_t} \) for ten simulation runs with the same set of coefficient matrices from the deterministic setting, RRR in orange (dashed), TVRRR in teal (solid), \( p = q = 10, \ d = 1 \), large variation, model A.
The difference in performance between RRR and TVRRR can be explained by considering the errors of the coefficient matrix estimates over time in Fig. 2. RRR averages over the coefficient matrices (see also the discussion on the starting values, Eq. (26)), and thus the smallest distance
is achieved at approximately $T/2$. This estimate is far off for time points that are closer to the start and end of the time series, and deviate even further for the validation data set $t = 201, ..., 400$. On the other hand, the estimation errors of TVRRR are consistently small throughout the time range. In particular, since the RRR estimates serve as starting values for the EM-algorithm, this further illustrates the robustness of TVRRR to "wrong" starting values.

Table 5 displays the respective results for the random walk coefficient evolution in Table 2. Even for the low variation random walk, we observe the failure of RRR both in terms of MSE and of estimation accuracy. Both in- and out-of-sample MSEs are significantly larger than the corresponding values for TVRRR. As in the deterministic setting, TVRRR gains from increasing variation in the coefficients.

### 3.3. Rank selection

We assess the performance of the BIC criterion (27) for rank selection in both models (A) and (B) and three scenarios for time varying parameters. In addition to the deterministic transition (with $s = 1$) and the random walk transition (with $s = 0.1$), as defined in Table 2, we also consider a structural break setting similar to (28): For model (A), we draw two coefficient matrices $\mathbf{a}_1, \mathbf{a}_2 \sim U\text{Stiefel}(p, d)$ and let $\mathbf{a}_i = \mathbf{a}_1$ for $i \leq \frac{T}{2}$ and $\mathbf{a}_i = \mathbf{a}_2$ for $i > \frac{T}{2}$. For model (B) we proceed analogously. We look at time series of length $T \in \{100, 200, 300\}$ and the same sets of values for $p, q$ and $d$ as before. Again, we carry out $10 \cdot 10 = 100$ repetitions of each setting.

The simulation results indicate that the BIC criterion is not sensitive to the coefficient transitions we consider. Table 6 reports the fraction of simulation runs where the rank was correctly estimated in both models for $p = q = 10$. Results for other combinations of $p$ and $q$ are comparable and can be found in the supplement. BIC rarely overestimates the true rank. For smaller samples, it
frequently underestimates the rank, but drastically improves as the sample size increases. Rank estimation seems to be the least successful for the random walk coefficient transition, where the tendency to underestimate the rank of the model is the strongest. However, the performance of the criterion also in this scenario markedly improves with increasing the length of the time series.

4. Data examples

We apply TVRRR and an adaptive version of RRR that we define in this section on two data sets. The first consists of seven main stock indices, and the second contains Covid-19 cases for 12 European countries.

4.1. Stock index data

We perform an out-of-sample forecasting experiment with stock index data. We retrieved closing prices for DAX Performance Index (DAX), Dow Jones Industrial Average (DJI), Hang Seng Index (HSI), Nikkei 225 (N225), NASDAQ Composite (NASDAQ), and S&P 500 (S&P 500) and TSE Weighted Index (TWII) for two different time periods. The first period ranges from January 15, 2008 to May 18, 2009, and the second from June 05, 2019 to October 05, 2020. Both time series include abrupt changes in global economy. The first period contains the Great Recession and the second the Covid-19 pandemic. While it is thought that stock index data are cointegrated, the relationships of the time series might have changed during both crises, possibly due to markets recovering from the shocks at different rates.

We fit a lag-1 vector error correction model (VECM) with possibly time-varying cointegrating relations $\beta_t y_{t-1}$,

$$\Delta y_t = \alpha \beta_t y_{t-1} + \Gamma \Delta y_{t-1} + \epsilon_t.$$  

Prior to the analysis, the stock index time series are logarithmized, centered and scaled so that the sample mean equals zero and the variance is equal to one in the training data set. Missing values are imputed using na_kalman from the imputeTS R-package (Moritz and Bartz-Beielstein, 2020).

Table 7. In- and out-of-sample MSEs of TVRRR, RRR and updated RRR for the 2008 and 2020 stock data sets, relative to the MSEs for the naive predictions evolving from the random walk assumption (34).

|          | DJI     | DAX     | HSI     | N225    | NASDAQ  | S+P 500 | TWII    |
|----------|---------|---------|---------|---------|---------|---------|---------|
| In-sample |         |         |         |         |         |         |         |
| TVRRR    | 0.9424  | 0.6555  | 0.4442  | 0.4641  | 0.9599  | 0.9509  | 0.6941  |
| RRR      | 0.8018  | 0.7035  | 0.5301  | 0.4938  | 0.8142  | 0.8101  | 0.7216  |
| Out-of-sample |     |         |         |         |         |         |         |
| TVRRR    | **0.9439** | 1.0503  | 1.0691  | **0.6923** | **0.9444** | 0.9517  | 0.9265  |
| RRR      | 3.3134  | 1.6374  | 2.0129  | 2.2012  | 3.8061  | 3.3488  | 1.6735  |
| RRR (upd.) | 1.1229  | 0.9969  | 0.7350  | 0.3340  | 0.7729  | 0.7269  | 0.5940  |

|          | DJI     | DAX     | HSI     | N225    | NASDAQ  | S+P 500 | TWII    |
|----------|---------|---------|---------|---------|---------|---------|---------|
| In-sample |         |         |         |         |         |         |         |
| TVRRR    | 0.8861  | 0.9691  | 0.7350  | 0.3340  | 0.7729  | 0.7269  | 0.5940  |
| RRR      | 0.8092  | 0.8819  | 0.8156  | 0.5048  | 0.7986  | 0.7749  | 0.6649  |
| Out-of-sample |     |         |         |         |         |         |         |
| TVRRR    | **0.9733** | 0.9997  | 1.0711  | **0.8085** | 0.9555  | 0.9648  | 0.9423  |
| RRR      | 1.9315  | 2.3440  | 3.5379  | 4.4262  | 2.4440  | 2.4903  | 2.5606  |
| RRR (upd.) | 1.3869  | 1.4096  | 1.3999  | 1.1272  | 1.3891  | 1.4010  | 1.0368  |

Note: For the out-of-sample MSEs, boldface numbers indicate significantly better prediction performance than the naive predictions according to the Diebold-Mariano test, and numbers in italics indicate significantly better performance of TVRRR than updated RRR at 0.05 significance level.

Table 7. In- and out-of-sample MSEs of TVRRR, RRR and updated RRR for the 2008 and 2020 stock data sets, relative to the MSEs for the naive predictions evolving from the random walk assumption (34).

3All time series retrieved from https://finance.yahoo.com/, accessed March 01, 2021.
We fit the models to the first 150 observations of the time series, and, using the estimated coefficients, predict the following 200 observations.

We compare the following three model fitting approaches: TVRRR, time-constant RRR and time-constant updated RRR. For TVRRR, we compute the one-step ahead predictions with the Kalman filter. However, for the estimation of the model parameters, we only use the training data. These predictions are compared to two versions of the time-constant VECM. For the first, we fit the model to the first 150 observations and obtain predictions for the following 200 data points using the estimated coefficients (time-constant RRR). The second, fairer comparison model is a VECM that is refitted with every new observation of the prediction period to include all the past information in the prediction (updated RRR).

The MSEs from in-sample and out-of-sample prediction of the time series are reported relative to benchmark (“naive”) predictions obtained from the assumption that stock index data follow a random walk (Fama, 1965; Kendall, 1953):

\[
\begin{align*}
\hat{y}_{t+1} &= y_t \\
\Delta y_{t+1} &= \hat{y}_{t+1} - y_t = 0.
\end{align*}
\]

We start with the analysis of the 2008 data. The seven time series before and after scaling are displayed in the left top and bottom panels of Fig. 3. The structural break in September 2008 is evident in both, with the scaled data capturing it better. The decline of the prices was also accompanied by an increased volatility, which can be clearly seen in the plots of the returns (log first differences). The training period is January 15, 2008 to August 11, 2008. When fitting the TVRRR model, the rank selected by the BIC criterion is \(d = 2\). The EM-algorithm converges after 54 iterations, and the likelihood reaches a plateau fairly quickly.

The relative in- and out-of-sample MSEs for both versions of time-constant RRR and TVRRR are reported in the top panel of Table 7. The performance of RRR and TVRRR is comparable in-sample. However, we are more interested in the out-of-sample performance of the methods. We thus have applied the (modified) Diebold-Mariano test (Harvey, Leybourne, and Newbold, 1997) to check, whether there are significant differences in the forecasting performance of the naive method, updated RRR and TVRRR. For the 2008 data, TVRRR yields significantly better predictions than the naive prediction in four cases, and it beats updated RRR for three series. Conversely, the prediction performance of TVRRR is never significantly worse than that of the competitors. Its capability to adapt even to abrupt changes in the coefficient estimates while retaining estimation precision when the parameters are constant pays off: The coefficient estimates adapt to the structural break.
As expected, RRR fails completely at prediction. The different behavior of the methods may also be seen in the plot of the $d = 2$ estimated cointegrating relationships $b_{1t}y_{t-1}$ and $b_{2t}y_{t-1}$ in Fig. 4, where $b_{it}$ denotes the $i$th row of $b_t$. The RRR components (top) show a trending behavior. In contrast, the cointegrating relationships inferred using TVRRR (middle) only vary around zero, but they clearly display the increased volatility during the crisis. Compare also the plot of the log first-differences.

We next analyze the 2020 data. The corresponding time series are displayed in Fig. 5. Here, the structural break is even more abrupt. Whereas the time series decreased slowly in 2008, the 2020 series exhibit a sudden drop around March 2020, with a fairly quick global recovery. The training period ranges from June 6, 2020 to December 30, 2020. When fitting TVRRR, the BIC criterion again selects $d = 2$ as the order of the model. The EM-algorithm converges after 104 iterations.

The corresponding MSEs are displayed in the bottom panel of Table 7. We observe a similar pattern as for the 2008 data set as regards the in-sample performance. According to the Diebold-Mariano test, the forecasting performance of TVRRR is significantly better than that of the naive prediction in one of the cases, and it is significantly better than updated RRR in three cases. In contrast, the forecasting performance of TVRRR is never significantly worse than that of the naive predictions and of updated RRR. Hence, also for the 2020 data, TVRRR shows significant performance improvement.

The cointegrating relations $b_{it}y_{t-1}$ exhibit a similar pattern as those for 2008 shown in Fig. 4. Therefore we do not include a corresponding figure for the 2020 data.

### 4.2. Covid-19 data

We analyze weekly Covid-19 cases per 100 000 capita for 12 European countries, namely Austria, Belgium, Czech Republic, France, Germany, Greece, Italy, Netherlands, Poland, Portugal, Sweden...
and Switzerland.\textsuperscript{4} Due to their proximity and frequent border crossings, the data can be assumed to be strongly correlated. Nevertheless, the relationships of the time series are likely to have changed during the pandemic.

We fit a time-varying VAR(1) model with reduced rank coefficient matrix of type (A),

\[
y_t = \alpha_t \beta y_{t-1} + \varepsilon_t. \tag{35}
\]

We do not preprocess the data. We fit the model starting at the end of March 2020 until February 2021, and predict the remaining values of the time series up to mid-November 2021, using \texttt{tvRRR} and updated \texttt{RRR}, as described in \textsc{Sec. 4.1}. More precisely, for updated \texttt{RRR} we use a time-constant model for the estimation period which is then refitted with every new observation in the validation period. The data were roughly split in two halves. The training and prediction period consist of 50 and 40 observations, respectively. Both models use latent rank $d = 2$.

The time series and the corresponding predictions are displayed in \textit{Fig. 6}. During the model fitting period, \texttt{tvRRR} captures the first and second wave of the pandemic nearly perfectly. For updated \texttt{RRR}, we can again observe the effect of averaging: The second peak leads to overestimation of the starting period. In the validation period from February 2021 to November 2021,

\footnote{Data retrieved from https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-19, accessed December 5, 2021.}
the one-step-ahead predictions of TVRRR closely track the progression of the number of Covid-19 cases in all countries. In contrast, the updated RRR performs badly in the validation period. This can also be seen from the mean squared prediction errors shown in Fig. 7. The overall MSE of TVRRR in the prediction period is less than one quarter of the MSE of updated RRR (3873 vs. 16,761). The Diebold-Mariano test finds statistically significantly better prediction performance of TVRRR for all countries except for the Netherlands.

We also tried model (B) for this data set. However, its performance is far worse than that of model (A), indicating that the time-invariant linear combinations \( \beta y_t \) contain essential information for prediction. However, these linear combinations have a time-varying effect \( x_t \) on the Covid-19 cases in the different countries.

5. Conclusions

Reduced-rank regression forms the basis for various models frequently used in applications, such as multivariate time-series regressions, cointegration models and methods for sufficient dimension reduction based on inverse regression. We propose a new time-varying coefficient reduced-rank regression modeling approach (TVRRR) by modeling the dynamics of the observations and the time-varying coefficients as a Gaussian linear state-space system. The estimation algorithm relies on classical methods from state-space modeling. We estimate the time-varying parameters using the Kalman filter, and fit the model to observed data using an EM-algorithm which relies on analytic updates and is numerically stable. Our EM-algorithm converges quickly most of the time. It needs on average 40 iterations in our simulation studies. It takes longer to reach convergence in real data applications, as expected. With respect to computer time, the estimation of the model for the Covid data takes about 10 seconds on a standard laptop with an Intel(R) i5 processor with 16 GB RAM. We use BIC to estimate the rank of the coefficient matrices with highly accurate results.

Our simulation studies and the data applications show the adaptive capabilities of the algorithm. It exhibits high estimation accuracy across different patterns of time-variation, as it accommodates both abrupt structural breaks and gradual changes in the time-varying coefficients. Moreover, TVRRR exhibits robustness to time-constant coefficients. An R-implementation of the model fitting algorithm is available on GitHub (https://github.com/b-brune/tvRRR). Additional simulation results are provided in the supplementary material.

Acknowledgements

We thank the editor, the associate editor and three anonymous referees for their comments and suggestions that helped us improve the paper. The simulation study presented in this paper was carried out using the Vienna Scientific Cluster (VSC).

Funding

This research was partially supported by the Austrian Science Fund (FWF P 30690-N35) and Vienna Science and Technology Fund (WWTF ICT19-018).

References

Ahn, S. K., Reinsel, G. C. (1988). Nested reduced-rank autoregressive models for multiple time series. Journal of the American Statistical Association 83(403):849–856.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In: Petrov, B. N., Csaki, F., eds., Second International Symposium on Information Theory. Budapest: Akademiai Kiado.
