High Dimensional Nonstationary Time Series Modelling with Generalized Dynamic Semiparametric Factor Model *

Song Song † Wolfgang K. Härdle ‡ Ya’acov Ritov§

November 28, 2010

Abstract

(High dimensional) time series which reveal nonstationary and possibly periodic behavior occur frequently in many scientific fields. In this article, we separate the modeling of high dimensional time series to time propagation of low dimensional time series and high dimensional time invariant functions via functional factor analysis. We propose a two-step estimation procedure. At the first step, we detect the deterministic trends of the time series by incorporating time basis selected by the group Lasso-type technique and choose the space basis based on smoothed functional principal component analysis. We show properties of this estimator under various situations extending current variable selection studies. At the second step, we obtain the detrended low dimensional stochastic process. However an important issue arises: is it justified, from an inferential point of view, to base further statistical inference on the estimated stochastic time series? We show that the difference between the inference based on the estimated time series and true unobserved time series is asymptotically negligible, which ultimately allows one to study the dynamics of the whole high-dimensional system with a low dimensional representation together with the deterministic trend. We apply the method to our motivating empirical problems: studies of the dynamic behavior of temperatures (further used for pricing weather derivatives), implied volatilities and risk patterns and correlated brain activities (neuro-economics related) using fMRI data, where a panel version model is also presented.

Keywords: Semiparametric model, Factor model, Group Lasso, Seasonality, Spectral Analysis, Periodic, Asymptotic inference, Weather, fMRI, Implied Volatility Surface

AMS 2000 subject classification: 62G08, 62G20, 62M10

JEL classification: C14, C32, G12

1 Introduction

Modeling high-dimensional data is a challenging task in statistics especially when the data come in a dynamic context and are observed at different time points with changing structure and different sample sizes. Such modeling challenges appear in many different fields. In meteorology and agricultural economics, one of the primary interests is to study fluctuations of temperatures at different locations,
for a recent summary, see Gleick et al. (2010). Such an analysis is essential for pricing weather derivatives and hedging weather risks, Odening et al. (2008). In neuro-economics, one uses (high dimensional) functional magnetic resonance imaging data (fMRI) to analyze the brain’s response to certain (economics related) stimuli as well as identifying its activation area, Worsley et al. (2002). In financial engineering, one studies the dynamics of the implied volatility surface for risk management, calibration and pricing purposes, Fengler et al. (2007). Other examples and research fields for very large dimensional time series include empirical macroeconomics, Stock and Watson (2005); mortality analysis, Lee and Carter (1992); bond portfolio risk management or derivative pricing, Nelson and Siegel (1987) and Diebold and Li (2006); limit order book dynamics, Hall and Hautsch (2006); yield curves, Hautsch and Ou (2008). In the biostatistical field, we refer to Martinussen and Scheike (2000) for bio-medical research; Kauermann (2000) for radiation treatment of prostate cancer; Gasser et al. (1987) for yield curves, Hautsch and Ou (2008). In the biostatistical field, we refer to Martinussen and Scheike (2000) for bio-medical research; Kauermann (2000) for radiation treatment of prostate cancer; Gasser et al. (1987) for yield curves, Hautsch and Ou (2008). In the biostatistical field, we refer to Martinussen and Scheike (2000) for bio-medical research; Kauermann (2000) for radiation treatment of prostate cancer; Gasser et al. (1987) for yield curves, Hautsch and Ou (2008).

The modeling challenge for high dimensional time series is that there are both high dimensionality (in space) and dynamics (in time). One approach utilizes a factor type model, which allows low-dimensional representation of the data by separating high dimensionality and dynamics, see Forni et al. (2005), Giannone et al. (2005), Stock and Watson (2002a), Stock and Watson (2002b). In an orthogonal L-factor model, a J-dimensional random vector \( Y_t = (Y_{t,1}, \ldots, Y_{t,J})^\top \) can be represented as

\[
Y_{t,j} = Z_{t,1} m_{1,j} + \cdots + Z_{t,L} m_{L,j} + \varepsilon_{t,j},
\]

where \( Z_{t,l} \) are common factors, \( \varepsilon_{t,j} \) are errors and the coefficients \( m_{l,j} \) are factor loadings. In the above described applications, the index \( t = 1, \ldots, T \) reflects the time evolution, and \( Y_t \) can be considered as a multidimensional not necessarily stationary time series. The study of the time behavior of the high-dimensional \( Y_t \) is then simplified to the modeling of \( Z_t = (Z_{t,1}, \ldots, Z_{t,L})^\top \), which is a more feasible task when \( L \ll J \). In a variety of applications, one has explanatory variables \( X_{t,j} \in \mathbb{R}^d \) at hand that may influence the factor loadings \( m_l \). An important refinement of the model (1) is to incorporate the existence of observable covariates \( X_{t,j} \). The factor loadings are then generalized to functions of \( X_{t,j} \), so that the model (1) is generalized to:

\[
Y_{t,j} = \sum_{l=1}^{L} Z_{t,l} \, m_l(X_{t,j}) + \varepsilon_{t,j}, \quad 1 \leq j \leq J, \quad 1 \leq t \leq T.
\]

where \( Z_t = (Z_{t,1}, \ldots, Z_{t,L})^\top \) (common factors) is an unobservable L-dimensional process (not necessarily stationary), \( m_l \) (factor loading functions) is an L-tuple \((m_1, \ldots, m_L)\) of unknown real-valued functions \( m_l \) defined on a subset of \( \mathbb{R}^d \) and \( \varepsilon_{t,j} \) are errors. The variables \( X_{1,1}, \ldots, X_{T,J_T}, \varepsilon_{1,1}, \ldots, \varepsilon_{T,J_T} \) are independent. Throughout the paper we assume that the \( X_{t,j} \) are deterministic. The errors \( \varepsilon_{t,j} \) are i.i.d., have zero mean and finite second moments. Park et al. (2009) consider this model when \( Z_t \) is stationary and call it a dynamic semiparametric factor model (DSFM). For simplicity of notation, we assume that the covariates \( X_{t,j} \) have support \([0,1]^d\), and also that \( J_t = J \) do not depend on \( t \) unless otherwise specified.

The approximation (2) involves unknown “space functions” \( m_l(\cdot) \) which in Park et al. (2009) are estimated via a B-Spline series:

\[
m_l(x) = \sum_{k=1}^{K} a_{lk} \psi_k(x)
\]

with a possibly multidimensional (as a tensor product of one dimensional) B-spline basis \( \{\psi_k\}_{k=1}^{K} \). Using the \( K \times J \) matrix \( \Psi_t = \{\psi_1(x_t), \ldots, \psi_K(x_t)\}^\top \) and the matrix \( A = (a_{lk}), l = 1, \ldots, L, k = 1, \ldots, K \), we can rewrite (2) as \( Y_t = Z_t^\top A \Psi_t + \varepsilon_t \). Expanding the time effect in a series leads us to modeling \( Z_t \)
as a sum of basis functions as well:

\[ Z_{it} = \sum_{r=1}^{R} \gamma_{it} u_r(t) \]  

(4)

Putting (3) and (4) together we obtain (5) and (6), i.e. we observe \((X_{t,j}, Y_{t,j})\) for \(j = 1, \ldots, J_t\) and \(t = 1, \ldots, T\) such that

\[
Y_{t,j} = \sum_{l=1}^{L} \sum_{r=1}^{R} u_r(t) \gamma_{rl} \sum_{k=1}^{K} a_{lk} \psi_k(X_{t,j}) + \varepsilon_{tj} \tag{5}
\]

\[
Y_t^\top = U_t^\top \Gamma^* A^* \Psi_t + \varepsilon_t \overset{\text{def}}{=} U_t^\top \beta^* \Psi_t + \varepsilon_t. \tag{6}
\]

Here \(U_t^\top = (u_1(t), \ldots, u_R(t))\) is a \(1 \times R\) matrix with \(u_r(t)\) as the pre-specified initial time basis, which we introduce to capture the global trend and periodic variations. \(\Psi_t = (\psi_1(X_t), \ldots, \psi_K(X_t))^\top\) is a \(K \times J\) matrix with \(\psi_k\) a space basis function. \(\Gamma^*, A^*\) and \(\beta^*\) are \(R \times L, L \times K\) and \(R \times K\) (unknown) underlying coefficient matrices consisting of \(\gamma_{rl}, a_{lk}\) and \(\beta_{rk}\) respectively. For every \(\beta\) matrix, we introduce \(\beta_r = (\beta_{rk}, 1 \leq k \leq K)\), that is, the column vector formed by the coefficients corresponding to the \(r\)-th time basis. Additionally we define \(\|\beta\|_{2,1} = \sum_{r=1}^{R} \sqrt{\sum_{k=1}^{K} \beta_{rk}^2}\). Finally we set \(R(\beta) = \{ r : \beta_r \neq 0 \}\) and \(M(\beta) = |R(\beta)|\) where \(|R(\beta)|\) denotes the cardinality of set \(R(\beta)\). For sake of simplicity and convenience, we sometimes use \(|\cdot|\) to denote the \(L_1\) norm for vectors and \(|| \cdot ||\) to denote the \(L_2\) norm for vectors or the mixed \((2,1)\) norm for matrices.

Since certainly not all initially included time basis are fully loading, to avoid overparametrization in time, basis or variable selection is necessary, i.e. some \(\beta_s\)s will be shrunk to 0 equivalently. A popular variable selection method is Lasso, Tibshirani (1996). An extension for factor structured models is the group Lasso, Yuan and Lin (2006), in which the penalty term is a mixed \((2,1)\)-norm of the coefficient matrix.

Under an additional Gaussian error assumption, we first show that this group Lasso type estimator enjoys sparsity inequalities (upper bounds on the prediction error and the distance between the estimator and the true regression matrix \(\beta^*\)) and variable selection properties. Finally, we show how our results can be extended to more general noise distributions, of which we only require the variance to be finite. Since the standard assumption on \(\varepsilon_t\) being independent is often not met in practice, we further extend our results into the dependent scenario. Since the original model (6) actually assumes that there is no randomness in time, we face some restrictions in practice. To this end, we consider an extension incorporating the stochasticity (in time) and call it a generalized dynamic semiparametric factor model (GDSFM). But it also poses an important question: is it justified, from an inferential point of view, to base further statistical inference on the detrended stochastic time series? We show that the difference of the inference based on the estimated time series and “true” unobserved time series is asymptotically negligible, which finally allows one to study the dynamics of the whole high-dimensional system with a low dimensional stochastic process representation together with the deterministic trend.

Another motivation of (4) (the expansion in time), is from the temperature analysis (across China over the past 50 years). Our data set is taken from Climatic Data Center (CDC), China Meteorological Administration (CMA), which contains daily observations from 159 weather stations across China (reduced from 202 after data cleaning) from Jan 1st, 1957 to Dec 31st, 2009, as can be seen from Figure 1 (left) (average over the 159 weather stations’ observations). Except the well known seasonality effect, we may expect a climate change related trend. If we take the moving average of 730 nearby days, which is \(\frac{1}{(159 \cdot 730)} \sum_{s=-365}^{s=365} \sum_{j=1}^{159} Y_{t+s,j}\) with \(Y_{t,j}\) being the temperature of the \(j\)th weather station at time \(t\), Figure 1 (right) shows a “large period” (around 10 years between peaks) and an upward trend of the Chinese temperatures. \(X_{t,j} = X_j\) is the three-dimensional geographical information of
the \( j \)th weather station. Studying the dynamics of temperatures in various places simultaneously using a well calibrated GDSFM model will enable us to forecast temperatures in time and space.

Figure 1: Temperatures of China from Jan 1st, 1957 to Dec 31st, 2009 (left) and the corresponding moving average (of 730 nearby days) view (right).

Another motivation for this research is from neuro-economics. Understanding which part of our brain is activated during risky decisions and whether there is a significant reaction to specific stimuli (neural processes underlying investment decisions) are important goals in neuroscience. We address this problem through the analysis of high dimensional, dynamic fMRI data recorded in an experiment (to be described in more detail later). The fMRI is a noninvasive technique of recording brain’s signals on spatial area in a given time period (2.5 sec for our data set). One obtains a series of three-dimensional images of the blood-oxygen-level-dependent (BOLD) fMRI signals, when an exercised person is subject to certain stimuli related with financial decisions (periodically), where \( Y_{t,j} \) is the BOLD value at voxel \( j \) and time \( t \). \( X_{t,j} = X_j \) is the three-dimensional geographical information of the \( j \)th voxel. An example of the images at one particular time point is presented in Figure 2.

Figure 2: Typical fMRI data in one particular time point. The brightness corresponds to the strength of the observed signals.

The third motivation for this modeling approach (especially the space part) comes from financial engineering, i.e. the dynamics of the implied volatility surface (IVS) (although considered as stationary time series here), as is observed in Figure 3. The IV is a volatility parameter that matches observed plain vanilla option prices with the ones given by the formula of Black and Scholes (1973), which is a key financial variable for trading, hedging and the risk management of option portfolios. Figure 3 shows the “string” structure of the IV data obtained from European option prices on the German stock index DAX (ODAX) for two different days from the whole data set - intraday observations from Jan 1, 2004 to Dec 30, 2004 from Bloomberg. The volatility strings shift towards expiry,
which is indicated by the bottom line in the figure. Moreover the shape of the IV strings is subject to stochastic deformation. Apart from the dynamic degeneration, one may also observe nonuniform frequency of the trades with significant greater market activities and the “smile” effect for the options closer to expiry or at-the-money. Fengler et al. (2007) first proposed to study the dynamics of the IV data, where $Y_{t,j}$ are the values of IV on the day $t$, and $X_{t,j}$ are the two-dimensional vectors of the moneyness and time-to-maturity, where the dimensionality $J$ (number of transactions) depends also on $t$.

Figure 3: The typical IV data design on two different days. In the maturity direction observations appear in the discrete points for each particular day. Bottom solid lines indicate the observed maturities. Left panel: observations on 20040701, $J_t = 5606$. Right panel: observations on 20040819, $J_t = 8152$.

The rest of the article is organized as follows. In the next section we present the estimation of (6) to extract the complex deterministic trends of the nonstationary time series using the group Lasso type technique. Its properties under various situations are presented in Section 3. Section 4 considers the general framework incorporating the stochasticity (in time) together with the corresponding asymptotic analysis. In Section 5 we present the results of simulation studies that illustrate the theoretical findings. In Section 6 we apply the model to the temperature, IVS and fMRI data, where a panel version of (6) is also presented. All technical proofs are sketched in Section 7.

2 Methodology

2.1 Choice of Time Basis

To capture the global trend in time, one may use an orthogonal Legendre polynomial basis: $u_1(t) = 1/C_1, u_2(t) = t/C_2, u_3(t) = (3t^2 - 1)/C_3, \ldots$ (throughout this paper, $C_i$ are generic constants). The rescaling is made here such that $\sum_{t=1}^p u_r^2(t)/C_r^2 = 1$. To capture periodic variations, we could use Fourier series, $u_4(t) = \sin(2\pi t/p)/C_4, u_5(t) = \cos(2\pi t/p)/C_5, u_6(t) = \sin(\{2\pi t/(p/2)\})/C_6, u_7(t) = \cos(\{2\pi t/(p/2)\})/C_7, \ldots$ with the given the period $p$. For example, in the fMRI application, we know that $p = 11.8$ (29.5s per trial & 2.5s per scan) and in the weather application, $p_1 = 365, p_2 = 365 \cdot 10$.

2.2 Choice of Space Basis

There are various choices for a space basis. For example, Park et al. (2009) use a series estimator as described in (3). However, it has some disadvantages. Firstly, since the B-spline basis $\{\psi_k\}_{k=1}^K$ is possibly multidimensional ($d > 1$), it is constructed as a tensor product of one dimensional ones. When $d \geq 3$, this may lead to quite large $K$, e.g. $K = 9 \times 9 \times 5 = 405$ in the fMRI application. More importantly, since the knots of the B-spline are equal-spaced, it could not capture some special
structure, e.g. the “smile” effect in the IVS modeling when the options are close to the maturity, as can be seen in Figure 4 from Park et al. (2009) (adaptive choice of the knots of the B-splines may solve this problem, but it is omitted here since not primary interest).

![Figure 4: Space basis using the series estimator for the IVS modeling.](image)

To this end, we propose a data driven method to estimate the space basis $\psi_1(x), \ldots, \psi_K(x)$, motivated by Hall et al. (2006), which combines smoothing techniques with ideas related to functional principal component analysis. We summarize the basic steps as follows:

1 Estimate the covariance operator. Write $X_{tj} = (X_{tj}^1, \ldots, X_{tj}^d)$, $u = (u^1, \ldots, u^d)$ and $v = (v^1, \ldots, v^d)$ (same for $b, \hat{b}, b_1, \hat{b}_1, b_2$ and $\hat{b}_2$). Given $u \in [0, 1]^d$, let $h_\mu$ and $h_\phi$ denote bandwidths, which could be selected as in the usual local polynomial regression setup and select $(\hat{a}, \hat{b}) = (a, b)$ to minimize

$$\sum_{t=1}^T \sum_{j=1}^{J_t} \{Y_{tj} - a - \sum_{c=1}^d b^c(u^c - X_{tj}^c)\}^2 K\left(\frac{X_{tj} - u}{h_\mu}\right),$$

and take $\hat{\mu}(u) = \hat{a}$. Then, given $u, v \in [0, 1]^d$, choose $(\hat{a}_0, \hat{b}_1, \hat{b}_2) = (a_0, b_1, b_2)$ to minimize

$$\sum_{t=1}^T \sum_{1 \leq j \neq k \leq J_t} \{Y_{tj}Y_{tk} - a_0 - \sum_{c=1}^d b_1^c(u^c - X_{tj}^c) - \sum_{c=1}^d b_2^c(v^c - X_{tk}^c)\}^2$$

$$\times K\left(\frac{X_{tj} - u}{h_\phi}\right) K\left(\frac{X_{tk} - v}{h_\phi}\right).$$

Denote $\hat{a}_0$ by $\hat{\phi}(u, v)$ and construct $\hat{\mu}(v)$ similarly with $\hat{\mu}(u)$. The estimate of the covariance operator is thus:

$$\hat{\psi}(u, v) = \hat{\phi}(u, v) - \hat{\mu}(u)\hat{\mu}(v).$$

Since the covariance operator is $J \times J$, where $J$ could be very large, to get its consistent estimates, various large covariance matrices regularization techniques, e.g. banding, Bickel and Levina (2008b) and thresholding, Bickel and Levina (2008a), could be further used.

2 Compute the principal space basis. Given the estimated operator, compute the largest $K$ eigenvalues and corresponding orthonormal eigenfunctions as the basis $\psi_1(X_{tj}), \ldots, \psi_K(X_{tj})$ ($\Psi, \Psi^T / J_t = I_K$ is thus valid). Computational methods could be found, for example, in Section 8.4 of Ramsay and Silverman (2005), where practical features regarding the operator-eigenfunction implementation are discussed in detail.
2.3 Estimation Procedure

We have now accumulated sufficient information to introduce the estimation method, which is summarized as below:

1. Find significantly loaded time basis functions by the group Lasso technique by minimizing:
\[
\min_{\beta}(JT)^{-1} \sum_{t=1}^{T} (Y_t^T - U_t^T \hat{\beta}^T \Psi_t) (Y_t^T - U_t^T \hat{\beta}^T \Psi_t)^T + 2\lambda \|\beta\|_{2,1}. \tag{7}
\]

2. Split the joint matrix \( \hat{\beta} \) into 2 separate coefficient matrices \( \hat{\Gamma}, \hat{A} \) by taking \( \hat{\Gamma} \) as the \( L \) eigenvectors of \( \hat{\beta} \beta^T \) with respect to the \( L \) largest eigenvalues, and \( \hat{A} = \hat{\Gamma}^T \hat{\beta} \).

To select \( K \) and \( L \) here, we could use either the classic “90%” rule in principal component analysis or the “explained variance” type selection method. Alternatively we could also sequentially test the \( \hat{\beta} \) as the initial value, with which usually a reasonable theoretical properties in a separate paper. Once \( K \) is selected, we assume that it is fixed (deterministic) afterwards. Otherwise we need also consider the additional risk from the space basis selection step.

In order to study the statistical properties of this estimator, it is useful to derive some optimality condition for a solution of (7). The solution path is therefore obtained by iteratively applying expression (10) to \( r = 1, \ldots, R \).

As a direct consequence of the Karush-Kuhn-Tucker conditions, we have a necessary and sufficient condition for \( \hat{\beta} \) to be a solution to expression (7) is
\[
(JT)^{-1} \sum_{t=1}^{T} \{\Psi_t(Y_t - \Psi_t^T \hat{\beta} U_t)U_t^T\}_r = \lambda \frac{\|\hat{\beta}_r\|}{\|\hat{\beta}_r\|}, \text{ if } \hat{\beta}_r \neq 0 \tag{8}
\]
\[
(JT)^{-1} \|\sum_{t=1}^{T} \{\Psi_t(Y_t - \Psi_t^T \hat{\beta} U_t)U_t^T\}_r\| \leq \lambda, \text{ if } \hat{\beta}_r = 0 \tag{9}
\]

Recall that \( \Psi_t U_t^T / J = I_K \). It can be easily verified that the solution to (8) and (9) is
\[
\hat{\beta}_r = \left(1 - \lambda / \|S_r\|\right) + S_r, \tag{10}
\]
where \( S_r = \sum_{t=1}^{T} \{\Psi_t(Y_t - \Psi_t^T \hat{\beta}_r U_t)U_t^T\}_r \), with \( \hat{\beta}_r = (\hat{\beta}_1, \ldots, \hat{\beta}_{r-1}, 0, \hat{\beta}_{r+1}, \ldots, \hat{\beta}_R) \). The solution to expression (7) can therefore be obtained by iteratively applying equation (10) to \( r = 1, \ldots, R \).

Since the group Lasso type estimates depend on the unknown tuning parameter parameter \( \lambda \), which needs to be estimated, to select the final models on the solution paths of the group selection methods, we introduce an easily computable \( C_p \)-type criterion as in Yuan and Lin (2006). The solution path is computed by evaluating on 100 equally spaced \( \lambda \)'s between 0 and \( \lambda_{max} = \max_r \|\sum_t \Psi_t U_{tr}\| / \sqrt{K} \).

We select the \( \lambda \) minimizing
\[
C_p(\lambda) = \sum_t \| Y_t^T - U_t^T \hat{\beta}_{OLS} \Psi_t \|^2 / \hat{\sigma}^2 - JT + 2df
\]
\[
\hat{\sigma}^2 = \sum_t \| Y_t^T - U_t^T \hat{\beta}_{OLS} \Psi_t \|^2 / JT - df
\]
\[
df = \sum_r 1{\| \hat{\beta}_r \| > 0} + \sum_r \frac{\| \hat{\beta}_r \|}{\| \hat{\beta}_{OLS} \|} (K - 1)
\]
Empirical evidence suggests that this approximation works fairly well. In our experience, the performance of this approximate $C_p$-criterion is generally comparable with that of computationally much more expensive (especially for the high-dimensional data) fivefold cross-validation, as already noted in Yuan and Lin (2006).

3 Estimates’ Properties

In this section, we first study the properties of this estimator as defined in (6) when the errors $\varepsilon_t$ are Gaussian. Our main results concern upper bounds on the prediction error and the distance between the estimator and the true matrix $\beta^*$, (Theorem 3.1). The techniques of proofs are closely build upon those of Lounici et al. (2009), Bickel et al. (2009) and Lounici (2008). In Theorem 3.2 we discuss how our results can be extended to more general noise distribution, of which we only require the variance to be finite. Since the standard assumption on $\varepsilon_t$ being independent is often not met in practice, in Theorem 3.3, we further extend our results into the dependent scenario.

**LEMMA 3.1** Consider the model (6) for $R \geq 2$ and $T,J \geq 1$. Assume that the random vectors $\varepsilon_1, \ldots, \varepsilon_T$ are i.i.d. Gaussian with zero mean and covariance matrix $\sigma^2 I_{J \times J}$, $\Psi_t \Psi_t^\top / J = I_K$, $\sum_{t=1}^T U_t^\top U_t / R = 1$, and $M(\beta^*) \leq s$. Let

$$\lambda = \frac{2\sigma}{\sqrt{JT}} \left(1 + A \log R/\sqrt{T}\right)^{1/2},$$

where $A > 8$ and let $q = \min(A \log R, \sqrt{T})$. Then with probability at least $1 - R^{1-q}$, for any solution $\hat{\beta}$ of problem (7) and $\forall \beta$ we have:

$$\sum_{t=1}^T \| \Psi_t^\top (\hat{\beta} - \beta^*) U_t \|^2 + \lambda \| \hat{\beta} - \beta \|_{2,1} \leq (JT)^{-1} \sum_{t=1}^T \| \Psi_t^\top (\beta - \beta^*) U_t \|^2 + 4\lambda \sum_{r \in \mathcal{R}(\beta)} \| \hat{\beta}_r - \beta_r \|,$$

(11)

$$\max_{1 \leq r \leq R} \| \sum_{t=1}^T \{ \Psi_t \Psi_t^\top (\hat{\beta} - \beta^*) U_t U_t^\top \}_r \| \leq \frac{3}{2} \lambda,$$

(12)

and

$$M(\hat{\beta}) \leq \frac{4\phi_{\text{max}}^2}{\lambda^2 T^2} \| \hat{\beta} - \beta^* \|_2,$$

(13)

where $\phi_{\text{max}}$ is the maximum eigenvalue of the matrix $\sum_{t=1}^T U_t U_t^\top$.

Before stating the first main result of this section, we make the following assumption first.

**ASSUMPTION 3.1** There exists a positive number $\kappa = \kappa(s)$ such that

$$\min \left\{ \frac{\sum_t \| \Psi_t^\top \Delta U_t \|}{\sqrt{J} \| \Delta_R \|} : |\mathcal{R}| \leq s, \Delta \in \mathbb{R}^{K \times R} \setminus \{0\}, \| \Delta_{\mathcal{R}^c} \|_{2,1} \leq 3 \| \Delta_R \|_{2,1} \right\} \geq \kappa,$$

where $\mathcal{R}^c$ denotes the complement of the set of indices $\mathcal{R}$, $\Delta_R$ denotes the matrix formed by stacking the rows of matrix $\Delta$ w.r.t. row index set $\mathcal{R}$.
Assumption 3.1 is essentially a restriction on the eigenvalues of $U_t$ as a function of sparsity $s$. It actually requires the initially involved time basis not to be too dependent, which is naturally satisfied by the orthogonal polynomials and Fourier series. Low sparsity means that $s$ is big and therefore $\kappa$ is small. $\kappa(s)$ is thus a decreasing function of $s$. For this reason we sometimes refer to it as Assumption RE(s), see also Bickel et al. (2009), but note that in their paper $l_1$ norms are used.

**THEOREM 3.1** Assume all conditions in Lemma 3.1 still hold and add Assumption 3.1. Then with probability at least $1 - R^{1-q}$, for any solution $\hat{\beta}$ of (7):

$$
(JT)^{-1} \sum_{t=1}^{T} \| \Psi_t^T (\hat{\beta} - \beta^*) U_t \|^2 \leq 64\sigma^2 s(1 + 4\log R/\sqrt{T})/(\kappa^2 J),
$$

$$
T^{-1/2} \| \hat{\beta} - \beta^* \|_{2,1} \leq 32\sigma s \sqrt{1 + 4\log R/\sqrt{T}}/(\kappa^2 \sqrt{J}),
$$

and

$$
M(\hat{\beta}) \leq 64\phi_{max}^2 s/\kappa^2
$$

Note that Theorem 3.1 is valid for any fixed $J, R, T$ and therefore yields non-asymptotic bounds. We could see that dependence on the number of initially specified time basis $R$ can be made negligible for large $T$. Additionally when the true coefficient matrix $\beta^*$‘s sparsity level is low ($s$ large, $\kappa$ small, $s/\kappa^2$ large), all the three bounds get larger and the number of nonzero rows of estimated one $\hat{\beta}^T$ is larger too correspondingly.

From now on, we only assume that the random variables $\varepsilon_{tj}$ are independent with zero mean and finite variance $\mathbb{E}(\varepsilon_{tj}^2) \leq \sigma^2$. In this case the results remain similar to those of the previous theorem, though the concentration effect is weaker. We use the following mild technical assumption.

**ASSUMPTION 3.2** The matrices $\Psi_t$ and $U_t$ are such that

$$(JT)^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J} \left( \max_{r} \left| \sum_{k=1}^{K} \Psi_{tkj} U_{tr} \right| \right)^2 \leq C,
$$

for a constant $C > 0$.

**THEOREM 3.2** Consider the DSFM (6) for $R \geq 3$ and $T, J \geq 1$. Assume that the random vectors $\varepsilon_1, \ldots, \varepsilon_T$ are independent with zero mean and finite variance $\mathbb{E}(\varepsilon_{tj}^2) \leq \sigma^2$, $\Psi_t^T I_K / J = I_K$, $\sum_{t=1}^{T} U_t^T U_t / R = 1$, and $M(\beta^*) \leq s$. Let also Assumption 3.2 be satisfied. Furthermore let $\kappa$ be defined as in Assumption 3.1 and $\phi_{max}$ is the maximum eigenvalue of the matrix $\sum_{t=1}^{T} U_t^T U_t$. Let

$$
\lambda = \sigma \sqrt{(\log R)^{1+\delta}/(JT)}, \quad \delta > 0.
$$

Then with probability at least $1 - (2e \log R - e)C/(\log R)^{1+\delta}$, for any solution $\hat{\beta}$ of (7) we have:

$$(JT)^{-1} \sum_{t=1}^{T} \| \Psi_t^T (\hat{\beta} - \beta^*) U_t \|^2 \leq 16\sigma^2 s (\log R)^{1+\delta}/(\kappa^2 J)
$$

$$
T^{-1/2} \| \hat{\beta} - \beta^* \|_{2,1} \leq 16\sigma s \sqrt{(\log R)^{1+\delta}/(\kappa^2 \sqrt{J})}
$$

and

$$
M(\hat{\beta}) \leq 64\phi_{max}^2 s/\kappa^2
$$
Thus, we see that under the finite variance assumption on the noise, the dependence on the dimension $R$ cannot be made negligible for large $T$.

Since the standard assumption on $\varepsilon_t$ being independent is often not met in practice, it is important to understand how the proposed estimator behaves under dependent error terms. As far as we know, our result is the first attempt with dependent error terms for (group) Lasso variable selection techniques. The other effort of getting rid of the independence assumption could be found in Jia et al. (2009), where they consider a sparse Possion-like model. Before moving on, similar to Janson (2004), we introduce the following definitions first.

Given a set $T$ and random variables $V_t$, $t \in T$, we say:

- A subset $T'$ of $T$ is **independent** if the corresponding random variables $\{V_t\}_{t \in T'}$ are independent.
- A family $\{T_j\}_j$ of subsets of $T$ is a **cover** of $T$ if $\bigcup_j T_j = T$.
- A family $\{(T_j, w_j)\}_j$ of pairs $(T_j, w_j)$, where $T_j \subseteq T$ and $w_j \in [0, 1]$ is a **fractional cover** of $T$ if $\sum_j w_j 1_{T_j} \geq 1_T$, i.e. $\sum_{j, t \in T_j} w_j \geq 1$ for each $t \in T$.
- A (fractional) cover is **proper** if each set $T_j$ in it is independent.
- $\mathcal{X}(T)$ is the size of the smallest proper cover of $T$, i.e. the smallest $m$ such that $T$ is the union of $m$ independent subsets.
- $\mathcal{X}^*(T)$ is the minimum of $\sum_j w_j$ over all proper fractional covers $\{(T_j, w_j)\}_j$.

Note that, in spite of our notation, $X$.

Moreover, consider the DSFM (6), where they consider a sparse Possion-like model. Before moving on, similar to Janson (2004), we introduce the following definitions first.

Given a set $T$ and random variables $V_t$, $t \in T$, we say:

- A subset $T'$ of $T$ is **independent** if the corresponding random variables $\{V_t\}_{t \in T'}$ are independent.
- A family $\{T_j\}_j$ of subsets of $T$ is a **cover** of $T$ if $\bigcup_j T_j = T$.
- A family $\{(T_j, w_j)\}_j$ of pairs $(T_j, w_j)$, where $T_j \subseteq T$ and $w_j \in [0, 1]$ is a **fractional cover** of $T$ if $\sum_j w_j 1_{T_j} \geq 1_T$, i.e. $\sum_{j, t \in T_j} w_j \geq 1$ for each $t \in T$.
- A (fractional) cover is **proper** if each set $T_j$ in it is independent.
- $\mathcal{X}(T)$ is the size of the smallest proper cover of $T$, i.e. the smallest $m$ such that $T$ is the union of $m$ independent subsets.
- $\mathcal{X}^*(T)$ is the minimum of $\sum_j w_j$ over all proper fractional covers $\{(T_j, w_j)\}_j$.

Note that, in spite of our notation, $\mathcal{X}(T)$ and $\mathcal{X}^*(T)$ depend not only on $T$ but also on the family $\{V_t\}_{t \in T}$. Note further that $\mathcal{X}^*(T) \geq 1$ (unless $T = \emptyset$) and that $\mathcal{X}^*(T) = 1$ if and only if the variables $V_t$, $t \in T$ are independent, i.e. $\mathcal{X}^*(T)$ is a measure of the dependence structure of $\{V_t\}_{t \in T}$. For example, if $V_t$ just depends on $V_{t-1}$ but independent of all $V_s$, $s < t - 1$, e.g. AR(1), $\mathcal{X}^*(T) = 2$.

We use the following mild technical assumption similar to Assumption 3.2.

**ASSUMPTION 3.3** The matrices $\Psi_t$ and $U_t$ and random variables $\varepsilon_t$ are such that

$$
(J^{-1} \sum_{k=1}^J \sum_{j=1}^J \Psi_{tkj}^{} \varepsilon_{tj}^{} U_{tr}^2)^2 \leq b_t^2 \quad \text{with a high probability}
$$

$$
\mathbb{E}(JT)^{-1} \left\{ \sum_{t=1}^T \left( \sum_{k=1}^K \sum_{j=1}^J \Psi_{tkj}^{} \varepsilon_{tj}^{} U_{tr}^2 \right)^2 \right\}^{1/2} \leq C' \sqrt{1/T}
$$

for $\forall r$ and some constants $b_t, C' > 0$, $t = 1, \ldots, T$. Note that dropping the sub-index $r$ for all constants here does not matter, since they could be taken as the maximum of all corresponding constants over different $r$s. Given $b_t, r = 1, \ldots, T$, $C'$ could be taken as $\max_t b_t$ for example.

We can now state our main result.

**THEOREM 3.3** Consider the DSFM (6) for $R \geq 3$, $T, J \geq 1$ and $T = \{1, \ldots, T\}$. Let also Assumption 3.3 be satisfied for the random vectors $\varepsilon_1, \ldots, \varepsilon_T$ and $\Psi_t^{} \Psi_t^\top / J = I_K$, $\sum_{t=1}^T U_t^\top U_t / R = 1$, and $M(\beta^s) \leq s$. Furthermore let $\kappa$ be defined as in Assumption 3.1 and $\phi_{max}$ is the maximum eigenvalue of the matrix $\sum_{t=1}^T U_t U_t^\top$. Let

$$
\lambda = \frac{C'}{\sqrt{T}} + \sqrt{\frac{\mathcal{X}^*(T) \sum_t b_t^2}{(\log R)^{1-\delta'} T^2}}, \quad \delta' > 0.
$$
Then with probability at least $p(1 - R^{-\delta})$, for any solution $\hat{\beta}$ of (7) we have:

$$(JT)^{-1} \sum_{t=1}^{T} \| \Psi_t^\top (\hat{\beta} - \beta^*) U_t \|_2^2 \leq 16 \left( C' + \sqrt{\frac{\mathcal{X}^*(T) \sum_i b_i^2}{(\log R)^{1-\delta} T}} \right) s/\kappa^2$$

$$T^{-1/2} \| \hat{\beta} - \beta^* \|_{2,1} \leq 16 \left( C' + \sqrt{\frac{\mathcal{X}^*(T) \sum_i b_i^2}{(\log R)^{1-\delta} T}} \right) s/\kappa^2$$

and

$$M(\hat{\beta}) \leq 64 \phi^2 \max_s s/\kappa^2$$

Not surprisingly, this theorem tells that the bounds get larger when the dependence level, i.e. $\mathcal{X}^*(T)$ increases and the bound is minimized when $\mathcal{X}^*(T) = 1$. The result is useful if the dependence level is not too strong, i.e. $\mathcal{X}^*(T) = \mathcal{O}(1)$. Otherwise if the dependence is strong, $\mathcal{X}^*(T) = \mathcal{O}(T)$.

## 4 Generalized Dynamic Semiparametric Factor Model

The original model (6) assumes that there is no stochastic evolution in time. To this end, we consider the following extension of (4) and (6):

$$Z_{0,t} = \sum_{r=1}^R \gamma_r u_r(t)$$

$$Y_t^\top = (Z_{0,t}^\top + U_t^\top \Gamma) A \Psi_t + \epsilon'_t = U_t^\top \Gamma A \Psi_t + (Z_{0,t}^\top A \Psi_t + \epsilon'_t), \quad (17)$$

with an unobservable $L$-dimensional random process $Z_{0,t}$ with $\mathbb{E}(Z_{0,t}|X_t) = 0$ and i.i.d. assumption on $\epsilon'_t$. We call (17) a generalized dynamic semiparametric factor model (GDSFM). If we concentrate on prediction, the trend represented by $U_t^\top \Gamma$ is enough. However, if we are interested in the stochasticity or dynamics of the original high dimensional time series, $Z_{0,t}$ comes into play, e.g. for pricing weather derivatives and various other financial engineering examples. The estimation procedure is now divided into 2 steps:

- For the model $Y_t^\top = U_t^\top \Gamma A \Psi_t + (Z_{0,t}^\top A \Psi_t + \epsilon'_t)$, treat $Z_{0,t}^\top A \Psi_t + \epsilon'_t$ as the $\epsilon_t$ in (6) and find the best parametric approximation according to the estimation procedure described in Subsection 2.3 to get the deterministic trend $U_t^\top \Gamma$.

- Based on $\hat{Y}_t^\top \overset{\text{def}}{=} Y_t^\top - U_t^\top \hat{\beta} \Psi_t$, $\hat{A}$ and $\Psi_t$, use the ordinary least square method to obtain the estimated random process $\hat{Z}_{0,t}$.

As we could see from step one here, since $\epsilon_t$ in (6) involves $Z_{0,t}^\top A \Psi_t + \epsilon'_t$, where $Z_{0,t}$ is a random process inhering dependence structure, Theorem 3.3 shows its necessity again. In the second step, $Z_{0,t}$ is estimated based on $\hat{\beta}$ instead of $\beta^*$, we need to show the influence of this plug-in estimate is negligible. Our first result this section relies on the following assumptions, which are similar to Assumptions (A1-8) in Park et al. (2009).

### Assumption 4.1

4.1.1 The variables $X_{1,1}, \ldots, X_{T,J}, \epsilon_{1,1}', \ldots, \epsilon_{T,J}',$ and $Z_{0,1}, \ldots, Z_{0,T}$ are independent.

4.1.2 For $t = 1, \ldots, T$ the variables $X_{t,1}, \ldots, X_{t,J}$ are identically distributed, have support $[0,1]^d$ and a density $f_t$ that is bounded from below and above on $[0,1]^d$, uniformly over $t = 1, \ldots, T$. 

11
4.1.3 We assume that \( E \varepsilon'_{t,j} = 0 \) for \( 1 \leq t \leq T, 1 \leq j \leq J \), and for \( c > 0 \) small enough \( \sup_{1 \leq t \leq T, 1 \leq j \leq J} E \exp\{c(\varepsilon'_{t,j})^2\} < \infty \).

4.1.4 The vector of functions \( m = (m_1, \ldots, m_L)^\top \) can be approximated by \( \Psi_k \), i.e.

\[
\delta_K \overset{\text{def}}{=} \sup_{x \in [0,1]^d} \inf_{A \in \mathbb{R}^{L \times K}} \|m(x) - A\Psi(x)\| \to 0
\]

as \( K \to \infty \). We denote \( A \) that fulfills \( \sup_{x \in [0,1]^d} \|m(x) - A\Psi(x)\| \leq 2\delta_K \) by \( A^* \).

4.1.5 There exist constants \( 0 < C_L < C_U < \infty \) such that all eigenvalues of the matrix \( T^{-1} \sum_{t=1}^T Z_0 Z_{0,t}^\top \) lie in the interval \([C_L, C_U]\) with probability tending to one.

4.1.6 The minimization \((7)\) runs over all values \( \beta \) with

\[
\sup_{x \in [0,1]^d} \max_{1 \leq t \leq T} \|Z_{0,t}^\top A\Psi(x)\| \leq M_T,
\]

where the constant \( M_T \) fulfills \( \max_{1 \leq t \leq T} \|Z_{0,t}\| \leq M_T/C_m \) (with probability tending to one) for a constant \( C_m \) such that \( \sup_{x \in [0,1]^d} \|m(x)\| < C_m \).

4.1.7 It holds that \( \rho^2 = (K + T) M_T^2 \log(JTM_T)/(JT) \to 0 \). The dimension \( L \) is fixed.

Assumption (4.1.6) and the additional bound \( M_T \) in the minimization is introduced for purely technical reasons.

**THEOREM 4.1** Suppose that model \((17)\), all assumptions in Theorem 3.3 and Assumption 4.1 hold. Then we have

\[
\frac{1}{T} \sum_{1 \leq t \leq T} \left\| \hat{Z}_{0,t}^\top A - Z_{0,t}^\top A^* \right\|^2 = O_P(\rho^2 + \delta_k^2). \tag{18}
\]

In the following we discuss how a statistical analysis differs if the inference of stochasticity on \( Z_{0,t} \) is based on \( \hat{Z}_{0,t} \) (note that the trend \( U_t^\top T \) is deterministic) instead of using (the unobserved) process \( Z_{0,t} \). We will show that the differences are asymptotically negligible (up to an orthogonal transformation). This is the content of the following theorem, where we consider estimators of autocovariances and show that these estimators differ only by second order terms. This asymptotic equivalence carries over to classical estimation and testing procedures in the framework of fitting a vector autoregresssive model.

For the statement of the theorem we need the following assumptions, which are similar to Assumptions (A9-11) in Park et al. (2009):

**ASSUMPTION 4.2** 4.2.1 \( Z_{0,t} \) is a strictly stationary sequence with \( E(Z_{0,t}) = 0 \), \( E(\|Z_{0,t}\|^\gamma) < \infty \) for some \( \gamma > 2 \). It is strongly mixing with 

\[
\sum_{i=1}^\infty \alpha(i)^{(\gamma-2)/\gamma} < \infty.
\]

The matrix \( E Z_{0,t} Z_{0,t}^\top \) has full rank. The process \( Z_{0,t} \) is independent of \( X_{11}, \ldots, X_T, \varepsilon_{11}, \ldots, \varepsilon_{T,J} \).

4.2.2 It holds that \( \|\log(KT)^2\{(KM_T/J)^{1/2} + T^{1/2}M_T^2J^{-2} + K^{3/2}J^{-1}
+ K^4/3J^{-2/3}T^{-1/6}\} + 1\|^2T^{1/2}(\rho^2 + \delta_K^2) = o(\rho^2 + \delta_K^2) \)

Assumption (4.2.2) poses very weak conditions on the growth of \( J,K,T \). Suppose, for example, that \( M_T \) is of logarithmic order and that \( K \) is of order \( (JT)^{1/5} \) so that the variance and the bias are balanced for twice differentiable functions. In this setting, (4.2.1) only requires that \( T/J^2 \) times a logarithmic factor converges to zero.
Furthermore, please note that the minimization problem (7) has only a unique solution up to $\beta$, but not to $\Gamma, A$. If $(\tilde{Z}_{0,t}, \tilde{A})$ is a minimizer, then also $(B^\top \tilde{Z}_{0,t}, B^{-1}A)$ is a minimizer, where $B$ is an arbitrary invertible matrix. In particular, with the choice $B = (\sum_{t=1}^{T} Z_{0,t} \tilde{Z}_{0,t})^{-1} \sum_{t=1}^{T} Z_{0,t}^T Z_{0,t}$, we get for $\tilde{Z}_{0,t} \equiv B^\top \tilde{Z}_{0,t}$ and $\tilde{A} \equiv B^{-1}A$ that $\sum_{t=1}^{T} Z_{0,t} (\tilde{Z}_{0,t} - Z_{0,t})^\top = 0$. Without loss of generality, we may assume $T^{-1} \sum_{s=1}^{T} \tilde{Z}_{0,s} = T^{-1} \sum_{s=1}^{T} Z_{0,s} = 0$. Additionally define

$$
\tilde{Z}_{n,t} = (T^{-1} \sum_{s=1}^{T} \tilde{Z}_{0,s} \tilde{Z}_{0,s}^\top)^{-1/2} \tilde{Z}_{0,t}
$$

$$
Z_{n,t} = (T^{-1} \sum_{s=1}^{T} Z_{0,s} Z_{0,s}^\top)^{-1/2} Z_{0,t}.
$$

**THEOREM 4.2** Suppose that model (17) holds. Besides all assumptions in Theorem 3.3, let also Assumption 4.1-4.2 be satisfied. Then there exists a random matrix $B$ such that for $h \geq 0$

$$
T^{-1} \min_{t=\max[1,-h+1]}^{\min[T,T-h]} \sum_{t=\max[1,-h+1]}^{T-h} \tilde{Z}_{0,t} (\tilde{Z}_{0,t+h} - \tilde{Z}_{0,t})^\top - Z_{0,t} (Z_{0,t+h} - Z_{0,t})^\top = o_p(T^{-1/2})
$$

and

$$
T^{-1} \min_{t=\max[1,-h+1]}^{\min[T,T-h]} \sum_{t=\max[1,-h+1]}^{T-h} \tilde{Z}_{n,t} \tilde{Z}_{n,t+h}^\top - Z_{n,t} Z_{n,t+h}^\top = o_p(T^{-1/2}).
$$

5 Simulation Study

We present three simulations which investigate how the spread of the sparsity level $M(\beta^*)$, the number of initial time basis $R$ and the dependence level of the error terms affect the performance. In the first example, we show how changing the values of $M(\beta^*)$ result in changing the two measures of estimation error in light of Theorem 3.1:

$$
L_{par} = 1 - \frac{\sum_{r=1}^{R} \| \hat{\beta}_r - \beta_r \|_\infty}{\sum_{r=1}^{R} \| \beta_r \|_\infty}
$$

$$
L_{prec} = 1 - \frac{\sum_{r=1}^{R} \| \Psi^\top I \hat{\beta}_r - \beta_r \|_\infty U_t}{\sum_{r=1}^{R} \| \Psi^\top I \beta_r \|_\infty U_t}
$$

All codes were done in Matlab and are available on the author’s homepage or www.quantlet.com. We applied the above algorithm (8), (9) to the following simulated data. We generate random $\beta_1, \ldots, \beta_{179} \in \mathbb{R}^5$ such that all coordinates are independent and consider an initial model with the parameters such that $\beta_{rk} \sim N\{0, \exp(-2k/5)\}, r = 1, \ldots, 179, k = 1, \ldots, 5$. We randomly pick $179 - M(\beta^*) \beta_s$ from $\beta_1, \ldots, \beta_{179}$ and assign them to be $0 \in \mathbb{R}^5$. We choose the same time basis as in Table 4. For the space part, inspired by Park et al. (2009), we considered $d = 2, L = 3$ and the following tuple of 2-dimensional functions:

$$
m_0(x_1, x_2) = 1, \quad m_1(x_1, x_2) = 3.46(x_1 - .5),
$$

$$
m_2(x_1, x_2) = 9.45 \{ (x_1 - .5)^2 + (x_2 - .5)^2 \} - 1.6,
$$

$$
m_3(x_1, x_2) = 1.41 \sin(2\pi x_2).
$$
The coefficients in these functions were chosen so that \( m_1, m_2, m_3 \) are close to orthogonal. The design points \( X_{t,j} \) were independently generated from a uniform distribution on the unit square. We generate 
\[
Y_t^T = U_t^T \beta^T \Psi_t + \varepsilon_t, t = 1, \ldots, 19345 \text{ where } \varepsilon_t \text{ is drawn as i.i.d. } N(0, 0.05).
\]

The convergence of the algorithm presented in (10) is usually achieved up to 5 iterations. Figure 5 is an illustration plot about how the group Lasso penalty shrinks the coefficients. With 250 repetitions, Table 1 displays different \( L_{\text{par}} \) and \( L_{\text{pre}} \) s w.r.t. different sparsity levels. Our theoretical results in the previous sections suggest that when \( M(\beta^*) (s) \) is small, \( L_{\text{par}} \) and \( L_{\text{pre}} \) will be large, which is confirmed by the simulation results.

\[
\begin{array}{ccc}
M(\beta^*) = 100 & M(\beta^*) = 50 & M(\beta^*) = 20 \\
L_{\text{par}} & 0.870 & 0.918 & 0.931 \\
L_{\text{pre}} & 0.710 & 0.835 & 0.859 \\
\end{array}
\]

Table 1: \( L_{\text{par}} \) and \( L_{\text{pre}} \) w.r.t. different sparsity levels.

The second experiment compares how \( L_{\text{par}} \) and \( L_{\text{pre}} \) react to changing the numbers of initial time basis \( R \), for \( M(\beta^*) = 50 \), if we additionally include the quartic term in the orthogonal polynomial and double the number of Fourier series, \( R = 53 \cdot 4 + 40 = 252 \) and if we remove the cubic term in the orthogonal polynomial and half the number of Fourier series, \( R = 53 \cdot 2 + 10 = 116 \). The \( L_{\text{par}} \) and \( L_{\text{pre}} \) s are presented in Table 2.

\[
\begin{array}{ccc}
R = 116 & R = 179 & R = 252 \\
L_{\text{par}} & 0.879 & 0.918 & 0.920 \\
L_{\text{pre}} & 0.695 & 0.835 & 0.841 \\
\end{array}
\]

Table 2: \( L_{\text{par}} \) and \( L_{\text{pre}} \) w.r.t. different number of initially involved time basis.

As we could see, when \( R \) increases, \( L_{\text{par}} \) and \( L_{\text{pre}} \) increase. This indicates us that in practice we need take a relatively large \( R \) value, i.e. involve as many as possible time basis.

The third experiment compares how \( L_{\text{par}} \) and \( L_{\text{pre}} \) are sensitive to the dependence level of the error items. We generated \( \varepsilon_t \) from a centered VAR(1) process \( \varepsilon_t = R \varepsilon_{t-1} + U_t \), where \( U_t \) is \( N_3(0, \Sigma_U) \) random vector, the rows of \( R \) from the top equal \((0.95, -0.2, 0), (0, 0.8, 0.1), (0.1, 0.0.6)\), and \( \Sigma_U = 10^{-4} I_3 \). We choose \( M(\beta^*) = 50, R = 179 \) as before. Besides the VAR(1) process indicated before, we also tried the VAR(2) to generate \( \varepsilon_t \). Table 3 displays the result, where we use VAR(0) to denote the independent
case. The performance decreases when the error terms are more dependent, which is consistent with Theorem 3.3.

<table>
<thead>
<tr>
<th></th>
<th>VAR(0)</th>
<th>VAR(1)</th>
<th>VAR(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L_{\text{par}})</td>
<td>0.918</td>
<td>0.854</td>
<td>0.783</td>
</tr>
<tr>
<td>(L_{\text{pre}})</td>
<td>0.835</td>
<td>0.774</td>
<td>0.712</td>
</tr>
</tbody>
</table>

Table 3: \(L_{\text{par}}\) and \(L_{\text{pre}}\) w.r.t. different levels of dependence of \(\varepsilon_t\).

For more Monte Carlo experiments concerning Theorem 4.2, we refer to Park et al. (2009).

6 Weather, Neuro-economics and IVS

This section presents three applications to the temperature, fMRI and IVS analysis. First, we fit the model to the daily temperature observations by Climatic Data Center (CDC), China Meteorological Administration (CMA), as introduced in Figure 1. To capture the upward trend, seasonal and “large period” effects, for time basis, similar to Racsko et al. (1991), Parton and Logan (1981) and Hedin (1991), we propose the following initial choice of time basis (rescaling factors omitted) in Table 4.

<table>
<thead>
<tr>
<th>Factors</th>
<th>Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trend</td>
<td>1</td>
</tr>
<tr>
<td>(Year by Year)</td>
<td>(t)</td>
</tr>
<tr>
<td></td>
<td>(3t^2 - 1)</td>
</tr>
<tr>
<td>Seasonal</td>
<td>(\sin \frac{2\pi t}{365})</td>
</tr>
<tr>
<td>Effect</td>
<td>(\cos \frac{2\pi t}{365})</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Initial choice of \(53 \cdot 3 + 20 = 179\) time basis.

Figure 6: Distribution of the eigenvalues and the relative proportion of variance explained by the first \(K\) basis.

For the space basis, consider the eigenvalues of the smoothed (with the usual optimal bandwidth for local polynomial regression) covariance operator (Figure 6) and also the climate types of China (Figure 7), the number of space basis \(K = 5\) seems to be satisfactory although \(K = 10\) is needed
to pass the “90%” rule. Please note that it is significantly smaller than the number of terms of a series estimator. Figure 8 displays the estimated coefficients of the first factor with respect to the $54 \cdot 3$ yearly polynomial time basis under the optimal choice of $\lambda$. The coefficients of constant, linear and quadratic terms are displayed as solid, dashed and dotted lines correspondingly. The coefficients estimates of the 20 Fourier series time basis corresponding to the optimal $\lambda$ are displayed in Table 5. It clearly indicates the 10-year period effect which, as some meteorologists claimed, are related to the solar activity. Figure 9 displays the extracted trends based on $U_t^T \hat{\beta}$, where the five lines correspond to the five factors. The characters of this kind of nonstationary time series further indicate that the autoregressive model may not be a proper tool to capture them. Firstly, since there exists the “stronger and stronger global warming” effect, if we use AR model, the constant, linear and quadratic coefficients should be time variant (increasing). Secondly, the existence of “large period” effect also poses the problem of lag or frequency selections there. Both of these actually introduce bigger technical challenges.

Since the eigenvalues of $\hat{\beta} \hat{\beta}^T$ are $(0.4683, 0.0106, 0.0068, 0.0040, 0.0007, 0.0000, \ldots)$, we choose $L = 5$ and estimated the remaining 5-dimensional random process $\hat{Z}_{0,t}$, e.g. $\hat{Z}_{0,t,1}$ as displayed in Figure 10 ($\hat{Z}_{0,t,2} - \hat{Z}_{0,t,5}$ are omitted due to the limited space here). The expectation of the random process is close to zero, which indicates our detrending using the group Lasso type technique works well.
Basis Estimates
\[
\begin{align*}
\sin 2\pi t/365 &{}: -0.1801, 0.1912, -0.1076, -0.0173, 0.3787 \\
\cos 2\pi t/365 &{}: -0.4651, 0.6220, -0.1794, 0.0615, -0.3170 \\
\sin 4\pi t/365 &{}: 0.0000, 0.0000, 0.0000, 0.0000, 0.0000 \\
\cos 4\pi t/365 &{}: -0.0408, 0.0564, 0.0357, -0.0735, -0.1158 \\
\ldots &{}: 0.0000, \ldots \\
\sin 10\pi t/365 &{}: 0.0000, \ldots \\
\cos 2\pi t/(365 \cdot 10) &{}: 0.0386, 0.0116, 0.1171, -0.2034, 0.1262 \\
\cos 2\pi t/(365 \cdot 10) &{}: 0.0000, \ldots \\
\ldots &{}: 0.0000, \ldots \\
\cos 10\pi t/(365 \cdot 10) &{}: 0.0000, \ldots
\end{align*}
\]

Table 5: Estimated coefficients of the 5 factors w.r.t. the 20 Fourier series time basis.

The residual multi-dimensional random process could be further modeled by multivariate time series techniques. For example, if we use VAR(1) process
\[
\hat{Z}_{0,t} = \mathcal{R}\hat{Z}_{0,t-1} + \varepsilon_{0,t},
\]
where \(\varepsilon_{0,t}\) is a random vector, the estimated coefficient matrix is:
\[
\begin{pmatrix}
0.8493 & -0.1240 & 0.0194 & -0.0361 & 0.0177 \\
0.2531 & 0.8424 & 0.0275 & 0.0300 & -0.0424 \\
-0.7896 & -0.3359 & 0.6789 & -0.0338 & 0.1227 \\
0.1411 & -0.0758 & -0.1755 & 0.6820 & 0.0032 \\
0.0413 & 0.7199 & -0.2709 & 0.0727 & 0.6541
\end{pmatrix}
\]

In comparison with the existing temperature modeling or weather derivatives pricing techniques, e.g. Benth and Benth (2005), we have the following advantages. Firstly, based on the high dimensional time series data, we offer integrated analysis considering space (high dimensionality) and time (dynamics) parts simultaneously, while forecasting at places different from the existing weather stations is also possible since the space basis are actually functions of the geographical location information. Secondly, we extract the trend more clearly. Thirdly, we provide the theoretical justification for further inferential analysis of \(\hat{Z}_{0,t}\) instead of \(Z_{0,t}\). However, if we have a closer look at the enlarged estimated stochastic process in Figure 10, we find that the volatility of the random process also has a seasonality, which is actually due to the fact that the variance of the noise (temperature, fMRI...
etc.) scale linearly with the expectation of the measurements. This motivates to consider (6) under heteroscedasticity (Poisson - like model) as follows:

\[ Y_t^T = U_t^T \Gamma A \Psi_t + \varepsilon_t, \quad \text{Cov}(\varepsilon_t) = \text{diag}(|U_t^T \Gamma A \Psi_t|), \]

which will be presented in a separate paper.

As a second application of the model, we consider a microeconomic experiment based on fitting an fMRI data set. Here we used a novel investment decision task that uses streams of (past) returns as stimuli to the exercised subjects, where the flowchart of the experiment is presented in Figure 11 (left), and obtain a series of three-dimensional images of the blood-oxygen-level-dependent (BOLD) fMRI signals. Our model helps to identify the corresponding brain’s activation areas and to simplify the inference to the analysis of time propagation of a few number of factors (low-dimensional representation). Additionally we classify the risk attitudes of different subjects based on the coefficients of time basis, which performed quite well compared to the classic risky decision making model (risk-return model) which is based on the subjects’ answers directly, where the risk attitude can be measured as value reduction in Euro for maximum risk (the case when the subjective perceived risk = 100), as described in Mohr et al. (2010). All subjects were classified as risk averse indicated by a positive risk weight as shown in Figure11 (right). However, for six subjects the risk attitude was quite low (risk weight< 5, colored with blue) resulting in only a small influence of risk on value. For the experimental procedure and the fMRI data description, we refer to Myšičková et al. (2010).

Since we are analyzing multi subjects $1 \leq i \leq I$ here, we obtain a panel version of the original
model (6) to
\[
\begin{align*}
Y_{i,t,j}^{i} &= \sum_{l=1}^{L} (\alpha_{i,t,l}^{i} + U_{t}^{i}\Gamma_{l}^{i})m_{l}(X_{t,j}) + \varepsilon_{t,j}, \quad 1 \leq j \leq J, 1 \leq t \leq T,
\end{align*}
\]
where the fixed effect \(\alpha_{i,t,l}^{i}\) is the individual effect on function \(m_{l}\) for subject \(i\) at time point \(t\). For identification purpose, we assume
\[
E\left(\sum_{i=1}^{I} \sum_{l=1}^{L} \alpha_{i,t,l}^{i}m_{l}(X_{t,j})|X_{t,j}\right) = 0.
\]
Please notice that assuming different subjects have the same basis function in space \(m_{l}\) makes sense here since the basis function is used to detect which part of the brain is activated for risky decisions, which should be homogeneous for human beings. Thus for this panel data, we have:
\[
Y_{t,j} = \sum_{l=1}^{L} (U_{t}^{i}\Gamma_{l}^{i})m_{l}(X_{t,j}) + \varepsilon_{t,j}, \quad 1 \leq j \leq J,
\]
and our 2-step estimation procedure is as follows:

1. Take the average of \(Y_{i,t,j}^{i}\) across different subjects \(i\), and estimate the common basis function in space \(m_{l}\) as in the original approach.

2. Given the common \(m_{l}\), for different subjects \(i\), estimate their specific factors in time \(Z_{i,t}^{i}\).

\[
Y_{t,j}^{i} = \sum_{l=1}^{L} U_{t}^{i}\Gamma_{l}^{i}m_{l}(X_{t,j}) + \varepsilon_{t,j}^{i}
\]

Since most of the technical details have been illustrated in the previous application, it is skipped here, while the differences will be emphasized. Since the significantly larger dimension \(J = 76176\) is observed here, computing eigenvalues of a \(76176 \times 76176\) matrix will encounter significant numerical difficulties. By using the fact that \(cc^{\top}\) has the same eigenvalues as \(c^{\top}c\) (where \(c\) is a \(J \times T\) matrix), we only need to compute eigenvalues of a \(722 \times 722\) matrix. If we additionally take the average of every 10 \(Y_{i,t,j}^{i}\)s over \(t\), we only need compute eigenvalues from a \(73 \times 73\) matrix.

The third factor loading function \(\hat{m}_{3}\) shown in Figure 12 could be identified as the Ventromedial prefrontal cortex (VMPFC) located in the bottom frontal part in the brain, which is the center for utility and conform herewith with our experiment (it is why it is presented here). The other functions \(m_{l}\), which also represent exactly those brain regions which we have expected to be involved during the experiment, are presented in Myšicková et al. (2010).

We use cubic orthogonal polynomials as time basis. Figure 13 displays the response curve (to the “decision phase” stimuli) \(U_{t}^{i}\Gamma_{2}^{i}\) for different subjects. Based on the estimated factors for different individuals, we could further develop a classification method which can predict the risk aversion only based on the measured fMRI signals. Observing that different probands’ response curves have different patterns and their corresponding \(\hat{Z}_{0,t}\) have different volatilities, for this purpose we use the estimated coefficients \(\Gamma_{3}^{i}\) since it correspond to the brain activity of the VMPFC, which is linked with utility. To provide the classification analysis, we apply Support Vector Machines (SVM), which is a widely used nonlinear method based on statistical learning theory. For the learning step, strongly risk averse subjects were labeled by \(-1\) and weakly risk averse subjects by \(1\). Then, we applied the leave-one-out method to first train and then estimate the classification rate of the SVM. The classification rates are 85% for strongly risk averse and 60% for weakly risk averse individuals. More importantly, these rates hold for a wide range of prior parameters: the radial basis coefficient \(r\) (0.25 – 0.35) and the capacity \(C\) (20 – 90).

In the analysis of IVS data, deterministic trends are not present, and do not make sense from a non arbitrage point of view. We may therefore assume stationarity. The first detrending step
Figure 12: Estimated function $\hat{m}_3$ shown in 12 axial slices (left) and as a 3D-plot in a posterior view (right) with highlighted Ventromedial prefrontal cortex (VMPFC).

Figure 13: Response curves (to stimuli) $U_t^\top \hat{\Gamma}_2$ for probands $i = 9$ (blue dashed) & $i = 19$ (black dash dotted) with periodic cubic polynomial as time basis.

is therefore omitted, alternatively, we could still use the dynamic semiparametric factor modeling approach proposed by Park et al. (2009) except for a different space basis. To this end, due to the limited space here, we only present the new space basis of the implied volatility surface (IVS) application in Figure 14 (left). We see that the “smile” effect is captured very well. The corresponding estimated time series of factors $\hat{Z}_{t,1}$, $\hat{Z}_{t,2}$ (stationary) are presented in Figure 14 (right).

7 Appendix

Here we collect one auxiliary result which is used in the proof of Lemma 3.1.

**Lemma 7.1** For any $I \times J$ matrix $A$ and any $J \times K$ matrix $B$, we have $\|AB\|_{2,1} \leq \|A\|_{2,1} \|B\|_{2,1}$.
Table 6: Classification rates of the SVM method using median(left) and mean (right) of volatilities of $\Delta \tilde{Z}_{t,2}$.

<table>
<thead>
<tr>
<th>MEAN</th>
<th>Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Strongly 0.85 0.14</td>
</tr>
<tr>
<td></td>
<td>Weakly 0.59 0.40</td>
</tr>
</tbody>
</table>

Proof With Cauchy Schwartz inequality it is not hard to derive:

$$\|AB\|_{2,1} = \sum_{i=1}^{I} \sqrt{\sum_{k=1}^{K} \left( \sum_{j=1}^{J} a_{ij} b_{jk} \right)^2}$$

$$\leq \sum_{i=1}^{I} \sqrt{K \sum_{k=1}^{K} \left( \sum_{j=1}^{J} a_{ij}^2 \sum_{j=1}^{J} b_{jk}^2 \right)}$$

$$\leq \left( \sum_{i=1}^{I} \sqrt{J \sum_{j=1}^{J} a_{ij}^2} \right) \left( \sum_{k=1}^{K} \sqrt{J \sum_{j=1}^{J} b_{jk}^2} \right)$$

$$= \|A\|_{2,1} \|B\|_{2,1}$$

Proof of Lemma 3.1 The proof is in a similar spirit of the one of Lemma 3.1 in Lounici et al. (2009). By the definition of $\hat{\beta}$ as a minimizer of (7), for $\forall \beta$ we have

$$(JT)^{-1} \sum_{t=1}^{T} \| \Psi_t^\top \beta U_t - Y_t \|^2 + 2\lambda \sum_{r=1}^{R} \| \hat{\beta}_r \|$$

$$\leq (JT)^{-1} \sum_{t=1}^{T} \| \Psi_t^\top \beta U_t - Y_t \|^2 + 2\lambda \sum_{r=1}^{R} \| \beta_r \|,$$
which, using $Y_t = \Psi_t^T \beta^* U_t + \varepsilon_t$, is equivalent to

$$(JT)^{-1} \sum_{t=1}^T \| \Phi_t^T (\hat{\beta} - \beta^*) U_t \|^2 \leq (JT)^{-1} \sum_{t=1}^T \| \Phi_t^T (\beta - \beta^*) U_t \|^2$$

$$+ 2(JT)^{-1} \sum_{t=1}^T \varepsilon_t^T \Phi_t^T (\hat{\beta} - \beta) U_t + 2\lambda \sum_{r=1}^R (\| \hat{\beta}_r \| - \| \hat{\beta}_r \|).$$

(20)

By Hölder’s inequality, we have that

$$\sum_{t=1}^T \varepsilon_t^T \Phi_t^T (\hat{\beta} - \beta) U_t \leq \| \sum_{t=1}^T \Phi_t \varepsilon_t U_t \|_{2,\infty} \| \hat{\beta} - \beta \|_{2,1}$$

(21)

where $\| \sum_{t=1}^T \Phi_t \varepsilon_t U_t \|_{2,\infty} = \max_{1 \leq r \leq R} \sqrt{\sum_{i=1}^T \sum_{k=1}^K \sum_{j=1}^J \Phi_{tkj} \varepsilon_{tj} U_{tr}}^2$.

Consider the random event

$$A = \left\{ 2(JT)^{-1} \| \sum_{t=1}^T \Phi_t \varepsilon_t U_t \|_{2,\infty} \leq \lambda \right\}.$$  

(22)

Since $\Phi_t \Phi_t^T / J = I_K$ and $\sum_{t=1}^T U_t^T U_t / R = 1$, the random variables

$V_{tr} = (\sqrt{J} \sigma)^{-1/2} \sum_{k=1}^K \sum_{j=1}^J \Phi_{tkj} \varepsilon_{tj} U_{tr}$, $t = 1, \ldots, T$, are i.i.d. standard Gaussian. Using this fact, we can write, for any $r = 1, \ldots, R$, and $\lambda = 2\sigma / \sqrt{JT} \left( 1 + A \log R / \sqrt{T} \right)^{1/2}$,

$$P \left\{ \sum_{t=1}^T \sum_{k=1}^K \sum_{j=1}^J \Phi_{tkj} \varepsilon_{tj} U_{tr} \right\} \geq \lambda^2 (JT)^2 / 4 \right\} = P \left\{ \chi_T^2 \geq \lambda^2 JT^2 / (4\sigma^2) \right\} = P \left( \chi_T^2 \geq T + A\sqrt{T} \log R \right)$$

where $\chi_T^2$ is a chi-square random variable with $T$ degrees of freedom. By the tail property of $\chi_T^2$ distribution (Lemma A.1 of Lounici et al. (2009)), and the fact that $A > 8$ we get:

$$P(A^c) \leq R \exp \{-A \log R / 8 \min(\sqrt{T}, A \log R) \} \leq R^{1-q}$$

with $q = \min(A \log R, \sqrt{T})$. It follows from (20) and (21) that, on the event $A$:

$$(JT)^{-1} \sum_{t=1}^T \| \Phi_t^T (\hat{\beta} - \beta^*) U_t \|^2 + \lambda \sum_{r=1}^R \| \hat{\beta}_r - \beta_r \|$$

$$\leq (JT)^{-1} \sum_{t=1}^T \| \Phi_t^T (\beta - \beta^*) U_t \|^2 + 2\lambda \sum_{r=1}^R (\| \hat{\beta}_r - \beta_r \| + \| \beta_r \| - \| \hat{\beta}_r \|)$$

$$\leq (JT)^{-1} \sum_{t=1}^T \| \Phi_t^T (\beta - \beta^*) U_t \|^2 + 2\lambda \sum_{r \in R(\beta)} (\| \hat{\beta}_r - \beta_r \| + \| \beta_r \| - \| \hat{\beta}_r \|)$$

$$+ 2\lambda \sum_{r \in R^c(\beta)} (\| \hat{\beta}_r - \beta_r \| + \| \beta_r \| - \| \hat{\beta}_r \|)$$

$$\leq (JT)^{-1} \sum_{t=1}^T \| \Phi_t^T (\beta - \beta^*) U_t \|^2 + 4\lambda \sum_{r \in R(\beta)} \| \hat{\beta}_r - \beta_r \|$$

(23)
which coincides with (11). To prove (12), we use (8) and (9) resulting in the inequality

\[(JT)^{-1} \max_{1 \leq r \leq R} \left\| \sum_{t=1}^{T} \{\Psi_t(Y_t - \Psi_t^\top \hat{\beta} U_t)U_t^\top\}_r \right\| \leq \lambda.\]  

(24)

Then

\[(JT)^{-1} \| \sum_{t=1}^{T} \{\Psi_t \Psi_t^\top (\hat{\beta} - \beta^*)U_t U_t^\top\}_r \| \leq (JT)^{-1} \| \sum_{t=1}^{T} \{\Psi_t(Y_t - \Psi_t^\top \hat{\beta} U_t)U_t^\top\}_r \| + (JT)^{-1} \| \sum_{t=1}^{T} (\Psi_t \varepsilon_t U_t^\top)_r \| \]

(25)

where we have used \(Y_t = \Psi_t^\top \beta^* U_t + \varepsilon_t\) and the triangle inequality. The derived bound (12) then follows by combining (25) with (24) and using the definition of the event \(A\). Finally, we prove (13). First, observe that,

\[\sum_{t=1}^{T} \Psi_t(Y_t - \Psi_t^\top \beta^* U_t)U_t^\top = \sum_{t=1}^{T} \Psi_t \Psi_t^\top (\hat{\beta} - \beta^*)U_t U_t^\top + \sum_{t=1}^{T} \Psi_t \varepsilon_t U_t^\top.\]

On the event \(A\), following from (8) and the triangle inequality, we have:

\[(JT)^{-1} \| \sum_{t=1}^{T} \{\Psi_t \Psi_t^\top (\hat{\beta} - \beta^*)U_t U_t^\top\}_r \| \geq \lambda/2, \text{ if } \hat{\beta}_r \neq 0.\]

The following arguments yield the bound (13) on the number of nonzero rows of \(\hat{\beta}_r^\top:\n
\[M(\hat{\beta}) \leq \frac{4}{\lambda^2 (JT)^2} \sum_{r \in R(\hat{\beta})} \| \sum_{t=1}^{T} \{\Psi_t \Psi_t^\top (\hat{\beta} - \beta^*)U_t U_t^\top\}_r \|^2\]

\[\leq \frac{4}{\lambda^2 (JT)^2} \sum_{r=1}^{R} \| \sum_{t=1}^{T} \{\Psi_t \Psi_t^\top (\hat{\beta} - \beta^*)U_t U_t^\top\}_r \|^2\]

\[= \frac{4}{\lambda^2 T^2} \| \sum_{t=1}^{T} \{J^{-1} \Psi_t \Psi_t^\top (\hat{\beta} - \beta^*)U_t U_t^\top\} \|^2_{2,1}\]

\[\leq \frac{4}{\lambda^2 T^2} \| \hat{\beta} - \beta^* \|_{2,1}^2 \sum_{t=1}^{T} U_t U_t^\top \|_{2,1}\]

\[\leq \frac{4 \phi_{\text{max}}^2}{\lambda^2 T^2} \| \hat{\beta} - \beta^* \|_{2,1}^2\]

which follows from Lemma 7.1, \(\Psi_t \Psi_t^\top / J = I_K\) and \(\phi_{\text{max}}\) is the maximum eigenvalues of the matrix \(\sum_{t=1}^{T} U_t U_t^\top\). □

**Proof of Theorem 3.1** We proceed similarly to the proof of Theorem 3.1 in Lounici et al. (2009) and Theorem 6.2 in Bickel et al. (2009). Let \(R = R(\beta^*) = \{r : \beta^*_r \neq 0\}\)

By inequality (11) in Lemma 3.1 with \(\beta = \beta^*\) we have, on the event \(A\) defined in (22):

\[(JT)^{-1} \sum_{t=1}^{T} \| \Psi_t^\top (\hat{\beta} - \beta^*)U_t \|^2 \leq 4\lambda \sum_{r \in R} \| \hat{\beta}_r - \beta^*_r \|\]

\[\leq 4\lambda \sqrt{s} \| (\hat{\beta} - \beta^*)_R \|\]

(26)
Moreover by the same inequality, on the event $A$, we have \( \sum_{r=1}^{R} \| \hat{\beta}_r - \beta^*_r \| \leq 4 \sum_{r \in \mathcal{R}} \| \hat{\beta}_r - \beta^*_r \| \), which implies that \( \sum_{r \in \mathcal{R}} \| \hat{\beta}_r - \beta^*_r \| \leq 3 \sum_{r \in \mathcal{R}} \| \hat{\beta}_r - \beta^*_r \| \). Thus, by Assumption 3.1 with $\Delta = (\hat{\beta} - \beta^*)$:

\[
\| (\hat{\beta} - \beta^*) \|_{\mathcal{R}} \leq \sum_{t=1}^{T} \| \Psi_t^\top (\hat{\beta} - \beta^*) U_t \| / (\kappa \sqrt{J}).
\]  

(27)

Now (14) follows from (26) and (27). Inequality (15) follows by noting that

\[
\sum_{r=1}^{R} \| \hat{\beta}_r - \beta^*_r \| \leq 4 \sum_{r \in \mathcal{R}} \| \hat{\beta}_r - \beta^*_r \| \leq 4 \sqrt{s} \| (\hat{\beta} - \beta^*) \|_{\mathcal{R}}
\]

and then using (14). Inequality (16) follows from (13) and (14). □

**Proof of Theorem 3.2** The proofs of this theorem are similar to the one of Theorem 3.1 up to a modification of the bound on $P(A^c)$ in Lemma 3.1. We consider now the event

\[
A = \left\{ \max_{1 \leq r \leq R} \left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr} \right\}^{1/2} \leq \lambda JT \right\}.
\]

The Markov inequality yields that

\[
P(A^c) \leq \sum_{t=1}^{T} E \left\{ \max_{1 \leq r \leq R} \left\{ \sum_{k=1}^{K} \sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr} \right\}^{1/2} / (\lambda JT)^2 \right\}.
\]

Then we use Nemirovski’s inequality, see Corollary 2.4 of Dümbgen et al. (2008)[p.5], with the random vectors

\[
W_{ij} = \left( \sum_{k=1}^{K} \Psi_{tkj} \varepsilon_{tj} U_{t1}/J, \ldots, \sum_{k=1}^{K} \Psi_{tkj} \varepsilon_{tj} U_{tR}/J \right) \in \mathbb{R}^R, \ \forall j, \ \forall t.
\]

We get that

\[
P(A^c) \leq \frac{2e \log R - e}{\lambda^2 JT} \sigma^2 (JT)^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J} \left( \max_{1 \leq r \leq R} \left| \sum_{k=1}^{K} \Psi_{tkj} U_{tr} \right| \right)^2.
\]

By the definition of $\lambda$ in Theorem 3.2 and Assumption 3.2 we obtain

\[
P(A^c) \leq \frac{(2e \log R - e)C}{(\log R)^{1+\delta}}. \ \ \ \ \ \□
\]

**Proof of Theorem 3.3** The proofs of this theorem are similar to the one of Theorem 3.1 up to a modification of the bound on $P(A^c)$ in Lemma 3.1. We consider now the event

\[
A = \left\{ \max_{1 \leq r \leq R} \left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr} \right\}^{1/2} \leq \lambda JT \right\}.
\]
Thus, following the fact that different space basis $\Psi_k$ and $\Psi_{k'}$ are independent, we have:

\[
P(A^c) = P \left[ \max_{1 \leq r \leq R} \left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{j=1}^{J} (\Psi_{tkj} \varepsilon_{tj} U_{tr})^2 \right\}^{1/2} > \lambda J T \right]
\]

\[
= P \left[ \max_{1 \leq r \leq R} \left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{j=1}^{J} (\Psi_{tkj} \varepsilon_{tj} U_{tr})^2 \right\}^{1/2} > \lambda J T \right]
\]

\[
\leq R P \left[ \left\{ \sum_{t=1}^{T} \left( \sum_{k=1}^{K} \sum_{j=1}^{J} (\Psi_{tkj} \varepsilon_{tj} U_{tr})^2 \right) \right\}^{1/2} > \lambda J T \right]
\]

\[
= R P \left[ \left\{ \sum_{t=1}^{T} \left( \sum_{k=1}^{K} \sum_{j=1}^{J} (\Psi_{tkj} \varepsilon_{tj} U_{tr})^2 \right) \right\}^{1/2} > \lambda J T \right] = R P \{ f(V) > \lambda \}
\]

where

\[
V_t \overset{\text{def}}{=} J^{-1} \sum_{k=1}^{K} \sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr}
\]

\[
V \overset{\text{def}}{=} (V_{1r}, \ldots, V_{Tr})
\]

\[
f(V) \overset{\text{def}}{=} T^{-1} \left( \sum_{t=1}^{T} V_t^2 \right)^{1/2}
\]

Since Assumption 3.3 holds, i.e. with a high probability $p$, for $\forall t$ and $v_{1r}, \ldots, v_{Tr}, v'_{tr}$,

\[
|f(v_{1r}, \ldots, v_{tr}, \ldots, v_{Tr}) - f(v_{1r}, \ldots, v'_{tr}, \ldots, v_{Tr})| \leq b_t^2 / T
\]

\[
\mathbb{E} f(V) \leq C' \sqrt{T}
\]

Then, by the (extended) Mcdiarmid inequality, see Theorem 2.1 of Janson (2004), with the random vectors $V$ and function $f$, we have

\[
P(A^c) \leq R P\{ f(V) > \lambda \} \leq R P\{ f(V) - \mathbb{E} f(V) > \lambda - C' \sqrt{T} \}
\]

\[
\leq R \exp \left\{ - \frac{(\lambda - C' \sqrt{T})^2 T^2}{X^*(T) \sum_t b_t^2} \right\} = R^{-\delta'}
\]

with $\lambda = C' \sqrt{T} + \sqrt{\frac{X^*(T) \sum_t b_t^2}{(\log R)^{1 - \delta'/2} T^2}}$, $\delta' > 0$. □

**Proof of Theorem 4.1** Similar to $\tilde{Y}_t \overset{\text{def}}{=} Y_t^T - U_t^T \beta \Psi_t$, define $\tilde{Y}^*_t \overset{\text{def}}{=} Y_t^T - U_t^T \beta^* \Psi_t$ with the corresponding estimate $\tilde{Z}_{0,t}$. Thus

\[
\frac{1}{T} \sum_{1 \leq t \leq T} \left\| \tilde{Z}_{0,t}^T \hat{A} - Z_{0,t}^T A^* \right\|^2 \leq \frac{1}{T} \sum_{1 \leq t \leq T} \left\| \tilde{Z}_{0,t}^T \hat{A} - \tilde{Z}_{0,t}^T \hat{A} \right\|^2 + \frac{1}{T} \sum_{1 \leq t \leq T} \left\| \tilde{Z}_{0,t}^T \hat{A} - Z_{0,t}^T A^* \right\|^2,
\]
where the second term is bounded by $O_P(\rho^2 + \delta_K^2)$ by Theorem 2 of Park et al. (2009). For the first term, since
\[
\hat{Z}_{0,t} = (A^T \Psi_t \Psi_t^T)^{-1} \hat{\Psi}_t Y_t
\]
\[
\tilde{Z}_{0,t} = (A^T \hat{\Psi}_t \hat{\Psi}_t^T)^{-1} \hat{\Psi}_t \hat{Y}_t
\]
\[
\tilde{Z}_{0,t} - \hat{Z}_{0,t} = (A^T \hat{\Psi}_t \hat{\Psi}_t^T)^{-1} \hat{\Psi}_t \{\Psi_t^T (\hat{\beta} - \beta^*) U_t\}
\]
and Theorem 3.3 tells us that $(JT)^{-1} \sum_{t=1}^T \| \Psi_t^T (\hat{\beta} - \beta^*) U_t \|^2$ could be arbitrary small, i.e. $\exists$ large enough $R$, s.t. the first term is dominated by the second one. $\square$

**Proof of Theorem 4.2** The proof is in a similar spirit of the one of Theorem 3 in Park et al. (2009). We will prove the first equation of the theorem for $h \neq 0$. The second equation follows from the first equation. We first prove that the matrix $T^{-1} \sum_{t=1}^T Z_{0,t} \hat{Z}_{0,t}^T$ is invertible. Suppose that the assertion is not true. We can choose a random vector $e$ such that $\|e\| = 1$ and $e^T \sum_{t=1}^T Z_{0,t} \hat{Z}_{0,t}^T = 0$. Note that
\[
\|T^{-1} \sum_{t=1}^T Z_{0,t} \hat{Z}_{0,t}^T - T^{-1} \sum_{t=1}^T Z_{0,t} Z_{0,t}^T A^* \|
\leq T^{-1} \sum_{t=1}^T \| Z_{0,t} (\hat{Z}_{0,t}^T \hat{A} - Z_{0,t}^T A^*) \|
\leq (T^{-1} \sum_{t=1}^T \| Z_{0,t} \|^2)^{1/2} (T^{-1} \sum_{t=1}^T \| \hat{Z}_{0,t}^T \hat{A} - Z_{0,t}^T A^* \|^2)^{1/2}
= O_P(\rho + \delta_K),
\]
because of Assumption (4.1.5) and Theorem 4.1. Thus with $f = T^{-1} \sum_{t=1}^T Z_{0,t} Z_{0,t}^T e$, we obtain
\[
\| f^T m \| = \| f^T (A^* \Psi) \| + O_P(\delta_K)
= \| e^T T^{-1} \sum_{t=1}^T Z_{0,t} Z_{0,t}^T \hat{A} \Psi \| + O_P(\rho + \delta_K)
= O_P(\rho + \delta_K).
\]
This implies that $m_1, \ldots, m_L$ are linearly dependent, contradicting to the construction that all space basis are independent. $Z_{0,t} = B^T \hat{Z}_{0,t}$ and $\hat{A} = B^{-1} A$ give with (28)
\[
\| \hat{A} - A^* \| = \| T^{-1} \sum_{t=1}^T Z_{0,t} Z_{0,t}^T (\hat{A} - A^*) \| O_P(1)
= \| T^{-1} \sum_{t=1}^T Z_{0,t} \tilde{Z}_{0,t}^T \tilde{A} - T^{-1} \sum_{t=1}^T Z_{0,t} Z_{0,t}^T A^* \| O_P(1)
= O_P(\rho + \delta_K)
\]

(29)
From Assumptions (4.1.4), (29) and Theorem 4.1, we get
\[
T^{-1} \sum_{t=1}^{T} \| \tilde{Z}_{t}^\top - Z_{0,t} \|^2
\]
\[
= T^{-1} \sum_{t=1}^{T} \| \tilde{Z}_{t}^\top (m_1, \ldots, m_L) - Z_{0,t}^\top (m_1, \ldots, m_L) \|^2 O_P(1)
\]
\[
= T^{-1} \sum_{t=1}^{T} \| \tilde{Z}_{t}^\top A^* - \tilde{Z}_{t}^\top \tilde{A} \|^2 O_P(1)
\]
\[
+ T^{-1} \sum_{t=1}^{T} \| \tilde{Z}_{t}^\top \tilde{A} - Z_{0,t}^\top A^* \|^2 O_P(1) + O_P(\delta_k^2)
\]
\[
\leq T^{-1} \sum_{t=1}^{T} \| \tilde{Z}_{0,t} - Z_{0,t} \|^2 \| \tilde{A} - A^* \|^2 O_P(1)
\]
\[
+ T^{-1} \sum_{t=1}^{T} \| Z_{0,t} \|^2 \| \tilde{A} - A^* \|^2 O_P(1)
\]
\[
+ T^{-1} \sum_{t=1}^{T} \| \tilde{Z}_{t}^\top \tilde{A} - Z_{0,t}^\top A^* \|^2 O_P(1) + O_P(\delta_k^2)
\]
\[
= O_P(\rho^2 + \delta_k^2).
\]

We will show that for \( h \neq 0 \)
\[
T^{-1} \sum_{t=h+1}^{T} \left\{ (\tilde{Z}_{0,t+h} - Z_{0,t+h}) - (\tilde{Z}_{0,t} - Z_{0,t}) \right\} Z_{0,t}^\top = O_P(T^{-1/2})
\]
\[
(31)
\]
This implies the first statement of Theorem 4.2, because by (30)
\[
T^{-1} \sum_{t=-h+1}^{T} (\tilde{Z}_{0,t} - Z_{0,t}) (\tilde{Z}_{0,t+h} - Z_{0,t+h}) = O_P(b^2) = O_P(T^{-1/2}).
\]

For the proof of (31), define
\[
\tilde{S}_{t,Z} = J^{-1} \sum_{j=1}^{J} \tilde{A} \Psi(X_{t,j}) \Psi(X_{t,j})^\top \tilde{A}^\top
\]
\[
S_{t,Z} = A^* E \left\{ \Psi(X_{t,j}) \Psi(X_{t,j})^\top \right\} A^*^\top
\]
\[
\tilde{S}_\alpha = (JT)^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J} \{ \Psi(X_{t,j}) \otimes \tilde{Z}_{0,t} \} \{ \Psi(X_{t,j}) \otimes \tilde{Z}_{0,t} \}^\top
\]
\[
S_\alpha = T^{-1} \sum_{t=1}^{T} E \left[ \{ \Psi(X_{t,j}) \otimes Z_{0,t} \} \{ \Psi(X_{t,j}) \otimes Z_{0,t} \}^\top \right] Z_{0,t}
\]
\[
S = J^{-1} A^* \left[ \Psi(X_{t,j}) \Psi(X_{t,j})^\top e - E \left\{ \Psi(X_{t,j}) \Psi(X_{t,j})^\top e \right\} \right],
\]
where \( e \in \mathbb{R}^K \) with \( \|e\| = 1 \). Let \( \tilde{a} \) be the stack form of \( \tilde{A} \). It can be verified that

\[
\tilde{Z}_{0,t} = \tilde{S}_{t,Z}^{-1}J^{-1} \sum_{j=1}^{J} \{Y_{t,j}A\Psi(X_{t,j})\},
\]

(32)

\[
\tilde{a} = \tilde{S}_\alpha^{-1}(JT)^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J} \{\Psi(X_{t,j}) \otimes \tilde{Z}_{0,t}\} Y_{t,j}.
\]

(33)

Let \( \gamma = T^{-1/2}/b \). We argue that

\[
\sup_{1 \leq t \leq T} \|\tilde{S}_{t,Z} - S_{t,Z}\| = o_P(\gamma), \quad \|\tilde{S}_\alpha - S_\alpha\| = o_P(\gamma).
\]

(34)

We show the first part of (34). The second part can be shown similarly. Since

\[
\tilde{A}\Psi_t\Psi_t^\top \tilde{A}^\top = (\tilde{A} - A^* + A^*) (\Psi_t\Psi_t^\top - E \Psi_t\Psi_t^\top + E \Psi_t\Psi_t^\top) (\tilde{A} - A^* + A^*)^\top,
\]

to prove the first part it suffices to show that, uniformly for \( 1 \leq t \leq T \),

\[
J^{-1} \sum_{j=1}^{J} A^* \{\Psi(X_{t,j})\Psi(X_{t,j})^\top - E \{\Psi(X_{t,j})\Psi(X_{t,j})^\top\}\} (\tilde{A} - A^*)^\top = o_P(\gamma)
\]

(35)

\[
J^{-1} \sum_{j=1}^{J} (\tilde{A} - A^*) \{\Psi(X_{t,j})\Psi(X_{t,j})^\top - E \{\Psi(X_{t,j})\Psi(X_{t,j})^\top\}\} (\tilde{A} - A^*)^\top = o_P(\gamma)
\]

(36)

\[
J^{-1} \sum_{j=1}^{J} A^* \{\Psi(X_{t,j})\Psi(X_{t,j})^\top - E \{\Psi(X_{t,j})\Psi(X_{t,j})^\top\}\} A^*^\top = o_P(\gamma)
\]

(37)

\[
J^{-1} \sum_{j=1}^{J} A^* E \{\Psi(X_{t,j})\Psi(X_{t,j})^\top\} (\tilde{A} - A^*)^\top = o_P(\gamma)
\]

(38)

\[
J^{-1} \sum_{j=1}^{J} (\tilde{A} - A^*) E \{\Psi(X_{t,j})\Psi(X_{t,j})^\top\} (\tilde{A} - A^*)^\top = o_P(\gamma)
\]

(39)

The proof of (35)-(37) follows by simple arguments. We now show (38). Claim (39) can be shown similarly. For the proof of (38), we use Bernstein’s inequality for the following sum:

\[
P \left( \left| \sum_{j=1}^{J} W_j \right| > x \right) \leq 2 \exp \left( - \frac{1}{2} \frac{x^2}{V + Mx/3} \right).
\]

(40)

Here for a value of \( t \) with \( 1 \leq t \leq T \), the random variable \( W_j \) is an element of the \( L \times 1 \)-matrix

\[
S = J^{-1} A^* \{\Psi(X_{t,j})\Psi(X_{t,j})^\top e - E \{\Psi(X_{t,j})\Psi(X_{t,j})^\top e\}\}
\]

where \( e \in \mathbb{R}^K \) with \( \|e\| = 1 \). In (40), \( V \) is an upper bound for the variance of \( \sum_{j=1}^{J} W_j \) and \( M \) is a bound for the absolute values of \( W_j \), i.e. \( |W_j| \leq M \) for \( 1 \leq j \leq J \), a.s. With some constants \( C_1 \) and \( C_2 \) that do not depend on \( t \) and the row number we get \( V \leq C_1 J^{-1} \) and \( M \leq C_2 K^{1/2} J^{-1} \). Application of Bernstein’s inequality gives that, uniformly for \( 1 \leq t \leq T \) and \( e \in \mathbb{R}^K \) with \( \|e\| = 1 \), all \( L \) elements of \( S \) are of order \( o_P(\gamma) \). This shows claim (35).
From (29), (30), (32), (33) and (34) it follows that uniformly for \(1 \leq t \leq T\),

\[
\tilde{Z}_{0,t} - Z_{0,t} = S_{t,Z}^{-1} J^{-1} \sum_{j=1}^{J} \tilde{\varepsilon}'_{t,j} A^* \Psi(X_{t,j}) + S_{t,Z}^{-1} J^{-1} \sum_{j=1}^{J} \tilde{\varepsilon}'_{t,j} (\tilde{A} - A^*) \Psi(X_{t,j}) + o_P(T^{-1/2})
\]

\[
\overset{\text{def}}{=} \Delta_{t,1,Z} + \Delta_{t,2,Z} + o_P(T^{-1/2}).
\]

(41)

For the proof of the theorem it remains to show that for \(1 \leq j \leq 2\)

\[
T^{-1} \sum_{t=-h+1}^{T} (\Delta_{t+h,j,Z} - \Delta_{t,j,Z}) Z_{0,t}^\top = o_P(T^{-1/2}).
\]

(42)

This can be easily checked for \(j = 1\). For \(j = 2\) it follows from \(\|\tilde{A} - A^*\| = O_P(\rho + \delta_K)\) and

\[
\mathbb{E} \left\{ \|(JT)^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J} \tilde{\varepsilon}'_{t,j} S_{t,Z}^{-1} \mathcal{M} \Psi(X_{t,j}) \|^2 \right\} = O(K(JT)^{-1}),
\]

for any \(L \times K\) matrix \(\mathcal{M}\) with \(\|\mathcal{M}\| = 1\). \(\square\)

References


