ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Song, Song; Härdle, Wolfgang Karl; Ritov, Ya'acov

Working Paper

High dimensional nonstationary time series modelling with generalized dynamic semiparametric factor model

SFB 649 Discussion Paper, No. 2010-039

Provided in Cooperation with:
Collaborative Research Center 649: Economic Risk, Humboldt University Berlin

Suggested Citation: Song, Song; Härdle, Wolfgang Karl; Ritov, Ya'acov (2010) : High dimensional nonstationary time series modelling with generalized dynamic semiparametric factor model, SFB 649 Discussion Paper, No. 2010-039, Humboldt University of Berlin, Collaborative Research Center 649 - Economic Risk, Berlin

This Version is available at:
https://hdl.handle.net/10419/56662

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

SFB 649 Discussion Paper 2010-039

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

Song Song*
Wolfgang K. Härdle**
Ya'acov Ritov***

*Humboldt-Universität zu Berlin & University of California, Berkeley
**Humboldt-Universität zu Berlin & National Central University
***The Hebrew University of Jerusalem

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

SFB 649 ECONOMIC RISK BERLIN

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

Song Song †, Wolfgang K. Härdle ‡, Ya'acov Ritov§

August 2, 2010

Abstract

(High dimensional) time series which reveal nonstationary and possibly periodic behavior occur frequently in many fields of science. 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, but it also poses an important question: 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 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 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

*Supported by Deutsche Forschungsgemeinschaft via SFB 649 "Ökonomisches Risiko", Humboldt-Universität zu Berlin. Ya'acov Ritov's research is supported by an ISF grant and a Humboldt Award. We especially would like to acknowledge the useful comments of Peter Bickel, Qiwei Yao and Enno Mammen, thank Wei Xu, Peter Mohr and Elena Silyakova for preparing the data and thank participants at numerous seminars and conferences for their discussions.

†Humboldt-Universität zu Berlin & University of California, Berkeley. Email: ssoonngg123@hotmail.com

‡Humboldt-Universität zu Berlin and National Central University

§The Hebrew University of Jerusalem

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. (1983) for Electroence-phalogram (EEG) analysis.

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

$$Y_{t,j} = Z_{t,1} m_{1,j} + \cdots + Z_{t,L} m_{L,j} + \varepsilon_{t,j}, \tag{1}$$

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:

$$\begin{aligned} Y_{t,j} &= \sum_{l=1}^{L} Z_{t,l}\ m_l(X_{t,j}) + \varepsilon_{t,j},\ \ 1 \le j \le J_t,\ 1 \le t \le T. \\ &\stackrel{\text{def}}{=} Z_t^\top m(X_{t,j}) + \varepsilon_{t,j} \end{aligned} \tag{2}$$

where $Z_t = (Z_{t,1}, \ldots, Z_{t,L})^\top$ (common factors) is an unobservable L-dimensional process (not necessarily stationary), m (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 $J_t \equiv J$ 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:

$$m_l(x) = \sum_{k=1}^{K} a_{lk} \psi_k(x) \tag{3}$$

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:

$$Z_{tl} = \sum_{r=1}^{R} \gamma_{rl} u_r(t) \tag{4}$$

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

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

$$Y_t^\top = \underbrace{U_t^\top \Gamma^*}_{Z_t^\top} \underbrace{A^* \Psi_t}_{m} + \varepsilon_t \stackrel{\text{def}}{=} U_t^\top \beta^{*\top} \Psi_t + \varepsilon_t. \tag{6}$$

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 ψ_k a space basis function. Γ^*, A^* and $\beta^{*\top}$ are $R \times L$, $L \times K$ and $R \times K$ (unknown) underlying coefficient matrices consisting of γ_{rl}, a_{lk} and β_{rk} respectively. For every β matrix, we introduce $\beta_r = (\beta_{kr}, 1 \le k \le K)$, 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 $\mathcal{R}(\beta) = \{r : \beta_r \neq 0\}$ and $M(\beta) = |\mathcal{R}(\beta)|$ where $|\mathcal{R}(\beta)|$ denotes the cardinality of set $\mathcal{R}(\beta)$. 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 β_rs 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 β^*) 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 ε_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 $(159 \cdot 730)^{-1} \sum_{s=-354}^{+365} \sum_{j=1}^{159} Y_{t+s,j}$ with $Y_{t,j}$ being the temperature of the jth 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 jth 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.

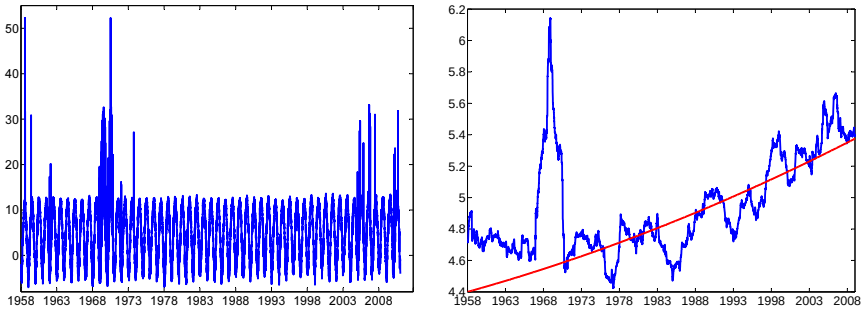

Figure 1: Temperatures of China from Jan 1st, 1957 to Dec 31st, 2009 (left) and the corresponding moving average (of 730 nearby days) view (right).

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 jth voxel. An example of the images at one particular time point is presented in Figure 2.

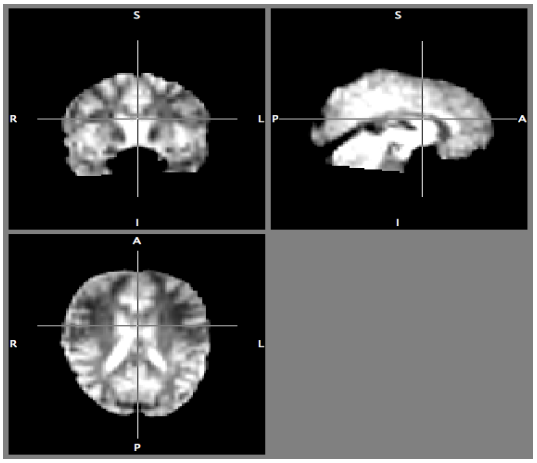

Figure 2: Typical fMRI data in one particular time point. The brightness corresponds to the strength of the observed signals.

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, heading 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.

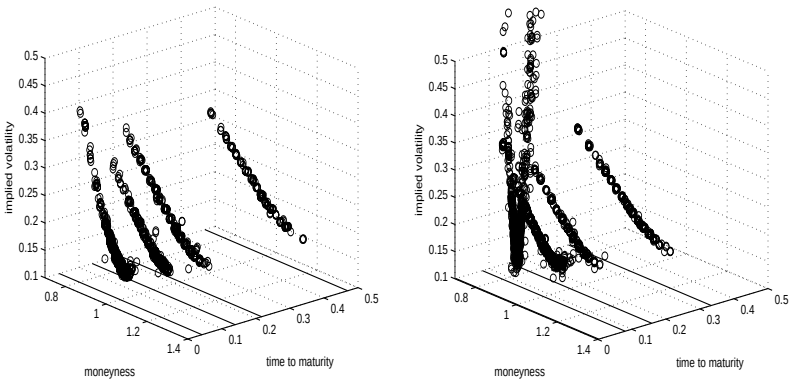

Figure 3: *The typical IV data design on two different days. In the maturity direction observations appear in the discrete points for each particular day. Bottom solid lines indicate the observed maturities. Left panel: observations on 20040701,* $J_t = 5606$. *Right panel: observations on 20040819,*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

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 $\sum_{t=1}^{T} u_r^2(t)/C_r^2 = 1$. 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 $d \geqslant 3$, 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).

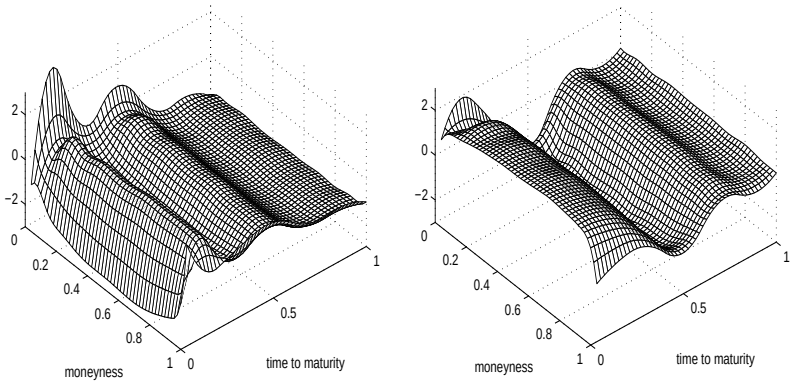

Figure 4: Space basis using the series estimator for the IVS modeling.

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, $\widehat{b}$, b_1, $\widehat{b}_1$, b_2 and $\widehat{b}_2$). Given $u \in [0,1]^d$, let h_μ and h_ϕ denote bandwidths, which could be selected as in the usual local polynomial regression setup and select $(\widehat{a}, \widehat{b}) = (a, b)$ to minimize

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

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

$$\sum_{t=1}^{T} \sum_{1 \leqslant j \neq k \leqslant J_t} \{Y_{tj} Y_{tk} - a_0 - \sum_{c=1}^{d} b_1^c (u^c - X_{tj}^c) - \sum_{c=1}^{d} b_2^c (v^c - X_{tk}^c)\}^2$$
$$\times K\Big(\frac{X_{tj} - u}{h_\phi}\Big) K\Big(\frac{X_{tj} - v}{h_\phi}\Big).$$

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

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

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 $\psi_1(X_{t,j}), \ldots, \psi_K(X_{t,j})$ ($\Psi_t \Psi_t^\top / J_t = I_K$ 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:

$$\min_\beta (JT)^{-1} \sum_{t=1}^{T} \left(Y_t^\top - U_t^\top \beta^\top \Psi_t\right) \left(Y_t^\top - U_t^\top \beta^\top \Psi_t\right)^\top + 2\lambda \|\beta\|_{2,1}. \tag{7}$$

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

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 size of the eigenvalues. But since it goes beyond the scope of this paper, we will therefore study its theoretical properties in a separate paper.

In order to study the statistical properties of this estimator, it is useful to derive some optimality condition for a solution of (7). Our implementation of the group Lasso-type estimator comes from Yuan and Lin (2006), which is an extension of the shooting algorithm of Fu (1998) for the lasso. As a direct consequence of the Karush-Kuhn-Tucker conditions, we have a necessary and sufficient condition for $\widehat{\beta}$ to be a solution to expression (7) is

$$(JT)^{-1} \sum_{t=1}^{T} \{\Psi_t (Y_t - \Psi_t^\top \widehat{\beta} U_t) U_t^\top\}_r = \lambda \frac{\widehat{\beta}_r}{\|\widehat{\beta}_r\|}, \text{ if } \widehat{\beta}_r \neq 0 \tag{8}$$

$$(JT)^{-1} \| \sum_{t=1}^{T} \{\Psi_t (Y_t - \Psi_t^\top \widehat{\beta} U_t) U_t^\top\}_r \| \leqslant \lambda, \qquad \text{if } \widehat{\beta}_r = 0 \tag{9}$$

Recall that $\Psi_t \Psi_t^\top / J = I_K$. It can be easily verified that the solution to (8) and (9) is

$$\widehat{\beta}_r = \left(1 - \lambda / \|S_r\|\right)_+ S_r, \tag{10}$$

where $S_r = \sum_{t=1}^{T} \{\Psi_t (Y_t - \Psi_t^\top \widehat{\beta}_{-r} U_t) U_t^\top\}_r$, with $\widehat{\beta}_{-r} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{r-1}, 0, \widehat{\beta}_{r+1}, \ldots, \widehat{\beta}_R)$. The solution to expression (7) can therefore be obtained by iteratively applying equation (10) to $r = 1, \ldots, R$. We choose the ordinary least square estimate $\widehat{\beta}_{OLS}$ as the initial value, with which usually a reasonable convergence tolerance is reached within 5 iterations. However, the computational burden increases dramatically as the number of initial basis increases.

Since the group Lasso type estimates depend on the unknown tuning parameter parameter λ, 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 λ's between 0 and $\lambda_{max} = \max_r \| \sum_t \Psi_t Y_t U_{tr} \| / \sqrt{K}$. We select the λ minimizing

$$\begin{aligned} C_p(\lambda) &= \frac{\sum_t \| Y_t^\top - U_t^\top \widehat{\beta}^\top \Psi_t \|^2}{\tilde{\sigma}^2} - JT + 2df \\ \tilde{\sigma}^2 &= \frac{\sum_t \| Y_t^\top - U_t^\top \widehat{\beta}_{OLS}^\top \Psi_t \|^2}{JT - df} \\ df &= \sum_r \mathbf{1}\{\| \widehat{\beta}_r \| > 0\} + \sum_r \frac{\| \widehat{\beta}_r \|}{\| \widehat{\beta}^{OLS} \|}(K-1) \end{aligned}$$

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 ε_t are Gaussian. Our main results concern upper bounds on the prediction error and the distance between the estimator and the true matrix β^*, (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 ε_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* $R \geqslant 2$ *and* $T, J \geqslant 1$. *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* $M(\beta^*) \leqslant s$. *Let*

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

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

$$(JT)^{-1}\sum_{t=1}^T \| \Psi_t^\top(\widehat{\beta} - \beta^*)U_t \|^2 + \lambda\|\widehat{\beta} - \beta\|_{2,1}$$
$$\leqslant (JT)^{-1}\sum_{t=1}^T \| \Psi_t^\top(\beta - \beta^*)U_t \|^2 + 4\lambda \sum_{r\in\mathcal{R}(\beta)} \|\widehat{\beta}_r - \beta_r\|, \tag{11}$$

$$(JT)^{-1}\max_{1\leqslant r\leqslant R} \|\sum_{t=1}^T \{\Psi_t\Psi_t^\top(\widehat{\beta} - \beta^*)U_tU_t^\top\}_r\| \leqslant \frac{3}{2}\lambda, \tag{12}$$

and

$$M(\widehat{\beta}) \leqslant \frac{4\phi_{max}^2}{\lambda^2 T^2}\|\widehat{\beta} - \beta^*\|_2^2, \tag{13}$$

where ϕ_{max} *is the maximum eigenvalue of the matrix* $\sum_{t=1}^T U_tU_t^\top$.

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*

$$\min\Big\{\frac{\sum_t \|\Psi_t^\top\Delta U_t\|}{\sqrt{J}\,\|\,\Delta_{\mathcal{R}}\,\|} : |\mathcal{R}| \leqslant s, \Delta \in \mathbb{R}^{K\times R}\backslash\{0\},$$
$$\|\,\Delta_{\mathcal{R}^c}\,\|_{2,1} \leqslant 3\,\|\,\Delta_{\mathcal{R}}\,\|_{2,1}\Big\} \geqslant \kappa,$$

where $\mathcal{R}^c$ *denotes the complement of the set of indices* $\mathcal{R}$, $\Delta_{\mathcal{R}}$ *denotes the matrix formed by stacking the rows of matrix* Δ *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 κ 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* $\widehat{\beta}$ *of* (7)*:*

$$(JT)^{-1} \sum_{t=1}^{T} \| \Psi_t^\top (\widehat{\beta} - \beta^*) U_t \|^2 \leqslant 64\sigma^2 s(1 + A \log R/\sqrt{T})/(\kappa^2 J), \tag{14}$$

$$T^{-1/2} \| \widehat{\beta} - \beta^* \|_{2,1} \leqslant 32\sigma s \sqrt{1 + A \log R/\sqrt{T}}/(\kappa^2 \sqrt{J}), \tag{15}$$

and

$$M(\widehat{\beta}) \leqslant 64\phi_{max}^2 s/\kappa^2 \tag{16}$$

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 β^*'s sparsity level is low (s large, κ small, s/κ^2 large), all the three bounds get larger and the number of nonzero rows of estimated one $\widehat{\beta}^\top$ is larger too correspondingly.

From now on, we only assume that the random variables ε_{tj} are independent with zero mean and finite variance $\mathsf{E}(\varepsilon_{tj}^2) \leqslant \sigma^2$. 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* Ψ_t *and* U_t *are such that*

$$(JT)^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J} \Big(\max_r | \sum_{k=1}^{K} \Psi_{tkj} U_{tr} | \Big)^2 \leqslant C,$$

for a constant $C > 0$*.*

THEOREM 3.2 *Consider the DSFM* (6) *for* $R \geqslant 3$ *and* $T, J \geqslant 1$*. Assume that the random vectors* $\varepsilon_1, \ldots, \varepsilon_T$ *are independent with zero mean and finite variance* $\mathsf{E}(\varepsilon_{tj}^2) \leqslant \sigma^2$*,* $\Psi_t \Psi_t^\top / J = I_K$*,* $\sum_{t=1}^{T} U_t^\top U_t / R = 1$*, and* $M(\beta^*) \leqslant s$*. Let also Assumption 3.2 be satisfied. Furthermore let* κ *be defined as in Assumption 3.1 and* ϕ_{max} *is the maximum eigenvalue of the matrix* $\sum_{t=1}^{T} U_t U_t^\top$*. 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* $\widehat{\beta}$ *of (7) we have:*

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

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

and

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

Since the standard assumption on ε_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 $\mathcal{T}$ and random variables V_t, $t \in \mathcal{T}$, we say:

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

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

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

ASSUMPTION 3.3 *The matrices Ψ_t and U_t and random variables ε_t are such that*

$$\begin{aligned} (J^{-1}\sum_{k=1}^{K}\sum_{j=1}^{J}\Psi_{tkj}\varepsilon_{tj}U_{tr})^2 &\leqslant b_t^2 \quad \textit{with a high probabilityp} \\ \mathsf{E}(JT)^{-1}\Big\{\sum_{t=1}^{T}(\sum_{k=1}^{K}\sum_{j=1}^{J}\Psi_{tkj}\varepsilon_{tj}U_{tr})^2\Big\}^{1/2} &\leqslant \frac{C'}{\sqrt{T}}. \end{aligned}$$

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 rs. Given $b_t, t = 1, \ldots, T$, 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 $R \geqslant 3$, $T, J \geqslant 1$ and $\mathcal{T} = \{1, \ldots, T\}$. 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 $M(\beta^*) \leqslant s$. Furthermore let κ be defined as in Assumption 3.1 and ϕ_{max} is the maximum eigenvalue of the matrix $\sum_{t=1}^{T} U_t U_t^\top$. Let*

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

Then with probability at least $p(1-R^{-\delta'})$, for any solution $\widehat{\beta}$ of (7) we have:

$$(JT)^{-1}\sum_{t=1}^{T}\parallel \Psi_t^{\top}(\widehat{\beta}-\beta^*)U_t \parallel^2 \leqslant 16\left(C'+\sqrt{\frac{\mathcal{X}^*(\mathcal{T})\sum_t b_t^2}{(\log R)^{1-\delta'}T}}\right)^2 s/\kappa^2$$

$$T^{-1/2}\parallel \widehat{\beta}-\beta^* \parallel_{2,1} \leqslant 16\left(C'+\sqrt{\frac{\mathcal{X}^*(\mathcal{T})\sum_t b_t^2}{(\log R)^{1-\delta'}T}}\right) s/\kappa^2$$

and

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

Not surprisingly, this theorem tells that the bounds get larger when the dependence level, i.e. $\mathcal{X}^*(\mathcal{T})$ increases, i.e. the bound is minimized when $\mathcal{X}^*(\mathcal{T})=1$.

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):

$$\begin{aligned} Z_{tl} &= \sum_{r=1}^{R}\gamma_{rl}u_r(t) \\ Y_t^{\top} &= (Z_{0,t}^{\top}+U_t^{\top}\Gamma)A\Psi_t+\varepsilon_t' = U_t^{\top}\Gamma A\Psi_t+(Z_{0,t}^{\top}A\Psi_t+\varepsilon_t'), \end{aligned} \tag{17}$$

with an unobservable L-dimensional random process $Z_{0,t}$ with $\mathsf{E}(Z_{0,t}|X_t)=0$ and i.i.d. assumption on ε_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 $Y_t^{\top}=U_t^{\top}\Gamma A\Psi_t+(Z_{0,t}^{\top}A\Psi_t+\varepsilon_t')$, treat $Z_{0,t}^{\top}A\Psi_t+\varepsilon_t'$ as the ε_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 $\widehat{\widehat{Y}}_t^{\top} \stackrel{\text{def}}{=} Y_t^{\top}-U_t^{\top}\widehat{\beta}\Psi_t$, $\widehat{A}$ and Ψ_t, use the ordinary least square method to obtain the estimated random process $\widehat{Z}_{0,t}$.

As we could see from step one here, since ε_t in (6) involves $Z_{0,t}^{\top}A\Psi_t+\varepsilon_t'$, 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 $\widehat{\beta}$ instead of β^*, 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 $X_{1,1},\ldots,X_{T,J}$, $\varepsilon_{1,1}',\ldots,\varepsilon_{T,J}'$, and $Z_{0,1},...,Z_{0,T}$ 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 $t=1,\ldots,T$.

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

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

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

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

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

4.1.6 The minimization (7) *runs over all values* β *with*

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

where the constant M_T *fulfils* $\max_{1\leqslant t\leqslant T}\|Z_{0,t}\| \leqslant M_T/C_m$ *(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*

$$\frac{1}{T}\sum_{1\le t\le T}\left\|\widehat{Z}_{0,t}^\top\widehat{A} - Z_{0,t}^\top A^*\right\|^2 = \mathcal{O}_P(\rho^2 + \delta_K^2). \tag{18}$$

In the following we discuss how a statistical analysis differs if the inference of stochasticity on $Z_{0,t}$ is based on $\widehat{Z}_{0,t}$ (note that the trend $U_t^\top\Gamma$ 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* $\mathsf{E}(Z_{0,t}) = 0$, $\mathsf{E}(\|Z_{0,t}\|^\gamma) < \infty$ *for some* $\gamma > 2$. *It is strongly mixing with* $\sum_{i=1}^\infty \alpha(i)^{(\gamma-2)/\gamma} < \infty$. *The matrix* $\mathsf{E}\,Z_{0,t}Z_{0,t}^\top$ *has full rank. The process* $Z_{0,t}$ *is independent of* $X_{11}, \ldots, X_{TJ}, \varepsilon'_{11}, \ldots, \varepsilon'_{TJ}$.

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

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 β, but not to Γ, A. If $(\widehat{Z}_{0,t}, \widehat{A})$ is a minimizer, then also $(B^\top \widehat{Z}_{0,t}, B^{-1}A)$ is a minimizer, where B is an arbitrary invertible matrix. In particular, with the choice $B = (\sum_{t=1}^T Z_{0,t}\widehat{Z}_{0,t})^{-1} \sum_{t=1}^T Z_{0,t}Z_{0,t}^\top$, we get for $\widetilde{Z}_{0,t} \stackrel{\text{def}}{=} B^\top \widehat{Z}_{0,t}$ and $\widetilde{A} \stackrel{\text{def}}{=} B^{-1}A$ that $\sum_{t=1}^T Z_{0,t}(\widetilde{Z}_{0,t} - Z_{0,t})^\top = 0$. Without loss of generality, we may assume $T^{-1}\sum_{s=1}^T \widehat{Z}_{0,s} = T^{-1}\sum_{s=1}^T Z_{0,s} = 0$. Additionally define

$$\begin{aligned}
\widetilde{Z}_{n,t} &= (T^{-1}\sum_{s=1}^T \widetilde{Z}_{0,s}\widetilde{Z}_{0,s}^\top)^{-1/2}\widetilde{Z}_{0,t} \\
Z_{n,t} &= (T^{-1}\sum_{s=1}^T Z_{0,s}Z_{0,s}^\top)^{-1/2}Z_{0,t}.
\end{aligned}$$

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 $h \geq 0$*

$$T^{-1}\sum_{t=\max[1,-h+1]}^{\min[T,T-h]} \widetilde{Z}_{0,t}\left(\widetilde{Z}_{0,t+h} - \widetilde{Z}_{0,t}\right)^\top - Z_{0,t}\left(Z_{0,t+h} - Z_{0,t}\right)^\top = o_P(T^{-1/2})$$

and

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

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:

$$\begin{aligned}
L_{par} &= 1 - \frac{\sum_{r=1}^R \|\widehat{\beta}_r - \beta_r\|_\infty}{\sum_{r=1}^R \|\beta_r\|_\infty} \\
L_{pre} &= 1 - \frac{\sum_{r=1}^R \|\Psi_t^\top(\widehat{\beta}_r - \beta_r)U_t\|_\infty}{\sum_{r=1}^R \|\Psi_t^\top \beta_r U_t\|_\infty}
\end{aligned}$$

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 $\beta_{rk} \sim \mathrm{N}\{0, \exp(-2k/5)\}, r = 1, \ldots, 179, k = 1, \ldots, 5$. We randomly pick $179 - M(\beta^*)$ β_rs 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:

$$\begin{aligned}
&m_0(x_1, x_2) = 1, \quad m_1(x_1, x_2) = 3.46(x_1 - .5), \\
&m_2(x_1, x_2) = 9.45\left\{(x_1 - .5)^2 + (x_2 - .5)^2\right\} - 1.6, \\
&m_3(x_1, x_2) = 1.41\sin(2\pi x_2).
\end{aligned}$$

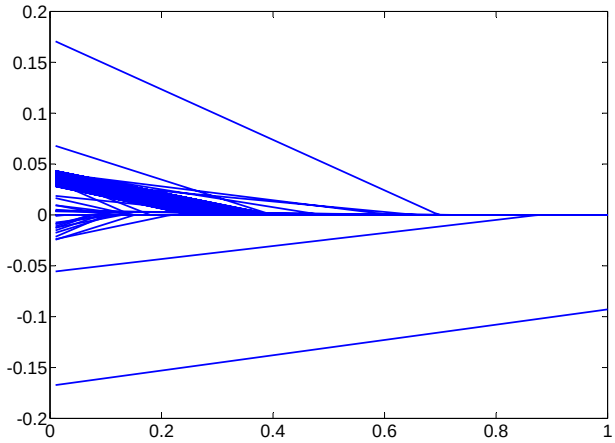

Figure 5: An illustration plot about how group Lasso penalty shrinks the coefficients.

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 $Y_t^\top = U_t^\top \beta^\top \Psi_t + \varepsilon_t, t = 1, \ldots, 19345$ where ε_t is drawn as i.i.d. $\mathrm{N}(0, 0.05)$.

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_{par} and L_{pre}s w.r.t. different sparsity levels. Our theoretical results in the previous sections suggest that when $M(\beta^*)$ (s) is small, L_{par} and L_{pre} will be large, which is confirmed by the simulation results.

	$M(\beta^*) = 100$	$M(\beta^*) = 50$	$M(\beta^*) = 20$
L_{par}	0.870	0.918	0.931
L_{pre}	0.710	0.835	0.859

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

The second experiment compares how L_{par} and L_{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_{par} and L_{pre}s are presented in Table 2.

	$R = 116$	$R = 179$	$R = 252$
L_{par}	0.879	0.918	0.920
L_{pre}	0.695	0.835	0.841

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

As we could see, when R increases, L_{par} and L_{pre}s 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_{par} and L_{pre} are sensitive to the dependence level of the error items. We generated ε_t from a centered VAR(1) process $\varepsilon_t = \mathcal{R}\varepsilon_{t-1} + U_t$, where U_t is $N_3(0, \Sigma_U)$ random vector, the rows of $\mathcal{R}$ 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 ε_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.

	$VAR(0)$	$VAR(1)$	$VAR(2)$
L_{par}	0.918	0.854	0.783
L_{pre}	0.835	0.774	0.712

Table 3: L_{par} and L_{pre} w.r.t. different levels of dependence of ε_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.

	Factors		Factors
Trend	1	Large	$\sin 2\pi t/(365 \cdot 10)$
(Year by Year)	t	Period	$\cos 2\pi t/(365 \cdot 10)$
	$3t^2 - 1$		$\sin 4\pi t/(365 \cdot 10)$
Seasonal	$\sin 2\pi t/365$		$\cos 4\pi t/(365 \cdot 10)$
Effect	$\cos 2\pi t/365$		$\sin 6\pi t/(365 \cdot 10)$
	$\cdots$		$\cdots$
	$\cos 20\pi t/365$		$\cos 20\pi t/(365 \cdot 10)$

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

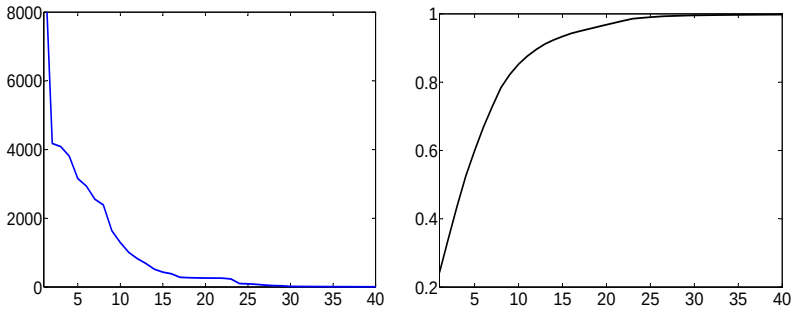

Figure 6: Distribution of the eigenvalues and the relative proportion of variance explained by the first K basis.

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

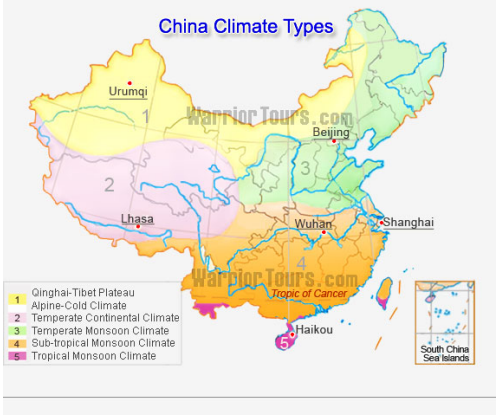

Figure 7: China Climate Types

series estimator. Figure 8 displays the estimated coefficients of the 5 factors with respect to the $54 \cdot 3$ yearly polynomial time basis under the optimal choice of λ. The coefficients of constant, linear and quadratic terms are displayed as solid, dashed and dotted lines correspondingly. As one may see, the fact that most of the coefficients are nonnegative (especially for $k = 1$) shows strong evidence of global warming effect (especially with a quadratic upward trend) in China during the past 50 years. In a climatological context this has also been observed by Karl et al. (1991), while the global climate change has been recently summarized by Gleick et al. (2010). The high estimates over the second half of 1960s are due to the high temperatures then in China (Figure 1). The pattern that all the coefficients display an upward trend further indicates the stronger and stronger warming effect. The coefficients estimates of the 20 Fourier series time basis corresponding to the optimal λ 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 $U_t^\top \widehat{\beta}$, 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.

Basis			Estimates		
$\sin 2\pi t/365$	-0.1777	0.0076	0.0177	-0.0136	0.0084
$\cos 2\pi t/365$	-0.6081	0.0126	0.0366	-0.0369	0.0114
$\sin 4\pi t/365$	0.0000	0.0000	0.0000	0.0000	0.0000
$\cos 4\pi t/365$	-0.0145	0.0028	0.0021	-0.0022	0.0029
$\ldots$	0.0000	$\ldots$			
$\cos 20\pi t/365$	0.0000	$\ldots$			
$\sin 2\pi t/(365 \cdot 10)$	0.0025	-0.0006	0.0009	-0.0008	-0.0001
$\cos 2\pi t/(365 \cdot 10)$	0.0000	$\ldots$			
$\ldots$	0.0000	$\ldots$			
$\cos 20\pi t/(365 \cdot 10)$	0.0000	$\ldots$			

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

Since the eigenvalues of $\widehat{\beta}\widehat{\beta}^\top$ 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 $\widehat{Z}_{0,t}$, e.g. $\widehat{Z}_{0,t,1}$ as displayed in Figure

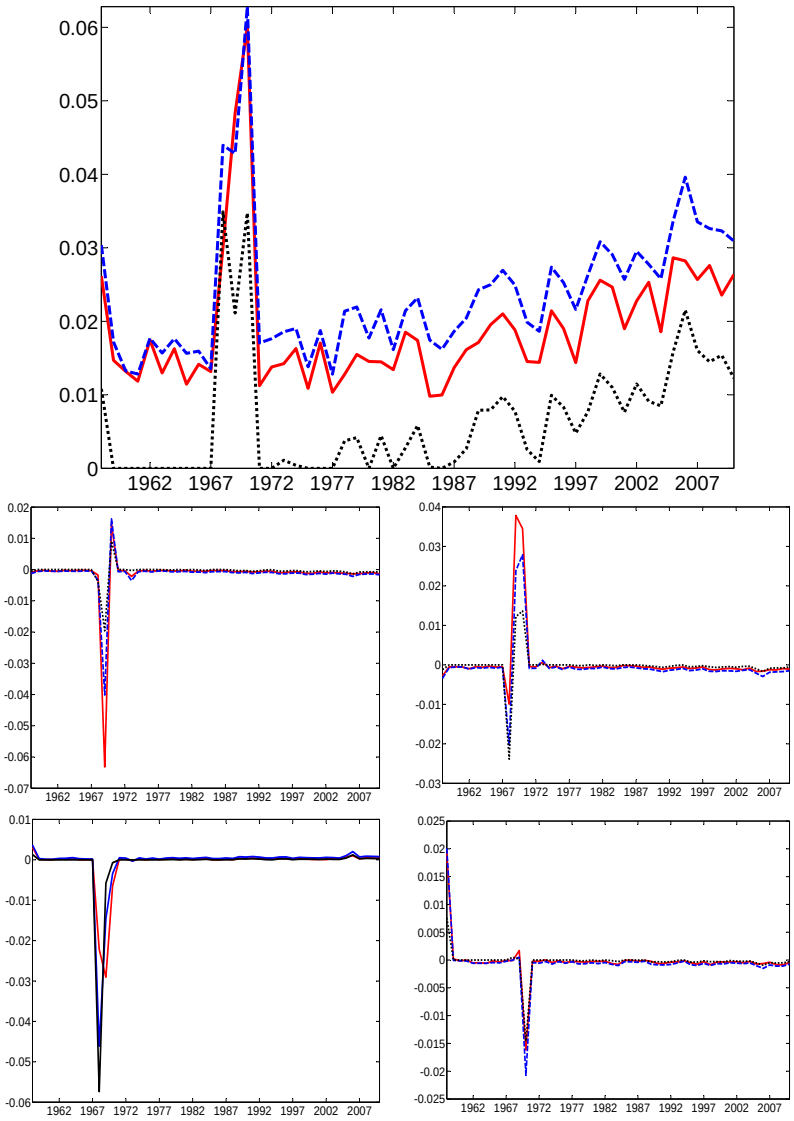

Figure 8: Estimated coefficients of the $54 \cdot 3$ yearly polynomial time basis w.r.t. $k = 1, \ldots, 5$ from up to down and left to right.

10 ($\widehat{Z}_{0,t,2}$ - $\widehat{Z}_{0,t,5}$ 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. The residual multi-dimensional random process could be further modeled by multivariate time series techniques. For example, if we use VAR(1) process $\widehat{Z}_{0,t} = \mathcal{R}\widehat{Z}_{0,t-1} + \varepsilon_{0,t}$, where $\varepsilon_{0,t}$ is a random vector, the estimated coefficient matrix is:

$$\begin{pmatrix} 0.9732 & -0.0135 & -0.0002 & -0.0006 & -0.0002 \\ 0.0127 & 0.1766 & -0.1824 & -0.0682 & -0.0009 \\ 0.0358 & -0.2867 & 0.4493 & -0.1138 & 0.0053 \\ -0.0001 & -0.1967 & -0.1962 & 0.8010 & -0.0052 \\ 0.0790 & 0.0492 & 0.0690 & -0.0225 & 0.8418 \end{pmatrix}.$$

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

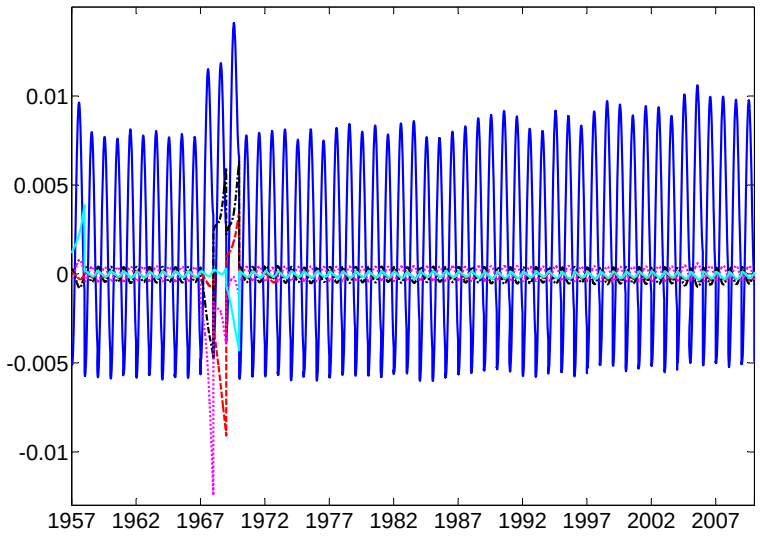

Figure 9: Extracted trends based on $U_t^\top \widehat{\beta}$.

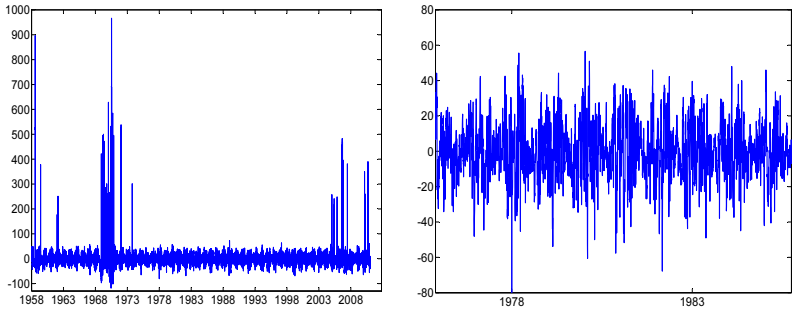

Figure 10: Estimated Stochastic Process $\widehat{Z}_{0,t,1}$ and a 10 year zoom.

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 $\widehat{Z}_{0,t}$ 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:

$$Y_t^\top = U_t^\top \Gamma A \Psi_t + \varepsilon_t, \quad \mathrm{Cov}(\varepsilon_t) = diag(|U_t^\top \Gamma A \Psi_t|),$$

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).

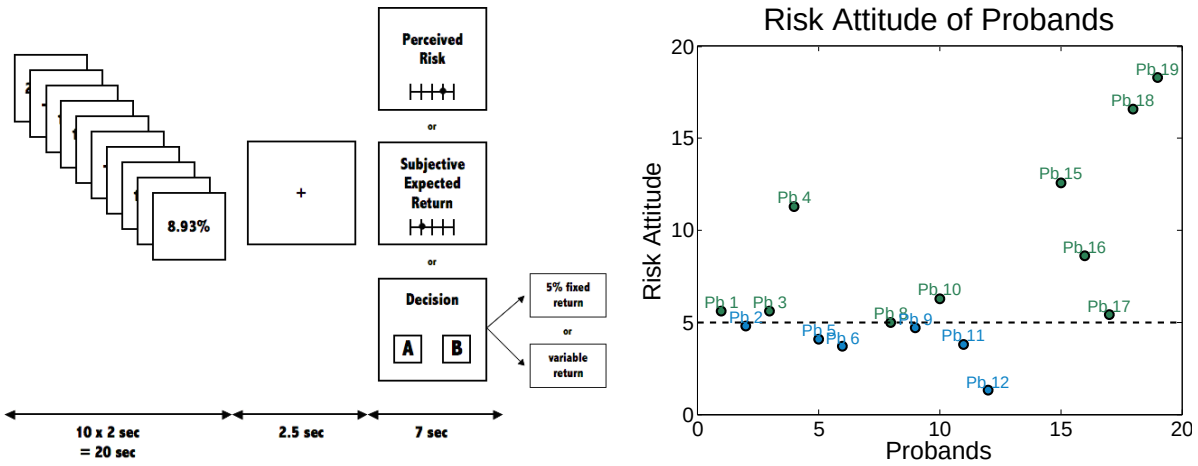

Figure 11: Flowchart of the experiment (left) "Returns Pause Decision" and risk attitudes of 16 subjects (right). Subjects with risk attitude < 5 are colored blue, otherwise green.

Since we are analyzing multi subjects $1 \leq i \leq I$ here, we obtain a panel version of the original model (6) to

$$Y_{t,j}^i = \sum_{l=1}^{L} (\alpha_{t,l}^i + U_t^\top \Gamma_l^i) m_l(X_{t,j}) + \varepsilon_{t,j}, \quad 1 \leq j \leq J_t, \; 1 \leq t \leq T,$$

where the fixed effect $\alpha_{t,l}^i$ is the individual effect on function m_l for subject i at time point t. For identification purpose, we assume
$\sum_{i=1}^{I} \sum_{l=1}^{L} \alpha_{t,l}^i m_l(X_{t,j}) = 0$. 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:

$$\overline{Y}_{t,j} = \sum_{l=1}^{L} (U_t^\top \overline{\Gamma}_l) m_l(X_{t,j}) + \varepsilon_{t,j} \; , \quad 1 \leq j \leq J,$$

and our 2-step estimation procedure is as follows:

1 Take the average of $Y_{t,j}^i$ 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 $Z_{t,l}^i$.

$$Y_{t,j}^i = \sum_{l=1}^{L} U_t^\top \Gamma_l^i \overline{m}_l(X_{t,j}) + \varepsilon_{t,j}^i$$

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 76176×76176 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 722×722 matrix. If we additionally take the average of every 10 $Y^i_{t,j}$s over t, we only need compute eigenvalues from a 73×73 matrix.

The third factor loading function $\widehat{m}_3$ shown in Figure 12 could be identified as the Ventromedial prefontal 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 Myšičková et al. (2010).

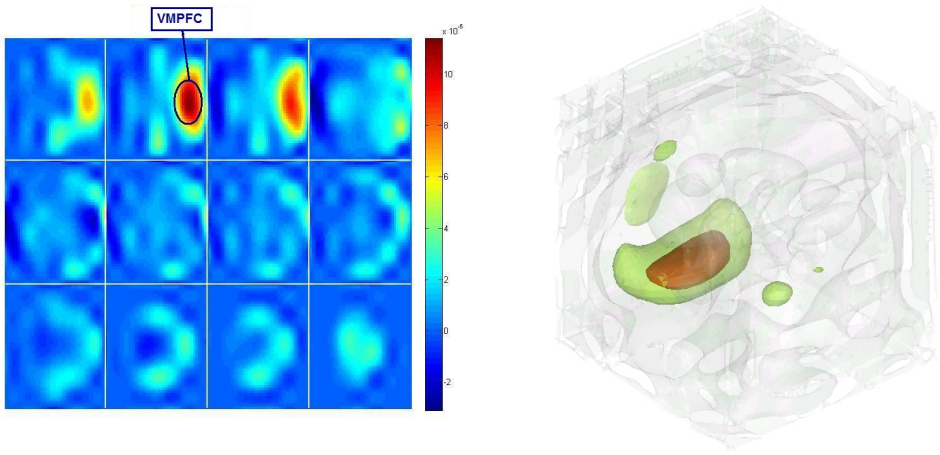

Figure 12: Estimated function $\widehat{m}_3$ shown in 12 axial slices (left) and as a 3D-plot in a posterior view (right) with highlighted Ventromedial prefrontal cortex (VMPFC).

We use the same 3 orthogonal polynomials and 10 Fourier series as before as time basis. Figure 13 displays the response curve (to stimuli) $U_t^\top \widehat{\Gamma}^i_2$ 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 $\widehat{Z}_{0,t}$ have different volatilities, for this purpose we use the estimated coefficients Γ^i_3 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 $(20 - 90)$.

MEAN		Estimated	
Data	Strongly	0.85	0.14
	Weakly	0.59	0.40

Table 6: Classification rates of the SVM method using median(left) and mean (right) of volatilities of $\Delta\widehat{Z}_{t,2}$.

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 is therefore omitted, alternatively, we could still use the dynamic semiparametric factor modeling

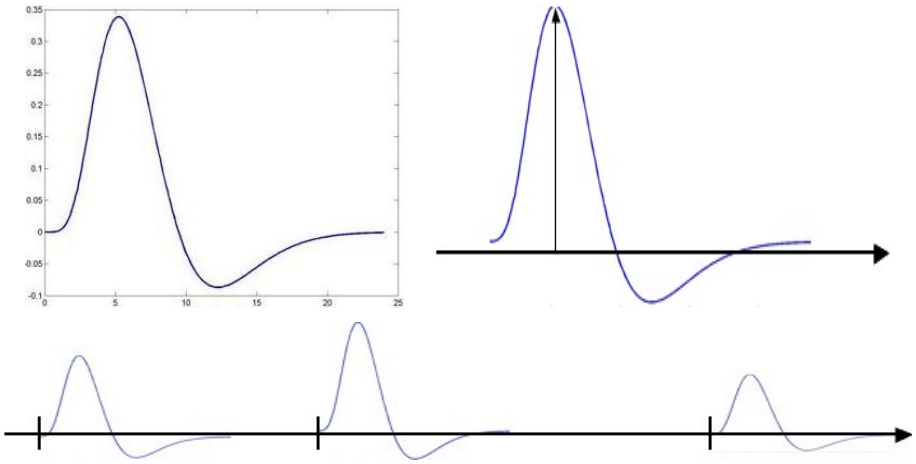

Figure 13: Response curve (to stimuli) $U_t^\top \widehat{\Gamma}_2^i$ for proband 18 (up) and 16, 19 and 11 (down, left to right) w.r.t $l = 3$.

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 $\widehat{Z}_{t,1}$, $\widehat{Z}_{t,2}$ (stationary) are presented in Figure 14 (right).

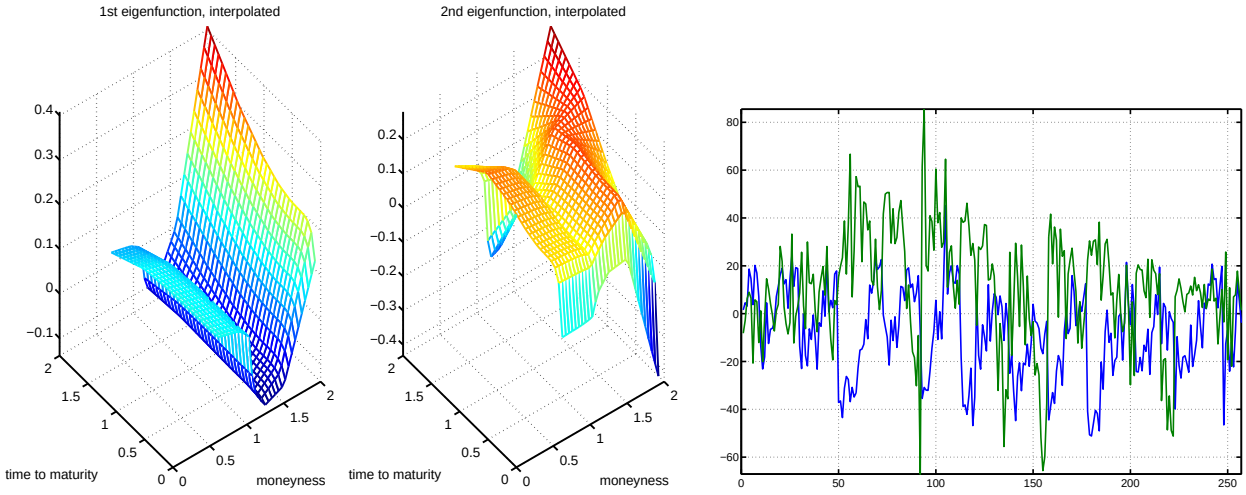

Figure 14: Space basis using the FPCA approach for IVS modeling and the estimated time series of factors $\widehat{Z}_{t,1}$, $\widehat{Z}_{t,2}$.

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 $\|AB\|_{2,1} \leqslant \|A\|_{2,1}\|B\|_{2,1}$.*

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

$$\begin{aligned}
\|AB\|_{2,1} &= \sum_{i=1}^{I}\sqrt{\sum_{k=1}^{K}(\sum_{j=1}^{J}a_{ij}b_{jk})^2} \\
&\leqslant \sum_{i=1}^{I}\sqrt{\sum_{k=1}^{K}(\sum_{j=1}^{J}a_{ij}^2\sum_{j=1}^{J}b_{jk}^2)} \\
&\leqslant (\sum_{i=1}^{I}\sqrt{\sum_{j=1}^{J}a_{ij}^2})(\sum_{k=1}^{K}\sqrt{\sum_{j=1}^{J}b_{jk}^2}) \\
&= \|A\|_{2,1}\|B\|_{2,1}
\end{aligned}$$

□

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 $\widehat{\beta}$ as a minimizer of (7), for $\forall\beta$ we have

$$\begin{aligned}
&(JT)^{-1}\sum_{t=1}^{T}\| \Psi_t^\top\widehat{\beta}U_t - Y_t \|^2 + 2\lambda\sum_{r=1}^{R}\|\widehat{\beta}_r\| \\
&\leqslant (JT)^{-1}\sum_{t=1}^{T}\| \Psi_t^\top\beta U_t - Y_t \|^2 + 2\lambda\sum_{r=1}^{R}\|\beta_r\|,
\end{aligned} \tag{19}$$

which, using $Y_t = \Psi_t^\top\beta^*U_t + \varepsilon_t$, is equivalent to

$$\begin{aligned}
(JT)^{-1}\sum_{t=1}^{T}\| \Psi_t^\top(\widehat{\beta}-\beta^*)U_t \|^2 &\leqslant (JT)^{-1}\sum_{t=1}^{T}\| \Psi_t^\top(\beta-\beta^*)U_t \|^2 \\
&+2(JT)^{-1}\sum_{t=1}^{T}\varepsilon_t^\top\Psi_t^\top(\widehat{\beta}-\beta)U_t + 2\lambda\sum_{r=1}^{R}(\|\beta_r\| - \|\widehat{\beta}_r\|).
\end{aligned} \tag{20}$$

By Hölder's inequality, we have that

$$\sum_{t=1}^{T}\varepsilon_t^\top\Psi_t^\top(\widehat{\beta}-\beta)U_t \leqslant \|\sum_{t=1}^{T}\Psi_t\varepsilon_tU_t^\top\|_{2,\infty}\|\widehat{\beta}-\beta\|_{2,1} \tag{21}$$

where $\|\sum_{t=1}^{T}\Psi_t\varepsilon_tU_t^\top\|_{2,\infty} = \max_{1\leqslant r\leqslant R}\sqrt{\sum_{t=1}^{T}\sum_{k=1}^{K}(\sum_{j=1}^{J}\Psi_{tkj}^\top\varepsilon_{tj}U_{tr})^2}$.

Consider the random event

$$\mathcal{A} = \Big\{2(JT)^{-1}\|\sum_{t=1}^{T}\Psi_t\varepsilon_tU_t^\top\|_{2,\infty} \leqslant \lambda\Big\}. \tag{22}$$

Since $\Psi_t\Psi_t^\top/J = I_K$ and $\sum_{t=1}^{T}U_t^\top U_t/R = 1$, the random variables $V_{tr} = (\sqrt{J}\sigma)^{-1/2}\sum_{k=1}^{K}\sum_{j=1}^{J}\Psi_{tkj}\varepsilon_{tj}U_{tr}$, $t = 1,\ldots,T$, are *i.i.d.* standard Gaussian. Using this fact, we can write, for any $r = 1,\ldots,R$, and $\lambda = 2\sigma/\sqrt{JT}\Big(1 + A\log R/\sqrt{T}\Big)^{1/2}$,

$$\begin{aligned}
\mathrm{P}\Big\{\sum_{t=1}^{T}\sum_{k=1}^{K}\Big(\sum_{j=1}^{J}\Psi_{tkj}\varepsilon_{tj}U_{tr}\Big)^2 \geqslant \lambda^2(JT)^2/4\Big\} &= \mathrm{P}\Big\{\mathcal{X}_T^2 \geqslant \lambda^2JT^2/(4\sigma^2)\Big\} \\
&= \mathrm{P}\Big(\mathcal{X}_T^2 \geqslant T + A\sqrt{T}\log R\Big)
\end{aligned}$$

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

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

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

$$\begin{aligned}
&(JT)^{-1}\sum_{t=1}^{T}\|\Psi_t^\top(\widehat{\beta}-\beta^*)U_t\|^2 + \lambda\sum_{r=1}^{R}\|\widehat{\beta}_r-\beta_r\| \\
&\leqslant (JT)^{-1}\sum_{t=1}^{T}\|\Psi_t^\top(\beta-\beta^*)U_t\|^2 + 2\lambda\sum_{r=1}^{R}(\|\widehat{\beta}_r-\beta_r\| + \|\beta_r\| - \|\widehat{\beta}_r\|) \\
&\leqslant (JT)^{-1}\sum_{t=1}^{T}\|\Psi_t^\top(\beta-\beta^*)U_t\|^2 + 2\lambda\sum_{r\in\mathcal{R}(\beta)}(\|\widehat{\beta}_r-\beta_r\| + \|\beta_r\| - \|\widehat{\beta}_r\|) \\
&\quad + 2\lambda\sum_{r\in\mathcal{R}^c(\beta)}(\|\widehat{\beta}_r-\beta_r\| + \|\beta_r\| - \|\widehat{\beta}_r\|) \\
&\leqslant (JT)^{-1}\sum_{t=1}^{T}\|\Psi_t^\top(\beta-\beta^*)U_t\|^2 + 4\lambda\sum_{r\in\mathcal{R}(\beta)}\|\widehat{\beta}_r-\beta_r\|
\end{aligned} \tag{23}$$

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

$$(JT)^{-1}\max_{1\leqslant r\leqslant R}\|\sum_{t=1}^{T}\{\Psi_t(Y_t-\Psi_t^\top\widehat{\beta}U_t)U_t^\top\}_r\| \leqslant \lambda. \tag{24}$$

Then

$$\begin{aligned}
&(JT)^{-1}\|\sum_{t=1}^{T}\{\Psi_t\Psi_t^\top(\widehat{\beta}-\beta^*)U_tU_t^\top\}_r\| \\
\leqslant\ &(JT)^{-1}\|\sum_{t=1}^{T}\{\Psi_t(\Psi_t^\top\widehat{\beta}U_t-Y_t)U_t^\top\}_r\| + (JT)^{-1}\|\sum_{t=1}^{T}(\Psi_t\varepsilon_tU_t^\top)_r\|
\end{aligned} \tag{25}$$

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 $\mathcal{A}$. Finally, we prove (13). First, observe that,

$$\sum_{t=1}^{T}\Psi_t(Y_t-\Psi_t^\top\beta^*U_t)U_t^\top = \sum_{t=1}^{T}\Psi_t\Psi_t^\top(\widehat{\beta}-\beta^*)U_tU_t^\top + \sum_{t=1}^{T}\Psi_t\varepsilon_tU_t^\top.$$

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

$$(JT)^{-1}\|\sum_{t=1}^{T}\{\Psi_t\Psi_t^\top(\widehat{\beta}-\beta^*)U_tU_t^\top\}_r\| \geqslant \lambda/2, \text{ if } \quad \widehat{\beta}_r \neq 0.$$

The following arguments yields the bound (13) on the number of nonzero rows of $\widehat{\beta}_r^\top$:

$$\begin{aligned} M(\widehat{\beta}) &\leqslant \frac{4}{\lambda^2(JT)^2} \sum_{r\in\mathcal{R}(\widehat{\beta})} \| \sum_{t=1}^T \{\Psi_t\Psi_t^\top(\widehat{\beta}-\beta^*)U_tU_t^\top\}_r\|^2 \\ &\leqslant \frac{4}{\lambda^2(JT)^2} \sum_{r=1}^R \| \sum_{t=1}^T \{\Psi_t\Psi_t^\top(\widehat{\beta}-\beta^*)U_tU_t^\top\}_r\|^2 \\ &= \frac{4}{\lambda^2T^2} \| \sum_{t=1}^T \{J^{-1}\Psi_t\Psi_t^\top(\widehat{\beta}-\beta^*)U_tU_t^\top\}\|_{2,1}^2 \\ &\leqslant \frac{4}{\lambda^2T^2} \|\widehat{\beta}-\beta^*\|_{2,1}^2 \| \sum_{t=1}^T U_tU_t^\top\|_{2,1}^2 \\ &\leqslant \frac{4\phi_{max}^2}{\lambda^2T^2} \|\widehat{\beta}-\beta^*\|_{2,1}^2, \end{aligned}$$

which follows from Lemma 7.1, $\Psi_t\Psi_t^\top/J = I_K$ and ϕ_{max} is the maximum eigenvalues of the matrix $\sum_{t=1}^T U_tU_t^\top$. □

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 $\mathcal{R} = \mathcal{R}(\beta^*) = \{r : \beta_r^* \neq 0\}$

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

$$\begin{aligned} (JT)^{-1} \sum_{t=1}^T \|\Psi_t^\top(\widehat{\beta}-\beta^*)U_t\|^2 &\leqslant 4\lambda \sum_{r\in\mathcal{R}} \|\widehat{\beta}_r - \beta_r^*\| \\ &\leqslant 4\lambda\sqrt{s}\|(\widehat{\beta}-\beta^*)_{\mathcal{R}}\| \end{aligned} \tag{26}$$

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

$$\|(\widehat{\beta}-\beta^*)_{\mathcal{R}}\| \leqslant \sum_{t=1}^T \|\Psi_t^\top(\widehat{\beta}-\beta^*)U_t\|/(\kappa\sqrt{J}). \tag{27}$$

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

$$\sum_{r=1}^R \|\widehat{\beta}_r - \beta_r^*\| \leqslant 4\sum_{r\in\mathcal{R}} \|\widehat{\beta}_r - \beta_r^*\| \leqslant 4\sqrt{s}\|(\widehat{\beta}-\beta^*)_{\mathcal{R}}\|$$

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 $\mathrm{P}(\mathcal{A}^c)$ in Lemma 3.1. We consider now the event

$$\mathcal{A} = \Big\{ \max_{1\leqslant r\leqslant R} \Big\{ \sum_{t=1}^T \sum_{k=1}^K (\sum_{j=1}^J \Psi_{tkj}\varepsilon_{tj}U_{tr})^2 \Big\}^{1/2} \leqslant \lambda JT \Big\}.$$

The Markov inequality yields that

$$\mathrm{P}(\mathcal{A}^c) \leqslant \sum_{t=1}^T \mathsf{E} \Big\{ \max_{1\leqslant r\leqslant R} (\sum_{k=1}^K \sum_{j=1}^J \Psi_{tkj}\varepsilon_{tj}U_{tr})^2 \Big\} / (\lambda JT)^2.$$

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

$$W_{tj} = \Big(\sum_{k=1}^{K} \Psi_{tkj} \varepsilon_{tj} U_{t1}/J, \ldots, \sum_{k=1}^{K} \Psi_{tkj} \varepsilon_{tj} U_{tR}/J \Big) \in \mathbb{R}^R, \quad \forall j, \forall t.$$

We get that

$$\mathrm{P}(\mathcal{A}^c) \leqslant \frac{2e \log R - e}{\lambda^2 JT} \sigma^2 (JT)^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J} \Big(\max_{1 \leqslant r \leqslant R} | \sum_{k=1}^{K} \Psi_{tkj} U_{tr} | \Big)^2.$$

By the definition of λ in Theorem 3.2 and Assumption 3.2 we obtain

$$\mathrm{P}(\mathcal{A}^c) \leqslant \frac{(2e \log R - e) C}{(\log R)^{1+\delta}}. \qquad \square$$

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 $\mathrm{P}(\mathcal{A}^c)$ in Lemma 3.1. We consider now the event

$$\mathcal{A} = \Big\{ \max_{1 \leqslant r \leqslant R} \Big\{ \sum_{t=1}^{T} \sum_{k=1}^{K} (\sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr})^2 \Big\}^{1/2} \leqslant \lambda JT \Big\}.$$

Thus, following the fact that different space basis Ψ_k and $\Psi_{k'}$ are independent, we have:

$$\begin{aligned}
\mathrm{P}(\mathcal{A}^c) &= \mathrm{P}\Big[\max_{1 \leqslant r \leqslant R} \Big\{ \sum_{t=1}^{T} \sum_{k=1}^{K} (\sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr})^2 \Big\}^{1/2} > \lambda JT \Big] \\
&= \mathrm{P}\Big[\max_{1 \leqslant r \leqslant R} \Big\{ \sum_{t=1}^{T} (\sum_{k=1}^{K} \sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr})^2 \Big\}^{1/2} > \lambda JT \Big] \\
&\leqslant R\,\mathrm{P}\Big[\Big\{ \sum_{t=1}^{T} (\sum_{k=1}^{K} \sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr})^2 \Big\}^{1/2} > \lambda JT \Big] \\
&= R\,\mathrm{P}\Big[\Big\{ \sum_{t=1}^{T} (\sum_{k=1}^{K} \sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr})^2 \Big\}^{1/2} > \lambda JT \Big] \\
&= R\,\mathrm{P}\{ f(\underline{V}) > \lambda \}
\end{aligned}$$

where

$$\begin{aligned}
V_t &\stackrel{\mathrm{def}}{=} J^{-1} \sum_{k=1}^{K} \sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr} \\
\underline{V} &\stackrel{\mathrm{def}}{=} (V_{1r}, \ldots, V_{Tr}) \\
f(\underline{V}) &\stackrel{\mathrm{def}}{=} T^{-1} \left(\sum_{t=1}^{T} V_t^2 \right)^{1/2}.
\end{aligned}$$

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

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

$$\mathsf{E}\, f(\underline{V}) \leqslant \frac{C'}{\sqrt{T}}.$$

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

$$\begin{aligned} \mathrm{P}(\mathcal{A}^c) &\leqslant R\,\mathrm{P}\{f(\underline{V}) > \lambda\} \leqslant R\,\mathrm{P}\{f(\underline{V}) - \mathsf{E}\, f(\underline{V}) > \lambda - \frac{C'}{\sqrt{T}}\} \\ &\leqslant R\exp\left\{-\frac{(\lambda - \frac{C'}{\sqrt{T}})^2 T^2}{\mathcal{X}^*(\mathcal{T})\sum_t b_t^2}\right\} = R^{-\delta'} \end{aligned}$$

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

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

$$\frac{1}{T}\sum_{1\leq t\leq T}\left\|\widehat{Z}_{0,t}^\top\widehat{A} - Z_{0,t}^\top A^*\right\|^2 \leqslant \frac{1}{T}\sum_{1\leq t\leq T}\left\|\widehat{Z}_{0,t}^\top\widehat{A} - \widetilde{Z}_{0,t}^\top\widehat{A}\right\|^2 + \frac{1}{T}\sum_{1\leq t\leq T}\left\|\widetilde{Z}_{0,t}^\top\widehat{A} - Z_{0,t}^\top A^*\right\|^2,$$

where the second term is bounded by $\mathcal{O}_P(\rho^2 + \delta_K^2)$ by Theorem 2 of Park et al. (2009). For the first term, since

$$\begin{aligned} \widehat{Z}_{0,t} &= (\widehat{A}\Psi_t\Psi_t^\top\widehat{A}^\top)^{-1}\widehat{A}\Psi_t\widehat{\widehat{Y}}_t \\ \widetilde{Z}_{0,t} &= (\widehat{A}\Psi_t\Psi_t^\top\widehat{A}^\top)^{-1}\widehat{A}\Psi_t\widetilde{Y}_t \\ \widetilde{Z}_{0,t} - \widehat{Z}_{0,t} &= (\widehat{A}\Psi_t\Psi_t^\top\widehat{A}^\top)^{-1}\widehat{A}\Psi_t\{\Psi_t^\top(\widehat{\beta} - \beta^*)U_t\} \end{aligned}$$

and Theorem 3.3 tells us that $(JT)^{-1}\sum_{t=1}^T \| \Psi_t^\top(\widehat{\beta} - \beta^*)U_t \|^2$ 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 $T^{-1}\sum_{t=1}^T Z_{0,t}\widehat{Z}_{0,t}^\top$ is invertible. Suppose that the assertion is not true. We can choose a random vector e such that $\|e\| = 1$ and $e^\top\sum_{t=1}^T Z_{0,t}\widehat{Z}_{0,t}^\top = 0$. Note that

$$\begin{aligned} &\|T^{-1}\sum_{t=1}^T Z_{0,t}\widehat{Z}_{0,t}^\top\widehat{A} - T^{-1}\sum_{t=1}^T Z_{0,t}Z_{0,t}^\top A^*\| \\ \leqslant\ & T^{-1}\sum_{t=1}^T \|Z_{0,t}(\widehat{Z}_{0,t}^\top\widehat{A} - Z_{0,t}^\top A^*)\| \\ \leqslant\ & (T^{-1}\sum_{t=1}^T \|Z_{0,t}\|^2)^{1/2}(T^{-1}\sum_{t=1}^T \|\widehat{Z}_{0,t}^\top\widehat{A} - Z_{0,t}^\top A^*\|^2)^{1/2} \\ =\ & \mathcal{O}_P(\rho + \delta_K), \end{aligned} \tag{28}$$

because of Assumption (4.1.5) and Theorem 4.1. Thus with $f = T^{-1}\sum_{t=1}^T Z_{0,t}Z_{0,t}^\top e$, we obtain

$$\begin{aligned} \|f^\top m\| &= \|f^\top(A^*\Psi)\| + \mathcal{O}_P(\delta_K) \\ &= \|e^\top T^{-1}\sum_{t=1}^T Z_{0,t}Z_t^\top\widehat{A}\Psi\| + \mathcal{O}_P(\rho + \delta_K) \\ &= \mathcal{O}_P(\rho + \delta_K). \end{aligned}$$

This implies that $m_1, \ldots, m_L$ are linearly dependent, contradicting to the construction that all space basis are independent.

$\widetilde{Z}_{0,t} = B^\top \widehat{Z}_{0,t}$ and $\widetilde{A} = B^{-1}A$ give with (28)

$$\begin{aligned}
\|\widetilde{A} - A^*\| &= \|T^{-1}\sum_{t=1}^{T} Z_{0,t}Z_t^\top(\widetilde{A} - A^*)\|\mathcal{O}_P(1) \\
&= \|T^{-1}\sum_{t=1}^{T} Z_{0,t}\widetilde{Z}_{0,t}^\top\widetilde{A} - T^{-1}\sum_{t=1}^{T} Z_{0,t}Z_{0,t}^\top A^*\|\mathcal{O}_P(1) \\
&= \mathcal{O}_P(\rho + \delta_K) \qquad (29)
\end{aligned}$$

From Assumptions (4.1.4), (29) and Theorem 4.1, we get

$$\begin{aligned}
&T^{-1}\sum_{t=1}^{T} \|\widetilde{Z}_t^\top - Z_{0,t}\|^2 \\
&= T^{-1}\sum_{t=1}^{T} \|\widetilde{Z}_t^\top(m_1, \ldots, m_L)^\top - Z_{0,t}^\top(m_1, \ldots, m_L)^\top\|^2\mathcal{O}_P(1) \\
&= T^{-1}\sum_{t=1}^{T} \|\widetilde{Z}_t^\top A^* - \widetilde{Z}_t^\top\widetilde{A}\|^2\mathcal{O}_P(1) \\
&\qquad + T^{-1}\sum_{t=1}^{T} \|\widetilde{Z}_t^\top\widetilde{A} - Z_{0,t}^\top A^*\|^2\mathcal{O}_P(1) + \mathcal{O}_P(\delta_K^2) \qquad (30) \\
&\leqslant T^{-1}\sum_{t=1}^{T} \|\widetilde{Z}_{0,t} - Z_{0,t}\|^2\|\widetilde{A} - A^*\|^2\mathcal{O}_P(1) \\
&\qquad + T^{-1}\sum_{t=1}^{T} \|Z_{0,t}\|^2\|\widetilde{A} - A^*\|^2\mathcal{O}_P(1) \\
&\qquad + T^{-1}\sum_{t=1}^{T} \|\widetilde{Z}_t^\top\widetilde{A} - Z_{0,t}^\top A^*\|^2\mathcal{O}_P(1) + \mathcal{O}_P(\delta_K^2) \\
&= \mathcal{O}_P(\rho^2 + \delta_K^2).
\end{aligned}$$

We will show that for $h \neq 0$

$$T^{-1}\sum_{t=h+1}^{T} \{(\widetilde{Z}_{0,t+h} - Z_{0,t+h}) - (\widetilde{Z}_{0,t} - Z_{0,t})\}Z_{0,t}^\top = o_P(T^{-1/2}) \qquad (31)$$

This implies the first statement of Theorem 4.2, because by (30)

$$T^{-1}\sum_{t=-h+1}^{T} (\widetilde{Z}_{0,t} - Z_{0,t})(\widetilde{Z}_{0,t+h} - Z_{0,t+h}) = \mathcal{O}_P(b^2) = \mathcal{O}_P(T^{-1/2}).$$

For the proof of (31), define

$$\begin{aligned}
\widetilde{S}_{t,Z} &= J^{-1}\sum_{j=1}^{J}\widetilde{A}\Psi(X_{t,j})\Psi(X_{t,j})^{\top}\widetilde{A}^{\top} \\
S_{t,Z} &= A^{*}E\left\{\Psi(X_{t,j})\Psi(X_{t,j})^{\top}\right\}A^{*\top} \\
\widetilde{S}_{\alpha} &= (JT)^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J}\{\Psi(X_{t,j})\otimes\widetilde{Z}_{0,t}\}\{\Psi(X_{t,j})\otimes\widetilde{Z}_{0,t}\}^{\top} \\
S_{\alpha} &= T^{-1}\sum_{t=1}^{T}\mathsf{E}\left[\{\Psi(X_{t,j})\otimes Z_{0,t}\}\{\Psi(X_{t,j})\otimes Z_{0,t}\}^{\top}\middle|Z_{0,t}\right] \\
S &= J^{-1}A^{*}\left[\Psi(X_{t,j})\Psi(X_{t,j})^{\top}e-\mathsf{E}\left\{\Psi(X_{tj})\Psi(X_{tj})^{\top}e\right\}\right],
\end{aligned}$$

where $e\in\mathbb{R}^{K}$ with $\|e\|=1$. Let $\tilde{a}$ be the stack form of $\widetilde{A}$. It can be verified that

$$\widetilde{Z}_{0,t} = \widetilde{S}_{t,Z}^{-1}J^{-1}\sum_{j=1}^{J}\{Y_{t,j}A\Psi(X_{t,j})\}, \tag{32}$$

$$\tilde{a} = \widetilde{S}_{\alpha}^{-1}(JT)^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J}\{\Psi(X_{t,j})\otimes\widetilde{Z}_{0,t}\}Y_{t,j}. \tag{33}$$

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

$$\sup_{1\le t\le T}\|\widetilde{S}_{t,Z}-S_{t,Z}\|=o_{P}(\gamma),\quad \|\widetilde{S}_{\alpha}-S_{\alpha}\|=o_{P}(\gamma). \tag{34}$$

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

$$\widetilde{A}\Psi_{t}\Psi_{t}^{\top}\widetilde{A}^{\top}=(\widetilde{A}-A^{*}+A^{*})(\Psi_{t}\Psi_{t}^{\top}-\mathsf{E}\,\Psi_{t}\Psi_{t}^{\top}+\mathsf{E}\,\Psi_{t}\Psi_{t}^{\top})(\widetilde{A}-A^{*}+A^{*})^{\top},$$

to prove the first part it suffices to show that, uniformly for $1\leqslant t\leqslant T$,

$$J^{-1}\sum_{j=1}^{J}A^{*}\left[\Psi(X_{t,j})\Psi(X_{t,j})^{\top}-\mathsf{E}\left\{\Psi(X_{t,j})\Psi(X_{t,j})^{\top}\right\}\right](\widetilde{A}-A^{*})^{\top}=o_{P}(\gamma) \tag{35}$$

$$J^{-1}\sum_{j=1}^{J}(\widetilde{A}-A^{*})\left[\Psi(X_{t,j})\Psi(X_{t,j})^{\top}-\mathsf{E}\left\{\Psi(X_{t,j})\Psi(X_{t,j})^{\top}\right\}\right](\widetilde{A}-A^{*})^{\top}=o_{P}(\gamma) \tag{36}$$

$$J^{-1}\sum_{j=1}^{J}A^{*}\left[\Psi(X_{t,j})\Psi(X_{t,j})^{\top}-\mathsf{E}\left\{\Psi(X_{t,j})\Psi(X_{t,j})^{\top}\right\}\right]A^{*\top}=o_{P}(\gamma) \tag{37}$$

$$J^{-1}\sum_{j=1}^{J}A^{*}\,\mathsf{E}\left\{\Psi(X_{t,j})\Psi(X_{t,j})^{\top}\right\}(\widetilde{A}-A^{*})^{\top}=o_{P}(\gamma) \tag{38}$$

$$J^{-1}\sum_{j=1}^{J}(\widetilde{A}-A^{*})\,\mathsf{E}\left\{\Psi(X_{t,j})\Psi(X_{t,j})^{T}\right\}(\widetilde{A}-A^{*})^{\top}=o_{P}(\gamma) \tag{39}$$

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:

$$\mathrm{P}\left(|\sum_{j=1}^{J}W_{j}|>x\right)\leqslant 2\exp\left(-\frac{1}{2}\frac{x^{2}}{V+Mx/3}\right). \tag{40}$$

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 $S = J^{-1}A^* \left[\Psi(X_{t,j})\Psi(X_{t,j})^\top e - \mathsf{E}\left\{\Psi(X_{tj})\Psi(X_{tj})^\top e\right\}\right]$ 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 $1 \leqslant t \leqslant T$,

$$\begin{aligned} \widetilde{Z}_{0,t} - Z_{0,t} &= S_{t,Z}^{-1} J^{-1} \sum_{j=1}^J \varepsilon'_{t,j} A^* \Psi(X_{t,j}) \\ &\quad + S_{t,Z}^{-1} J^{-1} \sum_{j=1}^J \varepsilon'_{t,j} (\widetilde{A} - A^*) \Psi(X_{t,j}) + o_P(T^{-1/2}) \qquad (41) \\ &\stackrel{\text{def}}{=} \Delta_{t,1,Z} + \Delta_{t,2,Z} + o_P(T^{-1/2}). \end{aligned}$$

For the proof of the theorem it remains to show that for $1 \leqslant j \leqslant 2$

$$T^{-1} \sum_{t=-h+1}^T (\Delta_{t+h,j,Z} - \Delta_{t,j,Z}) Z_{0,t}^\top = o_P(T^{-1/2}). \qquad (42)$$

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

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

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

References

Benth, F. and Benth, J. (2005). Stochastic modelling of temperature variations with a view towards weather derivatives. Applied Mathematical Finance, 12(1):53–85.

Bickel, J. and Levina, E. (2008a). Covariance regularization by thresholding. Ann. Statist, 36(6):2577–2604.

Bickel, J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. Ann. Statist, 36(1):199–227.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. Annals of Statists, 37(4):1705–1732.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. Journal of Political Economy, 81:637–654.

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. Journal of Econometrics, 130:337–364.

Dümbgen, L., van de Geer, S., Veraar, M., and Wellner, J. A. (2008). Nemirovski's Inequalities Revisited. ArXiv e-prints.

Fengler, M. R., Härdle, W., and Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics. Journal of Financial Econometrics, 5(2):189–218.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. Journal of the American Statistical Association, 100:830–840.

Fu, W. J. (1998). Penalized regressions: The bridge versus the Lasso. Journal of Computational and Graphical Statistics, 7(3):397–416.

Gasser, T., Möcks, R., and Verleger, R. (1983). Selavco: A method to deal with trial-to-trial variability of evoked potential. Electroencephalography and Clinical Neurophysiology, 55:717–723.

Giannone, D., Reichlin, L., and Sala, L. (2005). Monetary policy in real time. In NBER Macroeconomics Annual 2004, Volume 19, NBER Chapters, pages 161–224. National Bureau of Economic Research, Inc.

Gleick et al., P. H. (2010). Climate change and the integrity of science. Science, 328:689–691.

Hall, A. and Hautsch, N. (2006). Order aggressiveness and order book dynamics. Empirical Economics, 30(4):973–1005.

Hall, P., Mller, H. G., and Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. The Annals of Statistics, 34(3):1493–1517.

Hautsch, N. and Ou, Y. (2008). Yield curve factors, term structure volatility, and bond risk premia. SFB 649 Discussion Papers SFB649DP2008-053, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany.

Hedin, A. E. (1991). Extension of the msis thermosphere model into the middle and lower atmosphere. Journal of Geophysical Research, 96:1159–1172.

Janson, S. (2004). Large deviations for sums of partly dependent random variables. Random Structures Algorithms, 24(3):234–248.

Jia, J., Rohe, K., and Yu, B. (2009). The lasso under heteroscadecity. Technical Report 783, Statistics Department, UC Berkeley.

Karl, T. R., Kukla, G., Razuvayev, V. N., Changery, M. J., Quayle, R. G., Heim, R. R., and Easterling, D. R. (1991). Global warming: Evidence for asymmetric diurnal temperature change. Geophysical Research Letters, 18:2253–2256.

Kauermann, G. (2000). Modeling longitudinal data with ordinal response by varying coefficients. Biometrics, 56(3):1692–698.

Lee, R. D. and Carter, L. (1992). Modeling and forecasting the time series of u.s. mortality. Journal of the American Statistical Association, 87(419):659–671.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. Electronic Journal of Statistics, 2:90–102.

Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. (2009). Taking advantage of sparsity in multi-task learning. Proceedings of Conference on Learning Theory (COLT) 2009.

Martinussen, T. and Scheike, T. (2000). A nonparametric dynamic additive regression model for longitudinal data. Annals of Statistics, 28(4):1000–1025.

Mohr, P. N. C., Biele, G., Krugel, L. K., Li, S.-C., and Heekeren, H. R. (2010). Neural foundations of risk-return trade-off in investment decisions. NeuroImage, 49(3):2556–2563.

Myšičková, A., Song, S., Mohr, P. N., Heekeren, H. R., and Härdle, W. K. (2010). Risk patterns and correlated brain activities. Submitted to Neuroimage.

Nelson, C. R. and Siegel, A. F. (1987). Parsimonious modeling of yield curves. Journal of Business, 60:473–489.

Odening, M., Berg, E., and Turvey, C. (2008). Management of climate risk in agriculture. Special Issue of the Agricultural Finance Review, 68(1):83C97.

Park, B. U., Mammen, E., Härdle, W., and Borak, S. (2009). Time series modelling with semiparametric factor dynamics. Journal of the American Statistical Association, 104(485):284–298.

Parton, W. J. and Logan, J. A. (1981). A model for diurnal variation in soil and air temperature. Agricultural Meteorology, 23:205 – 216.

Racsko, P., Szeidl, L., and Semenov, M. (1991). A serial approach to local stochastic weather models. Ecological Modelling, 57(1-2):27 – 41.

Ramsay, J. and Silverman, B. (2005). Functional Data Analysis, 2nd Edition. Springer.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association, 97:1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics, 20(2):147–62.

Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for var analysis. NBER Working Papers 11467, National Bureau of Economic Research, Inc. available at http://ideas.repec.org/p/nbr/nberwo/11467.html.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, 58(1):267–288.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002). A general statistical analysis for fmri data. NeuroImange, 15:1–15.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B, 68(1):49–67.

SFB 649 Discussion Paper Series 2010

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Volatility Investing with Variance Swaps" by Wolfgang Karl Härdle and Elena Silyakova, January 2010.
002 "Partial Linear Quantile Regression and Bootstrap Confidence Bands" by Wolfgang Karl Härdle, Ya'acov Ritov and Song Song, January 2010.
003 "Uniform confidence bands for pricing kernels" by Wolfgang Karl Härdle, Yarema Okhrin and Weining Wang, January 2010.
004 "Bayesian Inference in a Stochastic Volatility Nelson-Siegel Model" by Nikolaus Hautsch and Fuyu Yang, January 2010.
005 "The Impact of Macroeconomic News on Quote Adjustments, Noise, and Informational Volatility" by Nikolaus Hautsch, Dieter Hess and David Veredas, January 2010.
006 "Bayesian Estimation and Model Selection in the Generalised Stochastic Unit Root Model" by Fuyu Yang and Roberto Leon-Gonzalez, January 2010.
007 "Two-sided Certification: The market for Rating Agencies" by Erik R. Fasten and Dirk Hofmann, January 2010.
008 "Characterising Equilibrium Selection in Global Games with Strategic Complementarities" by Christian Basteck, Tijmen R. Daniels and Frank Heinemann, January 2010.
009 "Predicting extreme VaR: Nonparametric quantile regression with refinements from extreme value theory" by Julia Schaumburg, February 2010.
010 "On Securitization, Market Completion and Equilibrium Risk Transfer" by Ulrich Horst, Traian A. Pirvu and Gonçalo Dos Reis, February 2010.
011 "Illiquidity and Derivative Valuation" by Ulrich Horst and Felix Naujokat, February 2010.
012 "Dynamic Systems of Social Interactions" by Ulrich Horst, February 2010.
013 "The dynamics of hourly electricity prices" by Wolfgang Karl Härdle and Stefan Trück, February 2010.
014 "Crisis? What Crisis? Currency vs. Banking in the Financial Crisis of 1931" by Albrecht Ritschl and Samad Sarferaz, February 2010.
015 "Estimation of the characteristics of a Lévy process observed at arbitrary frequency" by Johanna Kappusl and Markus Reiß, February 2010.
016 "Honey, I'll Be Working Late Tonight. The Effect of Individual Work Routines on Leisure Time Synchronization of Couples" by Juliane Scheffel, February 2010.
017 "The Impact of ICT Investments on the Relative Demand for High-Medium-, and Low-Skilled Workers: Industry versus Country Analysis" by Dorothee Schneider, February 2010.
018 "Time varying Hierarchical Archimedean Copulae" by Wolfgang Karl Härdle, Ostap Okhrin and Yarema Okhrin, February 2010.
019 "Monetary Transmission Right from the Start: The (Dis)Connection Between the Money Market and the ECB's Main Refinancing Rates" by Puriya Abbassi and Dieter Nautz, March 2010.
020 "Aggregate Hazard Function in Price-Setting: A Bayesian Analysis Using Macro Data" by Fang Yao, March 2010.
021 "Nonparametric Estimation of Risk-Neutral Densities" by Maria Grith, Wolfgang Karl Härdle and Melanie Schienle, March 2010.

SFB 649 Discussion Paper Series 2010

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

022 "Fitting high-dimensional Copulae to Data" by Ostap Okhrin, April 2010.
023 "The (In)stability of Money Demand in the Euro Area: Lessons from a Cross-Country Analysis" by Dieter Nautz and Ulrike Rondorf, April 2010.
024 "The optimal industry structure in a vertically related market" by Raffaele Fiocco, April 2010.
025 "Herding of Institutional Traders" by Stephanie Kremer, April 2010.
026 "Non-Gaussian Component Analysis: New Ideas, New Proofs, New Applications" by Vladimir Panov, May 2010.
027 "Liquidity and Capital Requirements and the Probability of Bank Failure" by Philipp Johann König, May 2010.
028 "Social Relationships and Trust" by Christine Binzel and Dietmar Fehr, May 2010.
029 "Adaptive Interest Rate Modelling" by Mengmeng Guo and Wolfgang Karl Härdle, May 2010.
030 "Can the New Keynesian Phillips Curve Explain Inflation Gap Persistence?" by Fang Yao, June 2010.
031 "Modeling Asset Prices" by James E. Gentle and Wolfgang Karl Härdle, June 2010.
032 "Learning Machines Supporting Bankruptcy Prediction" by Wolfgang Karl Härdle, Rouslan Moro and Linda Hoffmann, June 2010.
033 "Sensitivity of risk measures with respect to the normal approximation of total claim distributions" by Volker Krätschmer and Henryk Zähle, June 2010.
034 "Sociodemographic, Economic, and Psychological Drivers of the Demand for Life Insurance: Evidence from the German Retirement Income Act" by Carolin Hecht and Katja Hanewald, July 2010.
035 "Efficiency and Equilibria in Games of Optimal Derivative Design" by Ulrich Horst and Santiago Moreno-Bromberg, July 2010.
036 "Why Do Financial Market Experts Misperceive Future Monetary Policy Decisions?" by Sandra Schmidt and Dieter Nautz, July 2010.
037 "Dynamical systems forced by shot noise as a new paradigm in the interest rate modeling" by Alexander L. Baranovski, July 2010.
038 "Pre-Averaging Based Estimation of Quadratic Variation in the Presence of Noise and Jumps: Theory, Implementation, and Empirical Evidence" by Nikolaus Hautsch and Mark Podolskij, July 2010.
039 "High Dimensional Nonstationary Time Series Modelling with Generalized Dynamic Semiparametric Factor Model" by Song Song, Wolfgang K. Härdle, and Ya'acov Ritov, July 2010.