aptara	ECTJ	ectj12024	**Dispatch:** March 13, 2014	**CE:** N / A
	Journal	**MSP No.**	**No. of pages:** 32	**PE:** David

Econometrics Journal (2014), volume **17**, pp. 1–32.
doi: 10.1111/ectj.12024

Generalized dynamic semi-parametric factor models for high-dimensional non-stationary time series

SONG SONG[†], WOLFGANG K. HÄRDLE[‡,§] AND YA'ACOV RITOV[‡,§]

[†]*Department of Mathematics, University of Alabama, 318B Gordon Palmer Hall, Tuscaloosa, AL 35487, USA*.
E-mail: ssoonngg123@gmail.com

[‡]*School of Business and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099, Berlin, Germany*.
E-mail: haerdle@wiwi.hu-berlin.de, yaacov@mscc.huji.ac.il

[§]*Department of Statistics, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel*.

First version received: xxxx 2013; final version accepted: xxxx 2013

Summary High-dimensional non-stationary time series, which reveal both complex trends and stochastic behaviour, occur in many scientific fields, e.g. macroeconomics, finance, neuroeconomics, etc. To model these, we propose a generalized dynamic semi-parametric factor model with a two-step estimation procedure. After choosing smoothed functional principal components as space functions (factor loadings), we extract various temporal trends by employing variable selection techniques for the time basis (common factors). Then, we establish this estimator's non-asymptotic statistical properties under the dependent scenario (β-mixing and m-dependent) with the weakly cross-correlated error term. At the second step, we obtain a detrended low-dimensional stochastic process that exhibits the dynamics of the original high-dimensional (stochastic) objects and we further justify statistical inference based on this. We present an analysis of temperature dynamics in China, which is crucial for pricing weather derivatives, in order to illustrate the performance of our method. We also present a simulation study designed to mimic it.

Keywords: *Asymptotic inference, Factor model, Group Lasso, Periodic, Seasonality, Semi-parametric model, Spectral analysis, Weather*.

1. INTRODUCTION

Over the past few decades, high-dimensional data analysis has attracted increasing attention in various fields. We often face a high-dimensional vector of observations evolving in time (a very large interrelated time process), which is also possibly controlled by an exogenous covariate. For example, in macroeconomic forecasting, people use very large dimensional economic and financial time series (Stock and Watson, 2005b). In meteorology and agricultural economics, one of the primary interests is to study the fluctuations of temperatures at different nearby locations; for a recent summary, see Gleick et al. (2010). Such an analysis is also essential for pricing weather derivatives and hedging weather risks in finance (Odening et al., 2008). In

neuroeconomics, high-dimensional functional magnetic resonance imaging (fMRI) data are used to analyse the brain's response to certain risk-related stimuli, as well as to identify its activation area (Worsley et al., 2002). In financial engineering, the dynamics of the implied volatility surface (IVS) are studied for risk management, calibration and pricing purposes (Fengler et al., 2007). Other examples include 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 (Bowsher and Meeks, 2006), and so on.

Empirical studies in economics and finance often involve non-stationary variables, such as real consumer price index, individual consumption, exchange rates, real gross domestic product, etc. For example, the large panel macroeconomic data, provided by Stock and Watson (2005a), contain some complex non-stationary behaviour, such as normal seasonality, large economic cycle and upward trend representing economic growth, etc. However, some studies have produced counterintuitive and contradictory results; see Campbell and Yogo (2006), Cai et al. (2009), Xiao (2009) and Wang and Phillips (2009a, b). This might partly be attributed to the use of methods that cannot capture non-stationarity or non-linear structural relations. In fact, in the econometrics literature, the study of such non-stationary time series is dominated by linear or, at most, parametric models, restricting non-stationarity to the unit root or long-memory autoregressive fractionally integrated moving average (ARFIMA) types of non-stationarity and restricting structural relations to linear or parametric types of cointegration models. General processes can be characterized by certain recurrence properties. These processes contain stationary, long-memory and unit-root type or nearly integrated processes as subclasses, and are more general than the class of locally stationary processes. As pointed out in the recent econometrics literature, when some covariates are non-stationary, conventional statistical tests are invalid, even though the predictive power in a non-parametric regression model can be improved if some covariates are non-stationary. While some asymptotic results for general non-parametric estimation methods for low-dimensional non-stationary time series have been obtained, semi-parametric modelling has hardly been investigated so far, especially for high-dimensional non-stationary time series. For the i.i.d. case, there have been many studies in the literature, including but not limited to Horowitz and Lee (2005), Horowitz et al. (2006) and Horowitz (2006) for the moderate-dimension case and Horowitz and Huang (2012) and Huang et al. (2010) for the high-dimension case.

In such situations, if we still use either high-dimensional static methods, which are initially designed for independent data or low-dimensional multivariate time series techniques (on a few concentrated series through naïve aggregation), we might lose potentially relevant information, such as the time dynamics or the space dependence structure. This might produce suboptimal forecasts and would be extremely inefficient. In macroeconomics studies, this potentially creates an omitted variable bias with adverse consequences for both structural analysis and forecasting. Christiano et al. (1999) has pointed out that the positive reaction of prices in response to a monetary tightening, the so-called price puzzle, is an artefact resulting from the omission of forward-looking variables, such as the commodity price index. The more scattered and dynamic the information is, the more severe this loss will be. To this end, an integrated solution addressing both issues is appealing. We need to analyse jointly time and space dynamics by simultaneously fitting a time series evolution and by fine tuning the factors involved. The solution we are seeking helps us to understand the spatial pattern, to gain strength from the different time points and, at the same time, to analyse the non-stationary temporal behaviour of the value at each spatial point. In this paper, we present and investigate the so-called generalized dynamic semi-parametric

factor model (GDSFM), together with its corresponding panel version, in order to address this problem.

Panel data have attracted much attention in econometrics; see, e.g. Baltagi (2005), Frees (2004) and Hsiao (1986). To address the above challenges in a large panel of economic and financial time series, some recent studies have proposed ways to impose restrictions on the covariance structure in order to limit the number of parameters to be estimated. Dynamic factor models introduced by Forni et al. (2000) and Stock and Watson (2002a, b), also discussed by Forni et al. (2005) and Giannone et al. (2005), have drawn upon the idea that the intertemporal dynamics can be explained and represented by a few common factors (low-dimensional time series). Another approach in this field has been presented by Park et al. (2009), where a latent L-dimensional process, $Z_1, \ldots, Z_T$, is introduced, and the J-dimensional random process $Y_t = (Y_{t,1}, \ldots, Y_{t,J})^\top$, $t = 1, \ldots, T$, is represented as

$$Y_{t,j} = Z_{t,1} m_{1,j} + \cdots + Z_{t,L} m_{L,j} + \varepsilon_{t,j}, \quad j = 1, \ldots, J, \quad t = 1, \ldots, T. \tag{1.1}$$

Here, $Z_{t,l}$ are the common factors depending on time, $\varepsilon_{t,j}$ are errors or specific factors, and the coefficients $m_{l,j}$ are factor loadings. The index $t = 1, \ldots, T$ reflects the time evolution, $\{Z_t\}_{t=1}^T$ ($Z_t = (Z_{t,1}, \ldots, Z_{t,L})^\top$) is assumed to be a stationary random process, and $m_l = (m_{l,1}, \ldots, m_{l,J})^\top$ captures the spatial dependency structure. The study of the time behaviour of the high-dimensional Y_t is then simplified to the modelling of Z_t, which is more feasible when $L \ll J$. Model (1.1) reduces to a special case of the generalized dynamic factor model (approximate factor model) considered by Forni et al. (2000, 2005) and Hallin and Liska (2007), when $Z_{t,l} = a_{l,1}(B)U_{t,1} + \cdots + a_{l,q}(B)U_{t,q}$. Here, the q-dimensional vector process $U_t = (U_{t,1}, \ldots, U_{t,q})^\top$ is an orthonormal white noise and B denotes the lag operator. In this case, model (1.1) is expressed as $Y_{t,j} = m_{0,j} + \sum_{k=1}^q b_{k,j}(B)U_{t,k} + \varepsilon_{t,j}$, where $b_{k,j}(B) = \sum_{l=1}^L a_{l,k}(B)m_{l,j}$. Less general models in the literature include the static factor models proposed by Stock and Watson (2002a, b) and the exact factor models suggested by Sargent and Sims (1977) and Geweke (1977).

Our goal of modelling high-dimensional non-stationary time series is achieved by using a sparse representation approach to regression. In fact, we combine spatio-temporal modelling with group Lasso (Yuan and Lin, 2006). We approximate both the temporal common factors and spatial factor loadings by a linear combination of series terms. Because the temporal non-stationarity behaviour might result from different sources, the choice of basis functions is important. We start by introducing an overparametrized model, which can capture (almost) any type of temporal behaviour, such as cyclic behaviour plus linear or quadratic trends, by utilizing series basis, such as powers, trigonometrics, local polynomials, periodic functions and B-splines. Then, we select a sparse submodel, using penalizing-Lasso and group-Lasso techniques.

In practice, there might be multiple subjects, each of which by itself corresponds to a set of high-dimensional time series. For example, in international economies, industrial organizations or financial studies, there are data for many countries, firms or assets, all of which are high-dimensional. Thus, we also need to provide a panel version of the high-dimensional time series model to address this issue. Compared with previous studies in the literature, the novelty of this paper lies in the following aspects.

1. When the time process is not stationary (i.e. the process has a non-linear, non-parametric temporal structure in time), using a skilful selection of time basis, we can handle such complex time series. To achieve a successful selection, the key assumption is that the initially proposed time basis should not be too dependent, even though the number can

be large (i.e. we should include as many orthogonal time basis functions as possible for the automatic selection). From the point of view of large panel time series modelling, we incorporate non-stationarity and non-linearity (complex trends) into time dynamics. We deviate from most of the current body of literature that still requires Z_t to be stationary and still needs a large number of observations (relative to dimensionality) to establish asymptotic properties.

2. The contribution lies in the way the time dynamics is introduced for variable selection and regularization methods. Under the assumption that the product of the time basis, space basis and error term has a bounded second moment, and the error term ε_t is only weakly cross-correlated, the non-asymptotic theoretical properties of existing methods are established under the scenario of independence. We extend it to a dependent scenario ($\boldsymbol{\beta}$-mixing and m-dependent process) with the weakly cross-correlated error term (the details are specified in Assumption 3.2), and we derive oracle sparsity inequalities (non-asymptotic risk bounds). The key assumption is that the temporal dependence level of the error term is controlled within some level. Also, this result is not built upon any specific forms of time and space basis.
3. When the space structure of m_l is complex, the low-dimensional parametrizations do not capture it properly. We employ a data-driven semi-parametric method, introduced by Hall et al. (2006), to capture the spatial dependence structure.
4. For the case that there might be multiple subjects, each of which corresponds to a set of high-dimensional time series, we provide a panel version of the model with a corresponding estimation method.

In a variety of applications, we have explanatory variables $X_{t,j} \in \mathbb{R}^d$ at hand, e.g. the geo coordinates of weather stations, the voxels (volume elements, representing values on regular grids) of fMRI, or the moneyness and time-to-maturity variables for implied volatility modelling, which can influence the factor loadings m_l. An important refinement of model (1.1) is to incorporate the existence of observable covariates $X_{t,j}$ from Park et al. (2009). The factor loadings are then generalized to functions of $X_{t,j}$. In the following, we write $X_t = (X_{t,1}, \ldots, X_{t,J})^\top$ and consider the generalization of (1.1),

$$Y_{t,j} = Z_t^\top m(X_{t,j}) + \varepsilon_{t,j}, \quad t = 1, \ldots, T, \tag{1.2}$$

where $Y_{tj}, \varepsilon_{tj} \in \mathbb{R}$, $X_{tj} \in \mathbb{R}^d$, $m : \mathbb{R}^d \to \mathbb{R}^L$ and $Z_t \in \mathbb{R}^{1\times L}$.

Our motivating example is from temperature analysis for pricing weather derivatives. The data set is taken from the Climatic Data Center (CDC) of the China Meteorological Administration (CMA). It contains daily observations from 159 weather stations across China from 1 January 1957 to 31 December 2009. We would not only like to address the question of whether there is a change in time, but also to permit a different trend in time, in different climate types, as shown by Figure 1 (left), which shows a map of the network of China's weather stations. Besides the well-known seasonality effect, we can expect a trend related to climate change. In Figure 1 (right), we show the moving average (of 730 nearby days) of temperatures in China from 1 January 1957 to 31 December 2009, which is $(159 \times 730)^{-1} \sum_{s=-354}^{+365} \sum_{j=1}^{159} Y_{t+s,j}$, where $Y_{t,j}$ is the temperature of the jth weather station at time t. From this figure, we can see that there is a large period (around 10 years) between peaks and an upward trend for China's temperatures. Besides these trends, there is also stochasticity inherent in the remaining time dynamics, which is essential for pricing weather derivatives and hedging weather risks. By simultaneously studying the dynamics of temperatures in various places w.r.t. $X_{t,j} = X_j$

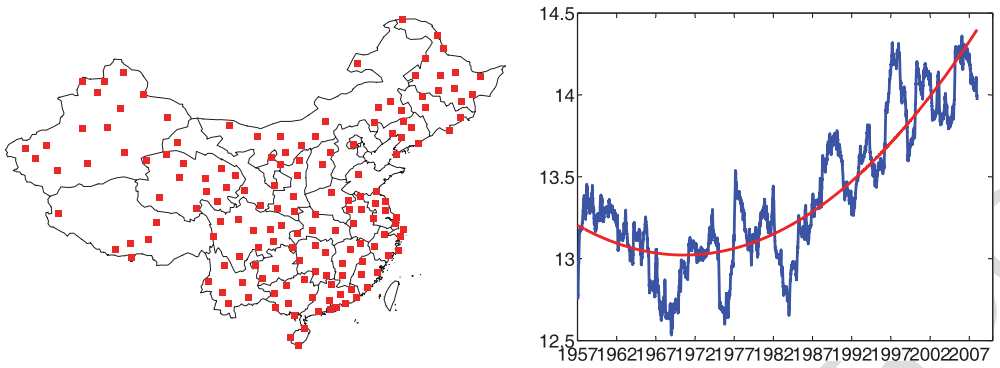


Figure 1. Map of China's weather stations and moving averages of temperature.

Q6

(the three-dimensional geographical information of the jth weather station), we will be able to estimate, forecast and price temperatures in time and space.

The rest of the paper is organized as follows. In the next section, we present details of the GDSFM, together with the corresponding basis selection and panel model. We present the estimator's properties under various scenarios in Section 3. In Section 4, we apply the method to the motivating problem: the dynamic behaviour of temperatures. In Section 5, we present the results of simulation studies that mimic the previous empirical example. Section 6 contains concluding remarks. The estimation procedure and all technical proofs are sketched in Appendices A and B, respectively.

2. GENERALIZED DYNAMIC SEMI-PARAMETRIC FACTOR MODELS

We observe $(X_{t,j}, Y_{t,j})$ for $j = 1, \ldots, J$ and $t = 1, \ldots, T$, $Y_{tj} \in \mathbb{R}$, $X_{tj} \in \mathbb{R}^d$, $\varepsilon_{tj} \in \mathbb{R}$ generated by

$$Y_{tj}^\top = Z_t^\top A^* \Psi(X_{tj}) + \varepsilon'_{tj} = \left(U_t^\top \Gamma^* + Z_{0,t}^\top\right) A^* \Psi(X_{tj}) + \varepsilon'_{tj},$$

where A^* and Γ^* are the $L \times K$ and $R \times L$ (unknown) underlying coefficient matrices and Z_t has two components $\Gamma^{*\top} U_t$ and $Z_{0,t}$. Let $Y_t = (Y_{t,1}, \ldots, Y_{t,J})^\top$, $X_t = (X_{t,1}, \ldots, X_{t,J})^\top$, $\varepsilon'_t = (\varepsilon'_{t,1}, \ldots, \varepsilon'_{t,J})^\top$ and $\Psi(X_t) = (\Psi(X_{t1}), \ldots, \Psi(X_{tJ}))$ (abbreviated as Ψ_t). We rewrite this in compact form as

$$\begin{aligned} Y_t^\top &= \left(U_t^\top \Gamma^* + Z_{0,t}^\top\right) A^* \Psi(X_t) + \varepsilon_t'^\top \\ &= U_t^\top \Gamma^* A^* \Psi(X_t) + Z_{0,t}^\top A^* \Psi_t + \varepsilon_t'^\top. \end{aligned} \tag{2.1}$$

Again, by introducing $\beta^{*\top} = \Gamma^* A^*$ (the $R \times K$ unknown underlying coefficient matrices consisting of β_{rk}) and $\varepsilon_t = Z_{0,t}^\top A^* \Psi_t + \varepsilon'_t$, we could further simplify this as

$$Y_t^\top \overset{\text{def}}{=} U_t^\top \beta^{*\top} \Psi_t + \varepsilon_t^\top. \tag{2.2}$$

Note the following.

(1) Time evolution/common factors. $Z_t = (Z_{t,1}, \ldots, Z_{t,L})^\top$ is an unobservable L-dimensional process consisting of both a deterministic portion, $\Gamma^{*\top} U_t$, and a stochastic portion, $Z_{0,t}$. Here, $\{Z_{0,t}\}_{t=1}^T$ is a stationary process (to be detailed later). A key difference between our method and that of Park et al. (2009) is this additional non-stationary component $\Gamma^{*\top} U_t$.

(2) Factor loading functions and error terms. $m(X_{tj}) = A^* \Psi(X_{tj})$ 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 $\varepsilon_t' = (\varepsilon_{t,1}', \ldots, \varepsilon_{t,J}')^\top$ are the errors. Throughout the paper, we assume that the covariates $X_{t,j}$ have support $[0, 1]^d$. The error terms ε_t and ε_t' only need to satisfy some mild condition (details specified in Assumptions 3.2 and 3.3(c)), which allows them to be weakly dependent (over time) and cross-correlated (over space).

(3) Time and space basis. We use a series expansion to capture the time trend and the space-dependent structure. Let $U_t^\top = (u_1(t), \ldots, u_R(t))$ be the $1 \times R$ vector of time basis functions (polynomial and trigonometric functions, etc.), which are selected and weighted by the matrix Γ^*. For the space basis, we take $\Psi_t = (\psi_1(X_t), \ldots, \psi_K(X_t))^\top$ ($K \times J$ matrix). For every β matrix, we introduce $\beta_r = (\beta_{kr}, 1 \le k \le K)$, which is the column vector formed by the coefficients corresponding to the rth time basis. Additionally, we define the mixed $(2, 1)$ norm $\|\beta\|_{2,1} = \sum_{r=1}^R \sqrt{\sum_{k=1}^K \beta_{rk}^2}$. 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 the sake of simplicity and convenience, we 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.

Because the non-stationary behaviour might be very complex, to ensure that all the trends causing the non-stationarity are considered, the dimension R of the initially included time basis might be large. For example, in the temperature analysis, because we never know the exact frequency (frequencies) of the period(s), at the beginning, we include all the basis functions. We think that this might be useful for capturing the non-stationary behaviour, e.g. 16 trigonometric functions w.r.t. different frequencies and 53×3 (year by year) cubic polynomial basis. Consequently, we end up with $R = 175$. However, to avoid overfitting, variable selection with regularization techniques is necessary. A popular variable selection method is the 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. Here, we assume that the vectors β_r are not only sparse, but also have the same sparsity pattern across different factors. We study the estimator's theoretical sparsity properties related to the time basis selection, and we take (2.1) to be the true model. Because group LASSO permits overparametrization, this is a mild assumption. We would also like to emphasize that our non-asymptotic sparse oracle inequality results are independent of specifications of time and space basis. They apply equally to local polynomials, periodic functions, such as sin and cos, and B-splines, etc., while we just assume that there is no additional approximation error for obtaining the space basis at this non-asymptotic analysis step.

2.1. A panel version with multiple individuals

Here, we just present a panel version of (2.1) based on assumptions closely related to the fMRI neuroeconomics study (Mohr et al., 2010). It is reasonable to assume that different subjects have

different patterns of brain activation (to the external stimuli) represented by the time series Z_t, but they (and all human beings) share essentially the same spatial structure of the brain represented by the space function $A^*\Psi_t$. With a panel of I subjects, we formulate the following generalization of (2.1) and (2.2),

$$Y^i_{t,j} = \sum_{l=1}^{L} (Z^i_{0,t,l} + U_t^\top \Gamma^i_l)\, m_l(X_{t,j}) + \varepsilon^i_{t,j}, \quad 1 \le j \le J_t,\ 1 \le t \le T,\ 1 \le i \le I, \tag{2.3}$$

where the fixed effects $Z^i_{0,t,l}$ and Γ^i_l are the individual effects on functions m_l for subject i at time point t. For identification purposes, assume

$$E\left[\sum_{i=1}^{I}\sum_{l=1}^{L} Z^i_{0,t,l}\, m_l(X_{t,j}) | X_{t,j}\right] = 0.$$

For this data structure, we use $\overline{Y}_{t,j}$ to denote the average of $Y^i_{t,j}$ across different subjects i. Thus, from (2.3), we have

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

The two-step estimation procedure for the panel version model is as follows.

STEP 1. Take the average of $Y^i_{t,j}$ across different subjects i, and estimate the common basis function in space $\hat{m}_l$ as in the original approach; see Appendix A for more details.

STEP 2. Given the common $\hat{m}_l$, estimate subject-specific time factors $Z^i_{t,l}$:

$$Y^i_{t,j} = \sum_{l=1}^{L} \left(Z^i_{0,t,l} + U_t^\top \Gamma^i_l\right) \widehat{m}_l(X_{t,j}) + \varepsilon^i_{t,j}.$$

Next, we discuss the choice of time basis U_t, space basis Ψ_t and the estimation procedure for (2.2).

2.2. *Choice of time basis*

To capture the global trend in time, we can use any orthogonal polynomial basis, e.g. $u_1(t) = 1/C_1$, $u_2(t) = t/C_2$, $u_3(t) = (3t^2 - 1)/C_3, \ldots$ (where C_i are generic constants with $T^{-1}\sum_{t=1}^{T} u_r^2(t)/C_r^2 = 1$). We can also use the fact that there are natural frequencies in the data, and thus start with a few trigonometric functions. In the temperature example, the yearly cycle and a large period are two clear phenomena. To capture these periodic variations, we can use trigonometric functions, $u_4(t) = \sin(2\pi t/p)/C_4$, $u_5(t) = \cos(2\pi t/p)/C_5$, $u_6(t) = \sin(2\pi t/(p/2))/C_6, u_7(t) = \cos(2\pi t/(p/2))/C_7, \ldots$, with the given period p: 365 and 10 for the yearly cycle and large period, respectively. In the fMRI application of Myšičková et al. (2013), the basic experiment is repeated every 29.5 seconds, and we have the period $p = 11.8$ (there is a fMRI scan every 2.5 seconds). In general, to adopt various types of non-linearities, various basis functions could be employed, such as powers, trigonometrics, local polynomials, periodic functions, B-splines, etc. The theory to be presented later for selecting the significant time basis selection is actually independent of their specific forms, and thus is very useful in practice.

2.3. Choice of space basis

There are various choices for space basis. Park et al. (2009) have proposed a multidimensional B-spline basis. Alternatively, functional principal component analysis (PCA; Hall et al., 2006) can be employed, which combines smoothing techniques with ideas related to functional PCA. The basic steps are as follows.

STEP 1. Calculate the covariance operator (in a functional sense). Denote $X_{tj} = (X_{tj}^1, \ldots, X_{tj}^d)$, $u = (u^1, \ldots, u^d)$ and $v = (v^1, \ldots, v^d)$ (as for $b, \widehat{b}, b_1, \widehat{b}_1, b_2$ and $\widehat{b}_2$). Given $u \in [0, 1]^d$, and bandwidths h_μ and h_ϕ, define $(\widehat{a}, \widehat{b})$ to minimize

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

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

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

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

$$\widehat{\psi}(u, v) = \widehat{\phi}(u, v) - \widehat{\mu}(u)\widehat{\mu}(v). \tag{2.4}$$

STEP 2. Compute the principal space basis. Obtain from (2.4) the largest K eigenvalues and corresponding orthonormal eigenfunctions as the basis $\hat{\psi}_1(x), \ldots, \hat{\psi}_K(x)$. For computational methods and practical considerations, we refer to Section 8.4 of Ramsay and Silverman (2005).

As remarked by Hall et al. (2006), the operator defined by (2.4) is not necessarily positive semi-definite, but it is assured to have real eigenvalues. Theorem 1 of Hall et al. (2006) provides theoretical foundations that the bandwidths h_μ and h_ϕ should be chosen as $\mathcal{O}(T^{-1/5})$ to minimize the distance between the estimates $\widehat{\psi}$ and the corresponding true $\widehat{\psi}$. In Section 4 (details presented later), we find that the performance of $\hat{\beta}$ is very robust to the choice of the smoothing parameter.

We would like to emphasize that the space basis function $\widehat{\Psi}_t$ is only an estimator of the true (unobservable) Ψ_t. However, in proving the properties of the time basis selection, as in Theorem 3.2 and Corollary 3.1, we assume that this space basis estimation does not affect the study of selecting the temporal basis, because, otherwise, the non-asymptotic theoretical deviation will be too complex. If we still stick to the B-spline basis as in Park et al. (2009), all the proofs afterwards do not need to be modified. For simplicity of notation, we continue to use Ψ_t to denote this estimate of space basis from now.

We apply this method to the implied volatility modelling problem, which has been discussed in detail by Park et al. (2009). Figure 2 displays the space basis modelling using the functional PCA approach, which could capture the special 'smiling' effect well, while the spline basis modelling cannot.

Q7

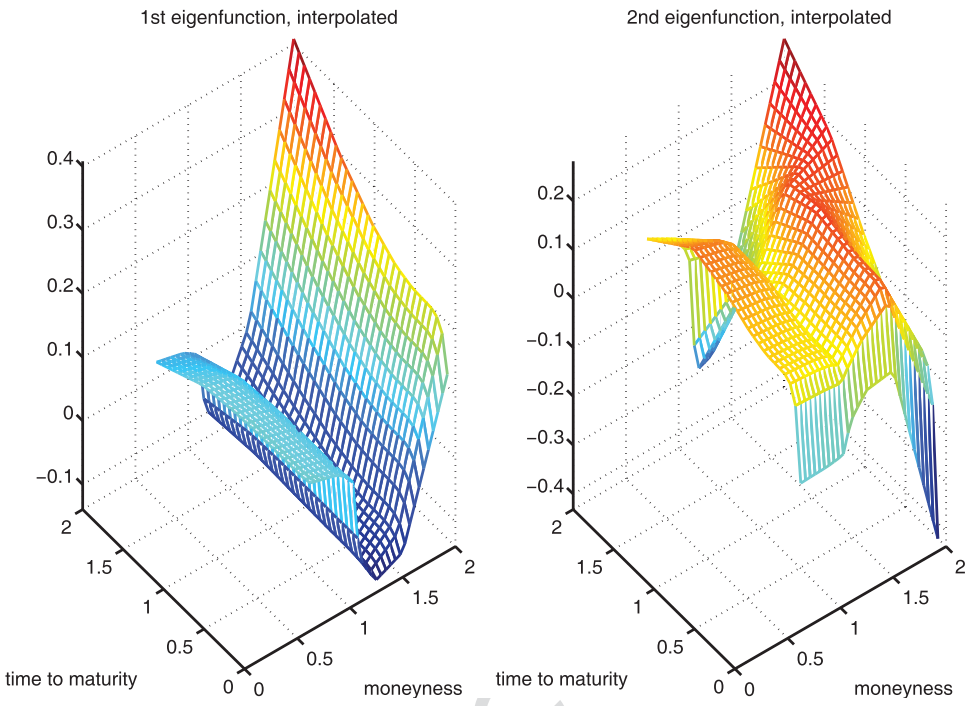


Figure 2. Space basis using the functional PCA approach for IVS modelling.

3. PROPERTIES OF ESTIMATES

In this section, we study sparse oracle inequalities for the estimate $\hat{\beta}$ defined in (A.1), assuming that the errors ε_t are dependent ($\boldsymbol{\beta}$-mixing in Theorem 3.2 and m-dependent in Corollary 3.1). This work extends those of Lounici et al. (2009), Bickel et al. (2009) and Lounici (2008) concerning upper bounds on the prediction error and the distance between the estimator and the true matrix β^*.

For the second step of the estimation procedure, an important question arises: is it justified, from an inferential point of view, to base further statistical inference on the detrended stochastic time series? Theorem 3.4 shows that the difference between the inference based on the estimated time series and true unobserved time series is asymptotically negligible.

Before stating the first theorem, we make the following assumption.

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

$$\min\left(\frac{\sqrt{\sum_t \|\Psi_t^\top \Delta U_t\|^2}}{\sqrt{T}\,\|\,\Delta_{\mathcal{R}}\,\|} : |\mathcal{R}| \leqslant s,\ \Delta \in \mathbb{R}^{K\times R}\backslash\{0\},\right.$$

$$\left.\|\,\Delta_{\mathcal{R}^c}\,\|_{2,1} \leqslant 3\,\|\,\Delta_{\mathcal{R}}\,\|_{2,1}\right) \geqslant \kappa,$$

where $\mathcal{R}^c$ denotes the complement of the set of indices $\mathcal{R}$ and $\Delta_{\mathcal{R}}$ 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 $\sum_{t=1}^{T} U_t U_t^\top$ as a function of sparsity s. In fact, it requires that the initially involved time basis is not too dependent, which is naturally satisfied by orthogonal polynomials and trigonometric functions. Low sparsity means that s is big and therefore κ is small. Thus, $\kappa(s)$ is a decreasing function of s; see also Lemma 4.1 of Bickel et al. (2009) for more details and related discussions.

THEOREM 3.1 (DETERMINISTIC PART). *Consider the model (2.2). Assume that $\Psi_t \Psi_t^\top = I_K$ (orthonormalized space basis), $T^{-1} \sum_{t=1}^{T} U_t^\top U_t / R = 1$, and the number of true non-zero time basis $M(\beta^*) \leqslant s$. If the random event*

$$\mathcal{A} = \left(2T^{-1} \max_{1 \leqslant r \leqslant R} \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{j=1}^{J} \Psi_{tkj}^\top \varepsilon_{tj} U_{tr} \leqslant \lambda \right) \tag{3.1}$$

holds for some $\lambda > 0$ and Assumption 3.1 is satisfied, then, for any solution $\widehat{\beta}$ of (A.1), we have

$$T^{-1} \sum_{t=1}^{T} \| \Psi_t^\top (\widehat{\beta} - \beta^*) U_t \|^2 \leqslant 16 s \lambda^2 \kappa^{-2}, \tag{3.2}$$

$$K^{-1/2} \| \widehat{\beta} - \beta^* \|_{2,1} \leqslant 16 s \lambda K^{-1/2} \kappa^{-2}, \tag{3.3}$$

$$M(\widehat{\beta}) \leqslant 64 \phi_{\max}^2 s \kappa^{-2}. \tag{3.4}$$

Note that Theorem 3.1 is valid for any J, R, T and any type of distribution of ε_t, and yields non-asymptotic bounds.

Because the standard assumption that ε_t is independent is often unsatisfied in practice, it is important to understand how the estimator behaves in a more general situation (i.e. with dependent error terms). As far as we know, our result is one of the first attempts to deal with dependent error terms for (group) Lasso variable selection techniques. We build it w.r.t. $\boldsymbol{\beta}$-mixing, which is an important measure of dependence between σ-fields (for time series). A detailed definition can be found in Appendix A. A very natural question to ask is, to what extent the degree of dependence (in terms of $\boldsymbol{\beta}$-mixing coefficients) is allowed, while we can still obtain certain sparse oracle inequalities (i.e. to study the relationship among high dimensionality R, moderate sample size T and $\boldsymbol{\beta}$-mixing coefficients β).

We use the following mild technical assumption similar to the typical bounded second-moment requirement for i.i.d. data.

ASSUMPTION 3.2. *The matrices Ψ_t and U_t and random variables ε_t are such that for $V_t \stackrel{\text{def}}{=} K^{-1/2} \sum_{k=1}^{K} \sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr}$, $\exists \sigma^2$ such that $\forall n, m$, $m^{-1} E[V_n + \ldots + V_{n+m}]^2 \leqslant \sigma^2$ and $\forall t$, $|V_t| \leqslant C''$, $\forall r$ and some constants $\sigma^2, C'' > 0$, $t = 1, \ldots, T$.*

Note that because V_t (as a function of ε_{tj}) is defined as a sum over j, it also indicates that the error term ε_t could be weakly cross-correlated. We can now state our main result.

THEOREM 3.2 ($\boldsymbol{\beta}$-MIXING). *Consider the model (2.2). Assume the sequence* $\{V_t\}_{t=1}^T$ *satisfies Assumption 3.2 and the* $\boldsymbol{\beta}$*-mixing condition with the* $\boldsymbol{\beta}$*-mixing coefficients*

$$\boldsymbol{\beta}\left(\left((3/8)\sigma\varepsilon^2 T^{1/2}(1-\varepsilon)^{1/2}C''^{-1}\log R^{-(1+\delta')/2}\right)-1\right)$$
$$\leqslant\left(24\sigma(1-\varepsilon)^{1/2}(R^{1+\delta'}\sqrt{\log R^{1+\delta'}TC''})^{-1}\right),$$

for any $\varepsilon>0$*, some* $\delta'>0$ *and* λ *defined below.* $\Psi_t\Psi_t^\top=I_K$*,* $T^{-1}\sum_{t=1}^T U_t^\top U_t/R=1$*, and* $M(\beta^*)\leqslant s$*. Furthermore, let* κ *be defined as in Assumption 3.1 and let* $\phi_{\max}$ *be the maximum eigenvalue of the matrix* $\sum_{t=1}^T U_t U_t^\top/T$*. Let*

$$\lambda=\sqrt{\frac{16\log R^{1+\delta'}K\sigma^2}{T(1-\varepsilon)}}.$$

Then, with a probability of at least $1-3R^{-\delta'}$*, for any solution* $\widehat{\beta}$ *of (A.1), we have*

$$T^{-1}\sum_{t=1}^T\parallel\Psi_t^\top(\widehat{\beta}-\beta^*)U_t\parallel^2\leqslant 256s\left(\frac{\log R^{1+\delta'}K\sigma^2}{T(1-\varepsilon)}\right)\kappa^{-2},\tag{3.5}$$

$$K^{-1/2}\parallel\widehat{\beta}-\beta^*\parallel_{2,1}\leqslant 96s\sqrt{\frac{\log R^{1+\delta'}\sigma^2}{T(1-\varepsilon)}}\kappa^{-2},\tag{3.6}$$

$$M(\widehat{\beta})\leqslant 64\phi_{\max}^2 s\kappa^{-2}.\tag{3.7}$$

REMARK 3.1. Before explaining the results, as also mentioned in Song and Bickel (2011), we would first like to discuss some related results. For technical simplicities, we consider the following simplest linear regression model with $R\to\infty$*:*

$$e_t=x_{t1}\theta_1+\ldots,x_{tR}\theta_R+\epsilon_t=x_t^\top\theta+\epsilon_t,\tag{3.8}$$

with the regressors $(x_{t1},\ldots,x_{tR})=x_t^\top$, the coefficients $(\theta_1,\ldots,\theta_R)=\theta^\top$ and the error term ϵ_t. Suppose x in (3.8) has full rank R and ϵ_t is $N(0,\sigma^2)$. Consider the least-squares estimate ($R\leqslant T$) $\hat{\theta}_{\mathrm{OLS}}=(xx^\top)^{-1}xe$. Then, from standard least-squares theory, we know that the prediction error $\|x^\top(\hat{\theta}_{\mathrm{OLS}}-\theta^*)\|_2^2/\sigma^2$ is χ_R^2-distributed, i.e.

$$E\left[\frac{\|x^\top(\hat{\theta}_{\mathrm{OLS}}-\theta^*)\|_2^2}{T}\right]=\frac{\sigma^2}{T}R.\tag{3.9}$$

In the sparse situation, if ϵ_t is $N(0,\sigma^2)$ (different from our case), Corollary 6.2 of Bühlmann and van de Geer (2011) shows that the Lasso estimate obeys the following oracle inequality:

$$\frac{\|x^\top(\hat{\theta}_{\mathrm{Lasso}}-\theta^*)\|_2^2}{T}\leqslant C_0\frac{\sigma^2\log R}{T}M(\theta^*),\tag{3.10}$$

with a large probability and some constant C_0. The additional $\log R$ factor here could be seen as the price to pay for not knowing the set $\{\theta_p^*,\theta_p^*\neq 0\}$ (Donoho and Johnstone, 1994). Similar to the i.i.d. Gaussian situation discussed above, the term $(\log R)^{1+\delta'}$ in (3.5) could be interpreted as

the price to pay for not knowing the set $\{\beta_r^*, \theta_r^* \neq 0\}$. Here, we have $(\log R)^{1+\delta'}$ instead of $\log R$ because we deviate from the typical i.i.d. Gaussian situation and establish the results under the more general Assumption 3.2, which can be thought of as the finite second-moment condition. Also, the δ' term is the price to pay for this deviation.

REMARK 3.2. Because

$$\boldsymbol{\beta}\left(\left((3/8)\sigma\varepsilon^2 T^{1/2}(1-\varepsilon)^{1/2}C''^{-1}\log R^{-(1+\delta')/2}\right)-1\right)$$
$$\leqslant\left(24\sigma(1-\varepsilon)^{1/2}(R^{1+\delta'}\sqrt{\log R^{1+\delta'}T}C'')^{-1}\right)$$

is required, when dimensionality R increases, the allowed dependence level reflected by the $\boldsymbol{\beta}$-mixing coefficients must decrease fast enough so that we still achieve similar risk bounds as in the independent case. Intuitively, this makes sense because if the dependence level inherent in $Z_{0,t}$ (or ε_t equivalently) is too strong (i.e. β exceeds some level), then the amount of information provided by these observations is less, and therefore the estimate does not perform well. However, strong dependence in $Z_{0,t}$ might be caused by some trend, which should be included in $U_t^\top\Gamma$, but is not, which results in the increased dependence. This tells us that at the beginning, we should include a large enough number R of pre-specified time basis functions such that it could include most of the deterministic (even though it could be segment by segment) time evolution and the remaining dependence level in $Z_{0,t}$ is controlled.

COROLLARY 3.1 (m-DEPENDENT). *Consider the model (2.2). Assume that the sequence $\{V_t\}_{t=1}^T$ is an m-dependent process with order k ($k \geqslant 1$) and satisfies the following conditions for some constants $\sigma_0^2, C'' > 0, t = 1, \ldots, T$: (a) $\forall t$, $E[V_t^2] \leqslant \sigma_0^2$, $|V_t| \leqslant C''$; (b) $((3/8)\sigma\varepsilon^2 T^{1/2}(1-\varepsilon)^{1/2}C''^{-1}\log R^{-(1+\delta')/2}) - 1 \geqslant k+1$ for any $\varepsilon > 0$ and some $\delta' > 0$. Also, $\Psi_t\Psi_t^\top = I_K$, $T^{-1}\sum_{t=1}^T U_t^\top U_t/R = 1$, and $M(\beta^*) \leqslant s$. Furthermore, let κ be defined as in Assumption 3.1, let $\phi_{\max}$ be the maximum eigenvalue of the matrix $\sum_{t=1}^T U_t U_t^\top/T$ and let λ be defined as in Theorem 3.2. Then, with a probability of at least $1 - 3R^{-\delta'}$, for any solution $\widehat{\beta}$ of (A.1), we have*

$$T^{-1}\sum_{t=1}^{T} \| \Psi_t^\top(\widehat{\beta}-\beta^*)U_t \|^2 \leqslant 512s\left(\frac{\log R^{1+\delta'}Kk\sigma_0^2}{T(1-\varepsilon)}\right)\kappa^{-2}, \tag{3.11}$$

$$K^{-1/2}\| \widehat{\beta}-\beta^* \|_{2,1} \leqslant 96\sqrt{2}s\sqrt{\frac{\log R^{1+\delta'}k\sigma_0^2}{T(1-\varepsilon)}}\kappa^{-2}, \tag{3.12}$$

$$M(\widehat{\beta}) \leqslant 64\phi_{\max}^2 s\kappa^{-2}. \tag{3.13}$$

REMARK 3.3. We can see that when k increases (i.e. the dependence in $\{V_t\}_{t=1}^T$ becomes stronger and stronger), the risk bounds become larger and larger. To ensure $((3/8)\sigma\varepsilon^2 T^{1/2}(1-\varepsilon)^{1/2}C''^{-1}) - 1 \geqslant k+1$, approximately we need $T^{1/2}\log R^{-(1+\delta')/2} \geqslant ((3/4)\sigma_0\varepsilon^2\sqrt{(1-\varepsilon)})^{-1}C''\sqrt{k}$, which gives the requirement on the sample size T (relative to the high dimensionality) and the amount of information from the data. Similar results could also be separately obtained for the generalized m-dependent process based on fractional cover theory and the (extended) McDiarmid inequality; see Theorem 2.1 of Janson (2004). At the second step, $Z_{0,t}$ is estimated based on $\widehat{\beta}$ instead of β^*, so we need to show that the influence of this

plug-in estimate is negligible. Our result relies on the following assumptions, which are similar to Assumptions (A1)–(A8) in Park et al. (2009).

ASSUMPTION 3.3. *(a) The sets of variables* $(X_{1,1}, \ldots, X_{T,J})$, $(\varepsilon'_{1,1}, \ldots, \varepsilon'_{T,J})$ *and* $(Z_{0,1}, \ldots, Z_{0,T})$ *are independent of each other; (b) 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* $t = 1, \ldots, T$; *(c) we assume that* $E[\varepsilon'_{t,j}] = 0$ *for* $1 \leq t \leq T$, $1 \leq j \leq J$, *and for* $c > 0$ *small enough,* $\sup_{1\leq t\leq T, 1\leq j\leq J} E[\exp(c(\varepsilon'_{t,j})^2)] < \infty$; *(d) 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 $K \to \infty$, *and we denote* A *that fulfils* $\sup_{x\in[0,1]^d} \|m(x) - A\Psi(x)\| \leq 2\delta_K$ *by* A^*; *(e) there exist constants* $0 < C_L < C_U < \infty$ *such that all eigenvalues of the matrix* $T^{-1}\sum_{t=1}^T Z_{0,t}Z_{0,t}^\top$ *lie in the interval* $[C_L, C_U]$ *with probability tending to one; (f) for all* β *and* A ($\beta^\top = \Gamma A$) *in (A.1), with probability tending to one, we have*

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

where the constant M_T *satisfies* $\max_{1\leqslant t\leqslant T} \|Z_{0,t}\| \leqslant M_T/C_m$ *for a constant* C_m *such that* $\sup_{x\in[0,1]^d} \|m(x)\| < C_m$; *(g) it holds that* $\rho^2 = (K + T)M_T^2 \log(JTM_T)/(JT) \to 0$, *and the dimension* L *is fixed.*

Assumption 3.3 (f) and the additional bound M_T in the minimization are introduced purely for technical reasons. They are similar to the assumption that V_t is upper bounded in Assumption 3.2 by noticing $V_t = K^{-1/2}\sum_{k=1}^K \sum_{j=1}^J \Psi_{tkj}\varepsilon_{tj}U_{tr}$ and $\varepsilon_t = Z_{0,t}^\top A^*\Psi_t + \varepsilon'_t$. Recall that given β, the number of parameters still needing to be estimated equals KT ($\{Z_{0,t}\}_{t=1}^T$) and KL (A) (given β, if A is fixed, Γ is also fixed). Because L is fixed, Assumption 3.3(g) basically requires that, neglecting the factors $M_T^2 \log(JTM_T)$, the number of parameters grows slower than the number of observations JT.

THEOREM 3.3. *Suppose that model (2.1), all assumptions in Theorem 3.2 and Assumption 3.3 hold. Then, we have*

$$\frac{1}{T}\sum_{1\leq t\leq T} \left\| \widehat{Z}_{0,t}^\top \widehat{A} - Z_{0,t}^\top A^*\right\|^2 = \mathcal{O}_P\left(\rho^2 + \delta_K^2\right). \tag{3.14}$$

In the following, we discuss how statistical analysis differs if the inference of stochasticity on $Z_{0,t}$ is based on $\widehat{Z}_{0,t}$ instead of using the (unobserved) process $Z_{0,t}$. We establish theoretical properties under a strong mixing condition, which is more general than the $\boldsymbol{\beta}$-mixing considered in Theorem 3.2. For the statement of the theorem, we need the following assumptions, which are similar to Assumptions (A9)–(A11) in Park et al. (2009).

ASSUMPTION 3.4. *(a) (i)* $Z_{0,t}$ *is a strictly stationary sequence with* $E[Z_{0,t}] = 0$, $E[\|Z_{0,t}\|^\gamma] < \infty$ *for some* $\gamma > 2$; *(ii) it is* $\boldsymbol{\alpha}$*-mixing with* $\sum_{i=1}^\infty \boldsymbol{\alpha}(i)^{(\gamma-2)/\gamma} < \infty$; *(iii) the matrix* $E[Z_{0,t}Z_{0,t}]^\top$ *has full rank; (iv) the process* $Z_{0,t}$ *is independent of* $X_{11}, \ldots, X_{TJ}, \varepsilon'_{11}, \ldots, \varepsilon'_{TJ}$. *(b) It holds that* $(\log(KT)^2((KM_T/J)^{1/2} + T^{1/2}M_T^4J^{-2} + K^{3/2}J^{-1} \quad + K^{4/3}J^{-2/3}T^{-1/6}) \quad +1)T^{1/2}(\rho^2 + \delta_K^2) = \mathcal{O}(1)$.

Assumption 3.4(b) imposes a very weak condition on the growth of J, K and T. Suppose, for example, that M_T is of logarithmic order and that K is of order $(JT)^{1/5}$, then the condition requires that T/J^2 times a logarithmic factor converges to zero. As remarked by Doukhan (1994), if a stochastic process is $\boldsymbol{\beta}$-mixing, then it is also $\boldsymbol{\alpha}$-mixing with $2\boldsymbol{\alpha}(\mathcal{A}, \mathcal{B}) \leqslant \boldsymbol{\beta}(\mathcal{A}, \mathcal{B})$. If the requirement on the $\boldsymbol{\beta}$-mixing coefficient in Theorem 3.2 is satisfied, then the requirement on the $\boldsymbol{\alpha}$-mixing coefficient in Assumption 3.4(a) is usually satisfied.

Furthermore, note that the minimization problem (A.1) only has a unique solution in β, but not in Γ and A. If $(\widehat{Z}_{0,t}, \widehat{A})$ is a minimizer, then so is $(B^\top \widehat{Z}_{0,t}, B^{-1}A)$, where B is an arbitrary invertible matrix. With the choice $B = (\sum_{t=1}^T Z_{0,t}\widehat{Z}_{0,t}^\top)^{-1} \sum_{t=1}^T Z_{0,t} Z_{0,t}^\top$, we obtain $\sum_{t=1}^T Z_{0,t}(\widetilde{Z}_{0,t} - Z_{0,t})^\top = 0$, where $\widetilde{Z}_{0,t} \stackrel{\text{def}}{=} B^\top \widehat{Z}_{0,t}$ and $\widetilde{A} \stackrel{\text{def}}{=} B^{-1}A$. Without loss of generality, we can assume $T^{-1}\sum_{s=1}^T \widehat{Z}_{0,s} = T^{-1}\sum_{s=1}^T Z_{0,s} = 0$. Additionally, we define

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

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

THEOREM 3.4. *Suppose that model (2.1) holds. Besides all assumptions in Theorem 3.2, also let Assumptions 3.3 and 3.4 be satisfied. Then, there exists a random matrix B specified above such that, for $h \geq 0$,*

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

and

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

In Theorem 3.4, we consider the autocovariances of the estimated stochastic process $\widehat{Z}_{0,t}$ and the (unobserved) process $Z_{0,t}$, and we show that these estimators differ only by second-order terms. Thus, the statistical analysis based on $\widehat{Z}_{0,t}$ is equivalent to that based on the (unobserved) process $Z_{0,t}$.

4. DYNAMICS OF TEMPERATURE ANALYSIS

Since the first transaction in the weather derivatives market in 1971, the market has expanded rapidly. Many companies, who faced the possibility of significant declines in earnings because of abnormal weather fluctuations, decided to hedge their seasonal weather risk. Thus, weather derivative contracts have become particularly attractive. One essential task is to model the fluctuations of temperatures at many different weather stations. Thus, in this section, we present the application to the analysis of temperature dynamics by fitting the daily temperature observations provided by the CDC of the CMA; see Figure 1. To capture the upward trend, seasonal and large-period effects, similar to Racsko et al. (1991), Parton and Logan (1981) and

Table 1. Initial choice of $53 \times 3 + 16 = 175$ time basis.

	Factors		Factors
Trend (year by year)	1	Large period	$\sin 2\pi t/(365 \times 15)$
	t		$\cos 2\pi t/(365 \times 15)$
	$3t^2 - 1$		$\sin 2\pi t/(365 \times 10)$
Seasonal effect	$\sin 2\pi t/365$		$\cos 2\pi t/(365 \times 10)$
	$\cos 2\pi t/365$		$\sin 2\pi t/(365 \times 5)$
	...		$\cos 2\pi t/(365 \times 5)$
	$\cos 10\pi t/365$		

Hedin (1991), we propose the following initial choice of time basis (rescaling factors omitted) in Table 1.

For the space basis, when we consider the relative proportion of variance explained by the first K basis (eigenvalues of the smoothed covariance operator) and the five climate types of China, as shown in Figure 3, the number of space basis $K = 5$ is appealing. As we discuss in Appendix A, the choice of tuning parameter λ is crucial here. Figure 4 presents the solution path of four different selection criteria, C_p, GCV, AIC and BIC, evaluated on 500 equally spaced values of λ, where the minimizer is marked as the red dot. As we can see, the minimizers of C_p, GCV and AIC are significantly smaller than that of BIC, which confirms previous discussions in the literature that the AIC-type criterion (including GCV and C_p) tends to overestimate the model size and thus overfits. Our estimate also involves the smoothing bandwidth in the smoothed functional PCA step, which, by Theorem 1 of Hall et al. (2006), should be chosen as $\mathcal{O}(T^{-1/5})$ in order to minimize the distance between the estimates of the $\widehat{\psi}$ eigenfunctions and the corresponding true ones. Figure 5 presents the BIC solution path w.r.t. four different (by a constant factor) values of the smoothing parameter for the same 500 values of λ as above. As we can see, the solution path is very stable w.r.t. the choice of the smoothing parameter.

Figure 6 displays the estimated coefficients of the first factor with respect to the 54×3 yearly polynomial time basis w.r.t. $k = 1$ under the optimal choice of λ selected by the BIC criterion. The coefficients of constant, linear and quadratic terms are displayed as solid, dashed and dotted lines, respectively, and they are also coupled with the corresponding 90% confidence intervals (based on year-by-year ordinary least-squares (OLS) estimates) represented by the thin lines (with the same colour and style). The fact that all these coefficients are non-negative indicates that over the past 50 years, there might have been a warming effect across China. The confidence intervals are computed using OLS polynomial fitting to the year-by-year time series after removing the normal seasonality and large-period effects. We observe an unusual large positive (w.r.t. the linear term) and negative (w.r.t. the quadratic term) variation for the OLS estimates at the end of the 1960s, caused by the extreme temperatures in China at that time. By employing shrinkage techniques, we can remove this disadvantage and produce stabler estimates. The estimated coefficients of the five factors w.r.t. the 16 trigonometric functions time basis corresponding to the optimal λ are displayed in Table 2. It clearly indicates that the 15-year period effect, as some meteorologists claim, is related to solar activity.

Because the eigenvalues of $\widehat{\beta}\widehat{\beta}^\top$ are $(10140, 208, 118, 44, 14, 0, 0, \ldots)$ (with the first five being non-zero and the rest being zero), we choose $L = 5$ and obtain the remaining five-dimensional random process $\widehat{Z}_{0,t}$, which could be further modelled by using multivariate time

Q9

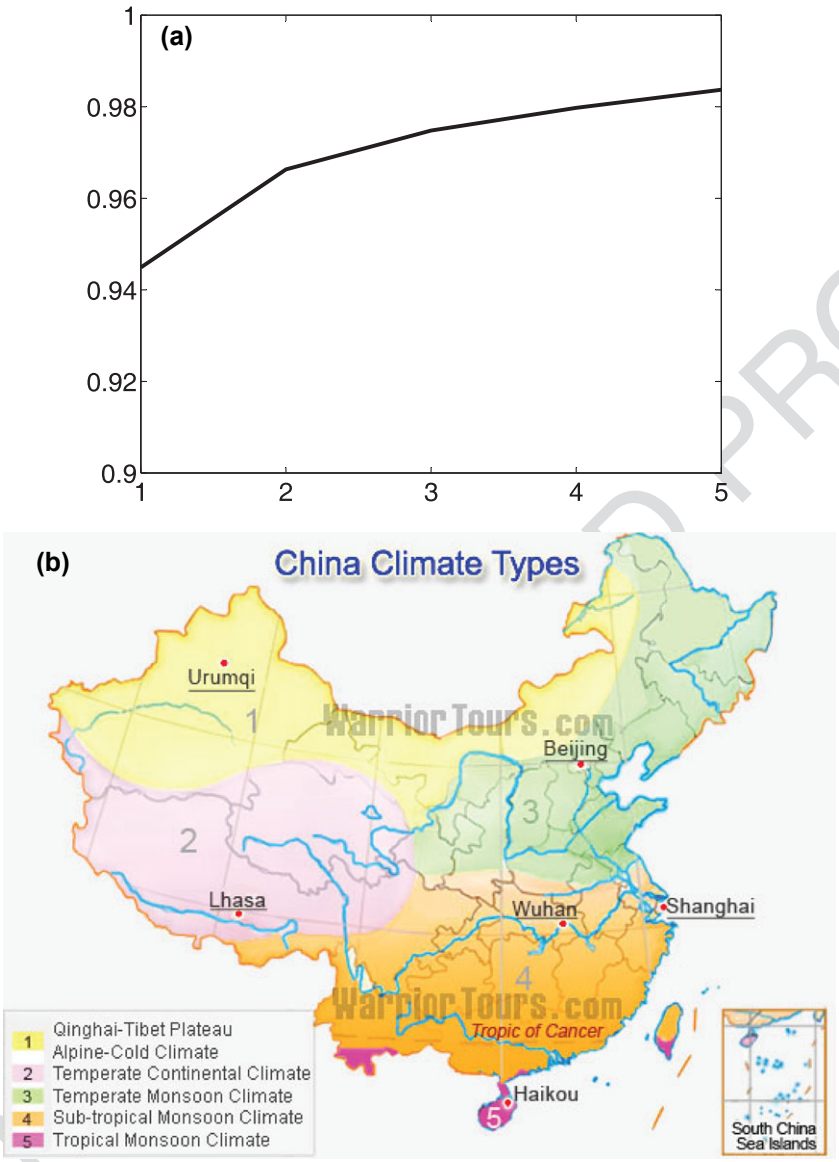


Figure 3. Relative proportion of variance and China's climate types.

series techniques. For example, if we use a VAR(1) process, $\widehat{Z}_{0,t} = \mathcal{S}\widehat{Z}_{0,t-1} + \varepsilon_{0,t}$, where $\varepsilon_{0,t}$ is a random vector, then the estimated coefficient matrix is

$$\begin{pmatrix} 0.7703 & 0.0103 & 0.0007 & 0.0015 & 0.0005 \\ -0.0552 & 0.1449 & -0.1841 & -0.0285 & 0.0003 \\ -0.3047 & -0.3419 & 0.3877 & -0.0436 & -0.0020 \\ 0.2078 & -0.1717 & -0.1337 & 0.8431 & 0.0071 \\ 0.6345 & -0.0484 & -0.0447 & 0.0184 & 0.8338 \end{pmatrix}.$$

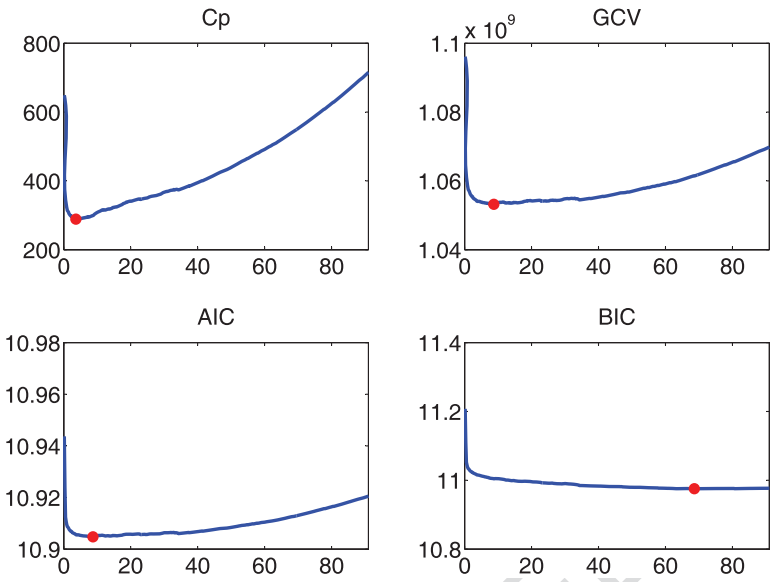


Figure 4. Comparison of C_p, GCV, AIC and BIC.

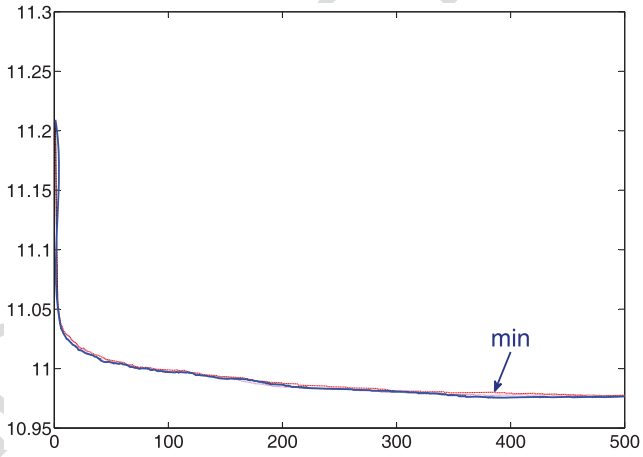


Figure 5. BIC solution path.

Compared with the existing temperature modelling (pricing weather derivatives) techniques (e.g. Benth and Benth, 2005), our approach possesses the following advantages. First, based on high-dimensional time series data, it offers integrated analysis considering space (high dimensionality) and time (dynamics) parts simultaneously, while forecasting at different places other than the existing weather stations is also possible because the space basis is actually a function of the geographical location information. Second, it extracts the trend more clearly. Third, it provides theoretical justification for further inferential analysis of $\widehat{Z}_{0,t}$ instead of $Z_{0,t}$.

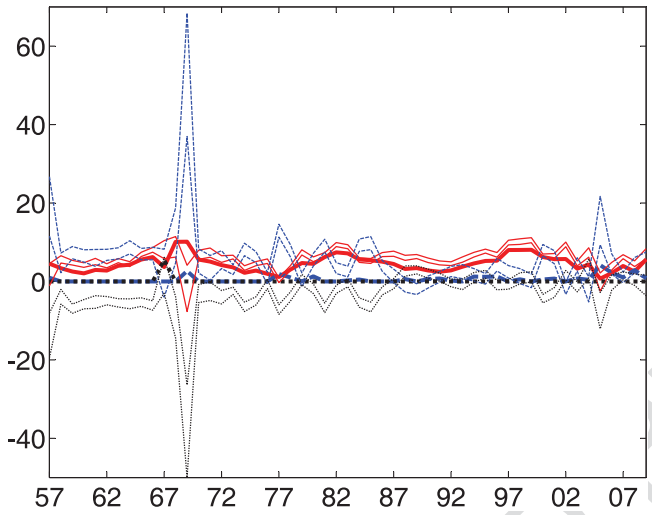


Figure 6. Estimated coefficients of the 54 × 3 yearly polynomial time basis.

Table 2. Estimated coefficients of the five factors.

Basis	Estimates				
$\sin 2\pi t/365$	−25.4922	1.1059	2.4129	−2.6985	1.2320
$\cos 2\pi t/365$	−87.3303	1.8228	5.3358	−5.0823	1.6284
$\sin 4\pi t/365$	0.0000	0.0000	0.0000	0.0000	0.0000
$\cos 4\pi t/365$	−4.5532	0.8761	0.6752	−0.6709	0.9163
...	0.0000	...			
$\cos 10\pi t/365$	0.0000	...			
$\sin 2\pi t/(365 \times 15)$	11.7818	−0.0053	−1.4026	0.4743	−0.0214
$\cos 2\pi t/(365 \times 15)$	0.0000	...			
...	0.0000	...			
$\cos 2\pi t/(365 \times 5)$	0.0000	...			

5. SIMULATION STUDY

Because the simulation results about the performance of the group-Lasso estimator have been well illustrated in the literature, to evaluate the overall fitting performance of the GDSFM, we conduct a Monte Carlo experiment designed to mimic the previous empirical example.

We generate random variables $\beta_1, \ldots, \beta_{175} \in \mathbb{R}^4$ such that all coordinates are i.i.d. standard normal random variables. We randomly pick 80% of the β_r coefficients from $\beta_1, \ldots, \beta_{175}$ and assign them to be $0 \in \mathbb{R}^4$. We choose the same time basis as in Table 1 with $p = 365$ and $T = 19345$. For the space part, inspired by Park et al. (2009), we consider $d = 2$ and the following

Table 3. Average values of $1 - R^2$.

	Independent	Weakly dependent	Strongly dependent
$1 - R^2$	5.30%	5.32%	5.40%

tuples of two-dimensional functions:

$$m_1(x_1, x_2) = 1, \quad m_2(x_1, x_2) = 3.46\,(x_1 - 0.5),$$
$$m_3(x_1, x_2) = 9.45\big((x_1 - 0.5)^2 + (x_2 - 0.5)^2\big) - 1.6,$$
$$m_4(x_1, x_2) = 1.41\sin(2\pi x_2).$$

These functions are chosen to be close to orthogonal. The design points $X_{t,j}$ are independently generated from a uniform distribution on the unit square. We generate $Y_t^\top = U_t^\top \beta^\top \Psi_t + \varepsilon_t$, $t = 1, \ldots, T$ with the following three types of error distributions:

(1) all coordinates of $\varepsilon_1, \ldots, \varepsilon_T$ are i.i.d. $N(0, 0.05)$ random variables;
(2) ε_t are generated from a centred VAR(1) process $\varepsilon_t = \mathcal{S}\varepsilon_{t-1} + \eta_t$, where $\mathcal{S}$ is a diagonal matrix with all diagonal entries equal to 0.4 and all entries of η_t are $N(0, 0.84 \times 0.05)$ random variables (such that $\mathrm{Var}(\varepsilon_t)$ is still the same as that of the independent case);
(3) the same as above except that all diagonal entries of $\mathcal{S}$ equal 0.8 (i.e. a stronger dependence level and η_t are $N(0, 0.36 \times 0.05)$ random variables).

The algorithm presented in (B.3) converges fast (with a tolerance of 10^{-3}). The values of β are estimated by the group-Lasso technique as in (A.1) with tuning parameter λ selected by the BIC-type criterion, as in (A.2). After obtaining $\hat{\beta}$, we further estimate the stochastic process $Z_{0,t}$ by a VAR(1) model. We take the remaining variation $(1 - R^2)$ as a measure of the fitting performance, where

$$1 - R^2 = \frac{\sum_{t=1}^{T} \left\| Y_t^\top - \big(U_t^\top \widehat{\Gamma} + \widehat{Z}_{0,t}^\top\big)\widehat{A}\widehat{\Psi}_t \right\|_2^2}{\sum_{t=1}^{T} \left\| Y_t^\top - \sum_{t=1}^{T}\sum_{j=1}^{J} Y_{t,j}/JT \right\|_2^2} \tag{5.1}$$

is the proportion of the remaining variation not explained by the model among total variation. We repeat this experiment 100 times and present the average values of $1 - R^2$ in Table 3 for the independent, weakly dependent and strongly dependent cases. As we can see, when the dependence level (in ε_t) increases, even though the remaining variation slightly increases because of the worse estimates of β, overall it is still relatively good.

6. CONCLUDING REMARKS

In this paper, we provide an integrated and yet flexible model for high-dimensional non-stationary time series that reveals both complex trends and stochastic components. When applying GDSFMs, we employ a non-parametric series expansion for both temporal and spatial components. After choosing smoothed (non-parametric) functional principal components as a space basis and extracting temporal trends utilizing time basis function selection techniques, the estimate's properties are investigated under the dependent scenario, together with the weakly

cross-correlated error term. This is not built upon any specific forms of time and space basis. This enables us to explore the interplay among the degree of time dependence, high dimensionality and moderate sample size (relative to dimensionality). The presented theory is an extension to the current regularization techniques. We further justify statistical inference, e.g. estimation and classification based on the detrended low-dimensional stochastic process. Applications to the dynamic behaviour analysis of temperatures confirm its power.

REFERENCES

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data* (3rded.). New York, NY: Wiley.

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

Bickel, P. J., Y. Ritov and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics 37*, 1705–32.

Bowsher, C. G. and R. Meeks (2006). High-dimensional yield curves: models and forecasting. Working Paper 2006-W12, Nuffield College, University of Oxford.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Heidelberg: Springer.

Cai, Z., Q. Li and J. Y. Park (2009). Functional-coefficient models for non-stationary time series data. *Journal of Econometrics 148*, 101–13.

Campbell, J. Y. and M. Yogo (2006). Efficient tests of stock return predictability. *Journal of Financial Economics 81*, 27–60.

Christiano, L., M. Eichenbaum and C. Evans (1999). Monetary policy shocks: what have we learned and to what end? In J. B. Taylor and M. Woodford (Eds.), *Handbook of Macroeconomics*, Volume *1*, 65–148. Amsterdam: Elsevier.

Diebold, F. X. and C. Li (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics 130*, 337–64.

Donoho, D. L. and I. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*, 425–55.

Doukhan, P. (1994). *Mixing: Properties and Examples*. Heidelberg: Springer.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–60.

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

Forni, M., M. Hallin, M. Lippi and L. Reichlin (2000). The generalized dynamic-factor model: identification and estimation. *Review of Economics and Statistics 82*, 540–54.

Forni, M., M. Hallin, M. Lippi and L. Reichlin (2005). The generalized dynamic factor model: one-sided estimation and forecasting. *Journal of the American Statistical Association 100*, 830–40.

Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge: Cambridge University Press.

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

Geweke, J. (1977). The dynamic factor analysis of economic time series. In D. J. Aigner and A. S. Goldberg (Eds.), *Latent Variables in Socio-Economic Models*, 365–83. Amsterdam: North-Holland.

Giannone, D., L. Reichlin and L. Sala (2005). Monetary policy in real time. In M. Gertler and K. Rogoff (Eds.), *NBER Macroeconomics Annual 2004, Volume 19*, 161–224. Cambridge, MA: National Bureau of Economic Research.

Gleick, P. H. et al. (2010). Climate change and the integrity of science. *Science 328*, 689–90.

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

Hall, P., H. G. Müller and J. L. Wang (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics 34*, 1493–517.

Hallin, M. and R. Liska (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association 102*, 603–17.

Hedin, A. E. (1991). Extension of the msis thermosphere model into the middle and lower atmosphere. *Journal of Geophysical Research 96*, 1159–72.

Horowitz, J. L. (2006). Testing a parametric model against a non-parametric alternative with identification through instrumental variables. *Econometrica 74*, 521–38.

Horowitz, J. and J. Huang (2012). Penalized estimation of high-dimensional models under a generalized sparsity condition. CWP 17/12, Centre for Microdata Methods and Practice, Institute for Fiscal Studies and University College London.

Horowitz, J. L. and S. Lee (2005). Non-parametric estimation of an additive quantile regression model. *Journal of the American Statistical Association 100*, 1238–49.

Horowitz, J., J. Klemelä and E. Mammen (2006). Optimal estimation in additive regression models. *Bernoulli 12*, 271–98.

Hsiao, C. (1986). *Analysis of Panel Data*. Econometric Society Monographs No. 11. Cambridge: Cambridge University Press.

Huang, J., J. L. Horowitz and F. Wei (2010). Variable selection in non-parametric additive models. *Annals of Statistics 38*, 2282–313.

Janson, S. (2004). Large deviations for sums of partly dependent random variables. *Random Structures Algorithms 24*, 234–48.

Lee, R. D. and L. Carter (1992). Modeling and forecasting the time series of U.S. mortality. *Journal of the American Statistical Association 87*, 659–71.

Leng, C., Y. Lin and G. Wahba (2006). A note on the Lasso and related procedures in model selection. *Statistica Sinica 16*, 1273–84.

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., M. Pontil, A. B. Tsybakov and S. van de Geer (2009). Taking advantage of sparsity in multi-task learning. In S. Dasgupta and A. Klivans (Eds.), *Proceedings of the 22nd Conference on Learning Theory (COLT) 2009*, 73–82. Madison, WI: Omnipress.

Mohr, P. N. C., G. Biele, L. K. Krugel, S-C. Li and H. R. Heekeren (2010). Neural foundations of risk-return trade-off in investment decisions. *NeuroImage 49*, 2556–63.

Myšičková, A., S. Song, P. N. Mohr, H. R. Heekeren and W. K. Härdle (2013). Risk patterns and correlated brain activities. Forthcoming in Psychometrika (doi:10.1007/s11336-013-9352-2).

Nelson, C. R. and A. F. Siegel (1987). Parsimonious modeling of yield curves. *Journal of Business 60*, 473–89.

Odening, M., E. Berg and C. Turvey (2008). Management of climate risk in agriculture. *Special Issue of the Agricultural Finance Review 68*, 83–97.

Park, B. U., E. Mammen, W. Härdle and S. Borak (2009). Time series modelling with semi-parametric factor dynamics. *Journal of the American Statistical Association 104*, 284–98.

Parton, W. J. and J. A. Logan (1981). A model for diurnal variation in soil and air temperature. *Agricultural Meteorology 23*, 205–16.

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

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis* (2nded.). Berlin: Springer.

Sargent, T. J. and C. A. Sims (1977). Business cycle modeling without pretending to have too much a priori economic theory. Working Paper 55, Federal Reserve Bank of Minneapolis.

Song, S. and P. Bickel (2011). Large vector auto regressions. Working Paper, University of California at Berkeley (arXiv:1106.3915).

Stock, J. H. and M. W. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–79.

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

Stock, J. H. and M. W. Watson (2005a). An empirical comparison of methods for forecasting using many predictors. Manuscript, Princeton University.

Stock, J. H. and M. W. Watson (2005b, July). Implications of dynamic factor models for VAR analysis. Working Paper 11467, National Bureau of Economic Research. 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*, 267–88.

Wang, H. and C. Leng (2008). A note on adaptive group Lasso. *Computational Statistics and Data Analysis 52*, 5277–86.

Wang, H., G. Li and C-L. Tsai (2007a). Regression coefficient and autoregressive order shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Series B *69*, 63–78.

Wang, H., R. Li and C-L. Tsai (2007b). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*, 553–68.

Wang, Q. and P. C. B. Phillips (2009a). Asymptotic theory for local time density estimation and non-parametric cointegrating regression. *Econometric Theory 25*, 710–38.

Wang, Q. and P. C. B. Phillips (2009b). Structural non-parametric cointegrating regression. *Econometrica 77*, 1901–48.

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

Xiao, Z. (2009). Functional-coefficient cointegration models. *Journal of Econometrics 152*, 81–92.

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

Q11

APPENDIX A: ESTIMATION PROCEDURE

First, we present the estimation method.

Step 1. Given the pre-specified time and space basis, find significantly loaded time basis functions (i.e. coefficients β) utilizing the group-Lasso technique by minimizing

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

Here, we use T^{-1} instead of $(JT)^{-1}$ because the space basis has been orthonormalized ($\widehat{\Psi}_t\widehat{\Psi}_t^\top = I_K$).

Step 2. Split the joint matrix $\widehat{\beta}$ into two separate coefficient matrices $\widehat{\Gamma}$ and $\widehat{A}$ by taking $\widehat{\Gamma}$ as the L eigenvectors of $\widehat{\beta}\widehat{\beta}^\top$ (w.r.t. the L largest eigenvalues) and $\widehat{A} = \widehat{\Gamma}^\top\widehat{\beta}$. Given $Y_t^\top - U_t^\top\hat{\beta}^\top\Psi_t$ and $\widehat{A}$, Ψ_t, estimate $Z_{0,t}$ by the OLS method.

It is worth noting that both Γ (and $Z_{0,t}$, respectively) and A are unidentifiable in model (2.1), because trivially $\Gamma^*A^* = (\Gamma^*B)(B^{-1}A^*)$. However, if we concentrate on prediction, the identification of β (as a product of Γ and A, as in (A.1)) is enough. Additionally, we show that for any version of $\{Z_{0,t}\}$, there exists a version of $\{\widehat{Z}_{0,t}\}$ whose lagged covariances are asymptotically the same as those of $\{Z_{0,t}\}$.

The group-Lasso estimates depend on the tuning parameter λ. We implement an easily computable BIC-type criterion. The solution path is computed by evaluating some criteria on equally spaced λ's between 0 and $\lambda_{\max} = \max_r \parallel \sum_t \Psi_t Y_t U_{tr} \parallel$. We select the λ that minimizes

$$\mathrm{BIC}(\lambda) = \log\left(\sum_t \parallel Y_t^\top - U_t^\top\widehat{\beta}^\top\Psi_t \parallel^2/T\right) + \log T \cdot df/T, \tag{A.2}$$

$$df = \sum_r \mathbf{1}\Big(\parallel \widehat{\beta}_r \parallel > 0\Big) + \sum_r \frac{\parallel \widehat{\beta}_r \parallel}{\parallel \widehat{\beta}^{\mathrm{OLS}} \parallel}(K-1).$$

For reference purposes, we also list the formulae of the C_p, GCV and AIC criteria:

$$C_p(\lambda) = \sum_t \parallel Y_t^\top - U_t^\top\widehat{\beta}^\top\Psi_t \parallel^2/\tilde{\sigma}^2 - T + 2df;$$

$$\tilde{\sigma}^2 = \sum_t \parallel Y_t^\top - U_t^\top\widehat{\beta}_{\mathrm{OLS}}^\top\Psi_t \parallel^2/(T-df);$$

$$\mathrm{GCV}(\lambda) = \sum_t \parallel Y_t^\top - U_t^\top\widehat{\beta}^\top\Psi_t \parallel^2/(1-df/T)^2;$$

$$\mathrm{AIC}(\lambda) = \log\left(\sum_t \parallel Y_t^\top - U_t^\top\widehat{\beta}^\top\Psi_t \parallel^2/T\right) + 2df/T.$$

As pointed out by Yuan and Lin (2006) (for i.i.d. data), the performance of this approximate information criterion is generally comparable with that of the computationally much more expensive (especially for the massive data) fivefold cross-validation. More importantly, because the data here are observed in time, the order of observations is significant, and hence a simple cross-validation procedure is inappropriate in a time series context. Besides BIC, there are other parameter selection criteria, such as C_p, GCV and AIC. In terms of variable selection, Wang and Leng (2008) have found that BIC is superior to C_p. The reason for this is that when there exists a true model, AIC-type criteria (including GCV and C_p) tend to overestimate the model size; see, e.g. Leng et al. (2006), Wang et al. (2007a) and Wang et al. (2007b). Subsequently, estimation accuracy using C_p can suffer. Wang et al. (2007b) have given a theoretical justification showing that GCV overfits the smoothly clipped absolute deviation (SCAD) method (Fan and Li, 2001). Analogous arguments also apply to the C_p methods.

APPENDIX B: TECHNICAL PROOFS

In order to study the statistical properties of this estimator, it is useful to derive some optimality conditions for a solution of (A.1). Our implementation of group-Lasso-type estimator comes from Yuan and Lin (2006), which is an extension of the shooting algorithm of Fu (1998). As a direct consequence of the Karush–Kuhn–

Tucker conditions, we have a necessary and sufficient condition for $\widehat{\beta}$ to be a solution of (A.1):

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

$$T^{-1}\|\sum_{t=1}^{T}(\Psi_t(Y_t-\Psi_t^\top\widehat{\beta}U_t)U_t^\top)_r\|\leqslant\lambda,\qquad \text{if } \widehat{\beta}_r=0. \tag{B.2}$$

Recall that $\Psi_t\Psi_t^\top=I_K$. It can be easily verified that the solution to (B.1) and (B.2) is

$$\widehat{\beta}_r=(1-\lambda/\|S_r\|)_+S_r, \tag{B.3}$$

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 (A.1) can therefore be obtained by applying (B.3) to $r=1,\ldots,R$ iteratively.

LEMMA B.1. *Consider model (2.2). Assume that $\Psi_t\Psi_t^\top=I_K$, $T^{-1}\sum_{t=1}^{T}U_t^\top U_t/R=1$, and $M(\beta^*)\leqslant s$. If the random event*

$$\mathcal{A}=\left(2T^{-1}\max_{1\leqslant r\leqslant R}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{j=1}^{J}\Psi_{tkj}^\top\varepsilon_{tj}U_{tr}\leqslant\lambda\right) \tag{B.4}$$

holds with high probability for some $\lambda>0$. Then, for any solution $\widehat{\beta}$ of problem (A.1) and $\forall\beta$, we have

$$\begin{aligned}&T^{-1}\sum_{t=1}^{T}\|\ \Psi_t^\top(\widehat{\beta}-\beta^*)U_t\ \|^2+\lambda\|\widehat{\beta}-\beta\|_{2,1}\\&\quad\leqslant T^{-1}\sum_{t=1}^{T}\|\ \Psi_t^\top(\beta-\beta^*)U_t\ \|^2+4\lambda\sum_{r\in\mathcal{R}(\beta)}\|\widehat{\beta}_r-\beta_r\|,\end{aligned} \tag{B.5}$$

$$T^{-1}\max_{1\leqslant r\leqslant R}\|\sum_{t=1}^{T}\left(\Psi_t\Psi_t^\top(\widehat{\beta}-\beta^*)U_tU_t^\top\right)_r\|\leqslant 3\lambda/2, \tag{B.6}$$

$$M(\widehat{\beta})\leqslant\frac{4\phi_{\max}^2}{\lambda^{-2}T^{-2}}\sum_{t=1}^{T}\|\ (\widehat{\beta}-\beta^*)U_t\ \|_2^2, \tag{B.7}$$

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

Proof: The proof involves similar thoughts given in Lemma 3.1 of Lounici et al. (2009). By the definition of $\widehat{\beta}$ as a minimizer of (A.1), $\forall\beta$ we have

$$\begin{aligned}&T^{-1}\sum_{t=1}^{T}\|\ \Psi_t^\top\widehat{\beta}U_t-Y_t\ \|^2+2\lambda\sum_{r=1}^{R}\|\widehat{\beta}_r\|\\&\quad\leqslant T^{-1}\sum_{t=1}^{T}\|\ \Psi_t^\top\beta U_t-Y_t\ \|^2+2\lambda\sum_{r=1}^{R}\|\beta_r\|,\end{aligned} \tag{B.8}$$

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

$$T^{-1}\sum_{t=1}^{T}\|\ \Psi_t^\top(\widehat{\beta}-\beta^*)U_t\ \|^2\leqslant T^{-1}\sum_{t=1}^{T}\|\ \Psi_t^\top(\beta-\beta^*)U_t\ \|^2$$

$$+2T^{-1}\sum_{t=1}^{T}\varepsilon_t^\top\Psi_t^\top(\widehat{\beta}-\beta)U_t+2\lambda\sum_{r=1}^{R}(\|\beta_r\|-\|\widehat{\beta}_r\|). \tag{B.9}$$

Using the Hölder inequality, we have

$$2T^{-1}\sum_{t=1}^{T}\varepsilon_t^\top\Psi_t^\top(\widehat{\beta}-\beta)U_t\leqslant 2T^{-1}\sum_{t=1}^{T}\|\Psi_t\varepsilon_tU_t^\top\|_{2,\infty}\|\widehat{\beta}-\beta\|_{2,1}, \tag{B.10}$$

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

If the random event

$$\mathcal{A}=\left(2T^{-1}\max_{1\leqslant r\leqslant R}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{j=1}^{J}\Psi_{tkj}^\top\varepsilon_{tj}U_{tr}\leqslant\lambda\right) \tag{B.11}$$

holds with high probability for some $\lambda>0$, which we specify afterwards, then it follows from (B.9) and (B.10), on the event $\mathcal{A}$, that

$$\begin{aligned}
&T^{-1}\sum_{t=1}^{T}\|\Psi_t^\top\left(\widehat{\beta}-\beta^*\right)U_t\|^2+\lambda\sum_{r=1}^{R}\|\widehat{\beta}_r-\beta_r\|\\
&\quad\leqslant T^{-1}\sum_{t=1}^{T}\|\Psi_t^\top(\beta-\beta^*)U_t\|^2+2\lambda\sum_{r=1}^{R}(\|\widehat{\beta}_r-\beta_r\|+\|\beta_r\|-\|\widehat{\beta}_r\|)\\
&\quad\leqslant T^{-1}\sum_{t=1}^{T}\|\Psi_t^\top(\beta-\beta^*)U_t\|^2+2\lambda\sum_{r\in\mathcal{R}(\beta)}(\|\widehat{\beta}_r-\beta_r\|+\|\beta_r\|-\|\widehat{\beta}_r\|)\\
&\qquad+2\lambda\sum_{r\in\mathcal{R}^c(\beta)}(\|\widehat{\beta}_r-\beta_r\|+\|\beta_r\|-\|\widehat{\beta}_r\|)\\
&\quad\leqslant T^{-1}\sum_{t=1}^{T}\|\Psi_t^\top(\beta-\beta^*)U_t\|^2+4\lambda\sum_{r\in\mathcal{R}(\beta)}\|\widehat{\beta}_r-\beta_r\|.
\end{aligned} \tag{B.12}$$

This proves (B.5).

To prove (B.4), we use (B.1) and (B.2), which yield the inequality

$$T^{-1}\max_{1\leqslant r\leqslant R}\left\|\sum_{t=1}^{T}\left(\Psi_t(Y_t-\Psi_t^\top\widehat{\beta}U_t)U_t^\top\right)_r\right\|\leqslant\lambda. \tag{B.13}$$

Then

$$\begin{aligned}
&T^{-1}\left\|\sum_{t=1}^{T}\left(\Psi_t\Psi_t^\top(\widehat{\beta}-\beta^*)U_tU_t^\top\right)_r\right\|\\
&\quad\leqslant T^{-1}\left\|\sum_{t=1}^{T}\left(\Psi_t(\Psi_t^\top\widehat{\beta}U_t-Y_t)U_t^\top\right)_r\right\|+T^{-1}\left\|\sum_{t=1}^{T}\left(\Psi_t\varepsilon_tU_t^\top\right)_r\right\|,
\end{aligned} \tag{B.14}$$

where we use $Y_t=\Psi_t^\top\beta^*U_t+\varepsilon_t$ and the triangle inequality. Then, the bound (B.4) follows by combining (B.14) with (B.13) and using the definition of the event $\mathcal{A}$.

Finally, we show (B.7). First, observe that

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

On the event $\mathcal{A}$, utilizing (B.1) and the triangle inequality, we have

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

The following arguments yield the bound (B.7) on the number of non-zero rows of $\widehat{\beta}_r^\top$:

$$\begin{aligned} M(\widehat{\beta}) &\leqslant \frac{4}{\lambda^2 T^2} \sum_{r \in \mathcal{R}(\widehat{\beta})} \left\| \sum_{t=1}^{T} \left(\Psi_t \Psi_t^\top (\widehat{\beta} - \beta^*) U_t U_t^\top\right)_r \right\|^2 \\ &\leqslant \frac{4}{\lambda^2 T^2} \sum_{r=1}^{R} \left\| \sum_{t=1}^{T} \left(\Psi_t \Psi_t^\top (\widehat{\beta} - \beta^*) U_t U_t^\top\right)_r \right\|^2 \\ &= \frac{4}{\lambda^2 T^2} \left\| \sum_{t=1}^{T} \left(\Psi_t \Psi_t^\top (\widehat{\beta} - \beta^*) U_t U_t^\top\right) \right\|_2^2 \\ &\leqslant \frac{4\phi_{\max}^2}{\lambda^2 T} \sum_{t=1}^{T} \left\| (\widehat{\beta} - \beta^*) U_t \right\|_2^2. \end{aligned}$$

Here, we use the fact that $\Psi_t \Psi_t^\top = I_K$ and $\phi_{\max}$ is the maximum eigenvalue of the matrix $\sum_{t=1}^{T} U_t U_t^\top / T$. □

Proof of Theorem 3.1: We proceed along the lines of Theorem 6.2 of Bickel et al. (2009) and Theorem 3.1 of Lounici et al. (2009). Let $\mathcal{R} = \mathcal{R}(\beta^*) = \{r : \beta_r^* \neq 0\}$.

Using inequality (B.5) in Lemma B.1 with $\beta = \beta^*$, on the event $\mathcal{A}$ defined in (3.1), we have

$$T^{-1} \sum_{t=1}^{T} \|\Psi_t^\top (\widehat{\beta} - \beta^*) U_t\|^2 \leqslant 4\lambda \sum_{r \in \mathcal{R}} \|\widehat{\beta}_r - \beta_r^*\| \leqslant 4\lambda \sqrt{s} \|(\widehat{\beta} - \beta^*)_{\mathcal{R}}\|. \tag{B.15}$$

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 $\Delta = (\widehat{\beta} - \beta^*)$,

$$\|(\widehat{\beta} - \beta^*)_{\mathcal{R}}\|^2 \leqslant \sum_{t=1}^{T} \|\Psi_t^\top (\widehat{\beta} - \beta^*) U_t\|^2 / (\kappa^2 T). \tag{B.16}$$

Now, $T^{-1} \sum_{t=1}^{T} \|\Psi_t^\top (\widehat{\beta} - \beta^*) U_t\|^2 \leqslant 16 s \lambda^2 \kappa^{-2}$ (3.2) follows from (B.15) and (B.16).

Inequality (3.3) follows by noting that

$$K^{-1/2} \sum_{r=1}^{R} \|\widehat{\beta}_r - \beta_r^*\| \leqslant 4K^{-1/2} \sum_{r \in \mathcal{R}} \|\widehat{\beta}_r - \beta_r^*\| \leqslant 4K^{-1/2} \sqrt{s} \|(\widehat{\beta} - \beta^*)_{\mathcal{R}}\| \leqslant 16 s \lambda \kappa^{-2} K^{-1/2}, \tag{B.17}$$

and then using (3.2). Inequality (3.4) follows from (B.7) and (3.2). □

Definition of $\boldsymbol{\beta}$-mixing: Following Doukhan (1994), let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{A}$ and $\mathcal{B}$ be two sub-σ algebras of $\mathcal{F}$. Various measures of dependence between $\mathcal{A}$ and $\mathcal{B}$ have been defined as

$$\boldsymbol{\beta}(\mathcal{A}, \mathcal{B}) = \sup \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |P(A_i \bigcap B_j) - P(A_i)P(B_j)|, \tag{B.18}$$

$$\boldsymbol{\alpha}(\mathcal{A}, \mathcal{B}) = \sup |P(A \bigcap B) - P(A)P(B)|, A \in \mathcal{A}, B \in \mathcal{B}, \tag{B.19}$$

where the supremum is taken over all pairs of (finite) partitions $\{A_1, \ldots, A_I\}$ and $\{B_1, \ldots, B_J\}$ of Ω such that $A_i \in \mathcal{A}$ for each i and $B_j \in \mathcal{B}$ for each j. Now suppose $\{V_t\}_{t\in\mathcal{T}}$ is a (not necessarily stationary) sequence of random variables. For $-\infty \leqslant i \leqslant j \leqslant \infty$, define the σ-field $\sigma_i^j = \sigma(V_t, i \leqslant t \leqslant j, t \in \mathcal{T})$. For each $a \geqslant 1$, define the following dependence coefficients:

$$\boldsymbol{\beta}(a) = \sup_{t\in\mathcal{T}} \boldsymbol{\beta}(\sigma_{-\infty}^t, \sigma_{t+a}^{\infty}), \qquad \boldsymbol{\alpha}(a) = \sup_{t\in\mathcal{T}} \boldsymbol{\alpha}(\sigma_{-\infty}^t, \sigma_{t+a}^{\infty}).$$

In the special case where the sequence $\{V_t\}_{t\in\mathcal{T}}$ is strictly stationary, they simply become

$$\boldsymbol{\beta}(a) = \boldsymbol{\beta}(\sigma_{-\infty}^t, \sigma_{t+a}^{\infty}), \qquad \boldsymbol{\alpha}(a) = \boldsymbol{\alpha}(\sigma_{-\infty}^t, \sigma_{t+a}^{\infty}).$$

A stochastic process is said to be $\boldsymbol{\beta}$-mixing (or $\boldsymbol{\alpha}$-mixing) if $\boldsymbol{\beta}(a) \to 0$ (or $\boldsymbol{\alpha}(a) \to 0$) as $a \to \infty$. By definition, when $\sigma_{-\infty}^t$ and σ_{t+a}^{∞} are independent of each other, $\boldsymbol{\beta}(a) = 0$; the closer $\boldsymbol{\beta}(a)$ gets to 0, the more independent the time series is.

Proof of Theorem 3.2: The proofs of this theorem are similar to those of Theorem 3.1 up to a specification of the bound on $P(\mathcal{A}^c)$ in Lemma B.1. Consider the event

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

Observe that

$$P(\mathcal{A}^c) \leqslant RP\left(\sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{j=1}^{J} \Psi_{tkj} \varepsilon_{tj} U_{tr} K^{-1/2} > 2^{-1} \lambda T K^{-1/2} \right)$$

$$\stackrel{\text{def}}{=} RP\left(\sum_{t=1}^{T} V_t > 2^{-1} \lambda T K^{-1/2} \right).$$

Because Assumption 3.2 holds, applying the Bernstein-type inequality for $\boldsymbol{\beta}$-mixing random variables $\{V_t\}_{t=1}^T$ (Theorem 4 of Doukhan, 1994, p. 36) yields that $\forall \varepsilon > 0$ and $\forall\, 0 < q \leqslant 1$,

$$P\left(\sum_{t=1}^{T} V_t \geqslant 2^{-1} \lambda T K^{-1/2} \right) \leqslant \underbrace{2 \exp\left(- \frac{(1-\varepsilon) 3 (1 + \varepsilon^2/4) \lambda^2 T K^{-1}}{4(6(1+\varepsilon^2/4)\sigma^2 + q C'' \lambda T K^{-1/2})} \right)}_{\stackrel{\text{def}}{=} T_1}$$

$$+ \underbrace{\frac{(1+\varepsilon^2/4) \boldsymbol{\beta}\left((qT\varepsilon^2/(4+\varepsilon^2)) - 1 \right)}{q}}_{\stackrel{\text{def}}{=} T_2}.$$

To make $T_1 \leqslant R^{-(1+\delta')}$, $\delta' > 0$ and $T_2 \leqslant R^{-(1+\delta')}$, we choose

$$\lambda = \sqrt{\frac{16 \log R^{1+\delta'} K\sigma^2}{T(1-\varepsilon)}}, \qquad qC''\lambda T K^{-1/2} = 6(1+\varepsilon^2/4)\sigma^2$$

and

$$\boldsymbol{\beta}\Big((qT\varepsilon^2/(4+\varepsilon^2)) - 1\Big) \leqslant qR^{-(1+\delta')}/(1+\varepsilon^2/4) = \Big(24\sigma(1-\varepsilon)^{1/2}(R^{1+\delta'}\sqrt{\log R^{1+\delta'} T}C'')^{-1}\Big),$$

with $qT\varepsilon^2/(4+\varepsilon^2) = (3/8)\sigma\varepsilon^2 T^{1/2}(1-\varepsilon)^{1/2}C''^{-1} \log R^{-(1+\delta')/2}$. Then, we have

$$P(\mathcal{A}^c) \leqslant RP\left(\sum_{t=1}^{T} V_t > \lambda T/K\right) \leqslant 3R^{-\delta'}.$$

□

Proof of Corollary 3.1: To prove this corollary, we need to show that Assumption 3.2 is satisfied, i.e. for an m-dependent process with order k, σ^2 in Assumption 3.2 is equal to $2k\sigma_0^2$. For simplicity, we assume that $n = 1$ and that m is divisible by $2k$. Then,

$$\begin{aligned} E\left[\sum_{i=1}^{m} V_i\right]^2 &= E\left[\sum_{i=1}^{k} V_i + \sum_{i=k+1}^{2k} V_i + \ldots + \sum_{i=m-k}^{m} V_i\right]^2 \\ &= E\left[\underbrace{\sum_{j=0}^{m/2k-1}\sum_{i=2jk+1}^{2jk+k} V_i}_{\stackrel{\text{def}}{=} C} + \underbrace{\sum_{j=0}^{m/2k-1}\sum_{i=2jk+k+1}^{2(j+1)k+k} V_i}_{\stackrel{\text{def}}{=} D}\right]^2 \\ &\leqslant 2E[C^2] + 2E[D^2]. \end{aligned}$$

Because for $j = 0, \ldots, m/2k-1$, $\sum_{i=2jk+1}^{2jk+k} V_i$ are independent of each other by the definition of V_t and the same argument holds for $\sum_{i=2jk+k+1}^{2(j+1)k+k} V_i$, we have

$$\begin{aligned} 2E[C^2] + 2E[D^2] &= 2\sum_{j=0}^{m/2k-1} E\left[\sum_{i=2jk+1}^{2jk+k} V_i\right]^2 + 2\sum_{j=0}^{m/2k-1} E\left[\sum_{i=2jk+k+1}^{2(j+1)k+k} V_i\right]^2 \\ &\leqslant m/kk^2\sigma_0^2 + m/kk^2\sigma_0^2 = 2mk\sigma_0^2. \end{aligned}$$

□

Proof of Theorem 3.3: 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, because

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

$$\widetilde{Z}_{0,t} - \widehat{Z}_{0,t} = \left(\widehat{A}\Psi_t\Psi_t^\top\widehat{A}^\top\right)^{-1}\widehat{A}\Psi_t\left(\Psi_t^\top(\widehat{\beta}-\beta^*)U_t\right),$$

Theorem 3.2 tells us that $T^{-1}\sum_{t=1}^{T}\parallel \Psi_t^\top(\widehat{\beta}-\beta^*)U_t \parallel^2$ is bounded by $\mathcal{O}(T^{-1})$. From the definitions of ρ^2 and δ_K, we know that the first term is dominated by the second term. □

Proof of Theorem 3.4: The proof shares ideas with Park et al. (2009). We prove the first equation of the theorem for $h \neq 0$. The second equation follows from the first. We start by proving 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, then 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

$$\begin{aligned}
&\|T^{-1}\sum_{t=1}^{T} Z_{0,t}\widehat{Z}_{0,t}^\top\widehat{A} - T^{-1}\sum_{t=1}^{T} Z_{0,t}Z_{0,t}^\top A^*\| \\
&\leqslant T^{-1}\sum_{t=1}^{T}\|Z_{0,t}(\widehat{Z}_{0,t}^\top\widehat{A} - Z_{0,t}^\top A^*)\| \\
&\leqslant (T^{-1}\sum_{t=1}^{T}\|Z_{0,t}\|^2)^{1/2}(T^{-1}\sum_{t=1}^{T}\|\widehat{Z}_{0,t}^\top\widehat{A} - Z_{0,t}^\top A^*\|^2)^{1/2} \\
&= \mathcal{O}_P(\rho+\delta_K), \qquad \text{(B.20)}
\end{aligned}$$

because of Assumption 3.3(e) and Theorem 3.3. 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 the construction that all space basis are independent.

Note that $\widetilde{Z}_{0,t} = B^\top\widehat{Z}_{0,t}$ and $\widetilde{A} = B^{-1}A$. With (B.20) this gives

$$\begin{aligned}
\|\widetilde{A} - A^*\| &= \|T^{-1}\sum_{t=1}^{T} Z_{0,t}Z_t^\top(\widetilde{A} - A^*)\|\mathcal{O}_P(1) \\
&= \|T^{-1}\sum_{t=1}^{T} Z_{0,t}\widetilde{Z}_{0,t}^\top\widetilde{A} - T^{-1}\sum_{t=1}^{T} Z_{0,t}Z_{0,t}^\top A^*\|\mathcal{O}_P(1) \\
&= \mathcal{O}_P(\rho+\delta_K). \qquad \text{(B.21)}
\end{aligned}$$

From Assumptions 3.3(d), (B.21) and Theorem 3.3, we obtain

$$T^{-1}\sum_{t=1}^{T}\|\widetilde{Z}_t^\top - Z_{0,t}\|^2 \qquad \text{(B.22)}$$

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

$$
\begin{aligned}
&+T^{-1}\sum_{t=1}^{T}\|\widetilde{Z}_t^\top\widetilde{A}-Z_{0,t}^\top A^*\|^2\mathcal{O}_P(1)+\mathcal{O}_P(\delta_K^2)\\
&\leqslant T^{-1}\sum_{t=1}^{T}\|\widetilde{Z}_{0,t}-Z_{0,t}\|^2\|\widetilde{A}-A^*\|^2\mathcal{O}_P(1)\\
&\quad+T^{-1}\sum_{t=1}^{T}\|Z_{0,t}\|^2\|\widetilde{A}-A^*\|^2\mathcal{O}_P(1)\\
&\quad+T^{-1}\sum_{t=1}^{T}\|\widetilde{Z}_t^\top\widetilde{A}-Z_{0,t}^\top A^*\|^2\mathcal{O}_P(1)+\mathcal{O}_P(\delta_K^2)\\
&=\mathcal{O}_P(\rho^2+\delta_K^2). \qquad \text{(B.23)}
\end{aligned}
$$

We show that for $h\neq 0$,

$$
T^{-1}\sum_{t=h+1}^{T}\Big((\widetilde{Z}_{0,t+h}-Z_{0,t+h})-(\widetilde{Z}_{0,t}-Z_{0,t})\Big)Z_{0,t}^\top=\mathcal{O}_P(T^{-1/2}). \qquad \text{(B.24)}
$$

This implies the first statement of Theorem 3.4 because by (B.23),

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

To prove (B.24), define

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

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

$$
\widetilde{Z}_{0,t}=\widetilde{S}_{t,Z}^{-1}J^{-1}\sum_{j=1}^{J}\Big(Y_{t,j}A\Psi(X_{t,j})\Big), \qquad \text{(B.25)}
$$

$$
\tilde{a}=\widetilde{S}_{\alpha}^{-1}(JT)^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J}\Big(\Psi(X_{t,j})\otimes\widetilde{Z}_{0,t}\Big)Y_{t,j}. \qquad \text{(B.26)}
$$

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

$$\sup_{1\le t\le T} \|\widetilde{S}_{t,Z} - S_{t,Z}\| = o_P(\gamma), \qquad \|\widetilde{S}_{\alpha} - S_{\alpha}\| = o_P(\gamma). \tag{B.27}$$

We show the first part of (B.27), and the second part can be obtained analogously. Because

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

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

$$J^{-1}\sum_{j=1}^{J} A^*\Big(\Psi(X_{t,j})\Psi(X_{t,j})^\top - E[\Psi(X_{t,j})\Psi(X_{t,j})^\top]\Big)(\widetilde{A} - A^*)^\top = o_P(\gamma), \tag{B.28}$$

$$J^{-1}\sum_{j=1}^{J} (\widetilde{A} - A^*)\Big(\Psi(X_{t,j})\Psi(X_{t,j})^\top - E[\Psi(X_{t,j})\Psi(X_{t,j})^\top]\Big)(\widetilde{A} - A^*)^\top = o_P(\gamma), \tag{B.29}$$

$$J^{-1}\sum_{j=1}^{J} A^*\Big(\Psi(X_{t,j})\Psi(X_{t,j})^\top - E[\Psi(X_{t,j})\Psi(X_{t,j})^\top]\Big)A^{*\top} = o_P(\gamma), \tag{B.30}$$

$$J^{-1}\sum_{j=1}^{J} A^* E[\Psi(X_{t,j})\Psi(X_{t,j})^\top](\widetilde{A} - A^*)^\top = o_P(\gamma), \tag{B.31}$$

$$J^{-1}\sum_{j=1}^{J} (\widetilde{A} - A^*)E[\Psi(X_{t,j})\Psi(X_{t,j})^T](\widetilde{A} - A^*)^\top = o_P(\gamma). \tag{B.32}$$

The proof of (B.28)–(B.30) follows by simple arguments. We now show (B.31). Claim (B.32) can be shown similarly. To prove (B.31), we use the Bernstein inequality for the following sum:

$$P\Big(|\sum_{j=1}^{J} W_j| > x\Big) \leqslant 2\exp\Big(-\frac{1}{2}\frac{x^2}{V + Mx/3}\Big). \tag{B.33}$$

For t with $1 \le t \le T$, the random variable W_j is an element of the $L \times 1$-matrix $S = J^{-1}A^*\Big(\Psi(X_{t,j})\Psi(X_{t,j})^\top e - E[\Psi(X_{tj})\Psi(X_{tj})^\top e]\Big)$, where $e \in \mathbb{R}^K$ with $\|e\| = 1$. In (B.33), 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. $|W_j| \le M$ for $1 \le j \le J$, a.s.). With some constants C_1 and C_2 that do not depend on t and the row number, we obtain $V \le C_1 J^{-1}$ and $M \le C_2 K^{1/2} J^{-1}$. The application of the Bernstein inequality gives that, uniformly for $1 \le t \le T$ and $e \in \mathbb{R}^K$ with $\|e\| = 1$, all L elements of S are of order $o_P(\gamma)$. This completes the proof of claim (B.28).

From (B.21), (B.25) and (B.25)–(B.27), it follows that uniformly for $1 \leqslant t \leqslant T$,

$$\begin{aligned}\widetilde{Z}_{0,t} - Z_{0,t} \;&=\; S_{t,Z}^{-1}J^{-1}\sum_{j=1}^{J}\varepsilon'_{t,j}A^*\Psi(X_{t,j}) \\ &\quad + S_{t,Z}^{-1}J^{-1}\sum_{j=1}^{J}\varepsilon'_{t,j}(\widetilde{A} - A^*)\Psi(X_{t,j}) + o_P(T^{-1/2}) \\ &\stackrel{\text{def}}{=} \Delta_{t,1,Z} + \Delta_{t,2,Z} + o_P(T^{-1/2}).\end{aligned} \tag{B.34}$$

To prove the theorem, it remains to show that for $1 \leqslant j \leqslant 2$,

$$T^{-1} \sum_{t=-h+1}^{T} (\Delta_{t+h,j,Z} - \Delta_{t,j,Z}) Z_{0,t}^{\top} = o_P(T^{-1/2}). \tag{B.35}$$

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

$$E[\|(JT)^{-1} \sum_{t=1}^{T} \sum_{j=1}^{J} \varepsilon'_{t,j} S_{t,Z}^{-1} \mathcal{M} \Psi(X_{t,j})\|^2] = \mathcal{O}(K(JT)^{-1}), \tag{B.36}$$

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

QUERIES

Journal: ECTJ
Paper: ectj12024

Dear Author

During the copy-editing of your paper, the following queries arose. Please respond to these by marking up your proofs with the necessary changes/additions. Please write your answers on the query sheet if there is insufficient space on the page proofs. Please write clearly and follow the conventions shown on the corrections sheet. If returning the proof by fax do not write too close to the paper's edge. Please remember that illegible mark-ups may delay publication.

Query No.	**Description**	**Remarks**
Q1	Author: Please note that the addresses have been given in full and that missing e-mail addresses have been added. Please confirm that these details are correct.	
Q2	Author: Some editorial/grammatical changes have been made, so please read through your paper carefully to check that the meaning has not been unintentionally altered.	
Q3	Author: Please check that the full meaning of the abbreviation 'IVS' used later in the paper has been correctly identified here.	
Q4	Author: The sentence 'For example, in international...' has been amended for clarity; please confirm that the meaning is correct.	
Q5	Author: Please confirm that inherent (rather than inheriting) was the intended meaning here.	
Q6	Author: Please note that figure and table captions need to be short.Where necessary, some details from the captions have been moved to the text. Please confirm that all figure and table captions are correct.	

Q7	Author: The sentence 'Figure 2 displays the...' has been amended for clarity; please confirm that the meaning is correct.	
Q8	Author: Please check that Appendix A is the correct appendix to cite here.	
Q9	Author: Please check the number '16', because in the original table caption the number '20' was used.	
Q10	Author: The abbreviation 'r.v.s' has been given in full as 'random variables' throughout the paper; please check that this is correct.	
Q11	Author: Can you provide any updated details for Stock and Watson (2005a)?	
Q12	Author: Please note that equation (B.16) has been amended slightly to avoid the use of a large \sqrt \dots ; please confirm that this is okay.	

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	••• under matter to remain	ⓛ
Insert in text the matter indicated in the margin	⅄	New matter followed by ⅄ or ⅄ⓧ²
Delete	/ through single character, rule or underline or ⊢⟷⊣ through all characters to be deleted	⅁ or ⅁ⓧ²
Substitute character or substitute part of one or more word(s)	/ through letter or ⊢⟷⊣ through characters	new character / or new characters /
Change to italics	— under matter to be changed	⌣
Change to capitals	≡ under matter to be changed	≡
Change to small capitals	= under matter to be changed	=
Change to bold type	~ under matter to be changed	~
Change to bold italic	≈ under matter to be changed	≈
Change to lower case	Encircle matter to be changed	⋕
Change italic to upright type	(As above)	⊬
Change bold to non-bold type	(As above)	⊬
Insert 'superior' character	/ through character or ⅄ where required	Y or Y under character e.g. Y^{2} or Y^{2}
Insert 'inferior' character	(As above)	⅄ over character e.g. $⅄_{2}$
Insert full stop	(As above)	⊙
Insert comma	(As above)	,
Insert single quotation marks	(As above)	⅄ or ⅄ and/or ⅄ or ⅄
Insert double quotation marks	(As above)	⅄ or ⅄ and/or ⅄ or ⅄
Insert hyphen	(As above)	⊢
Start new paragraph	⌐	⌐
No new paragraph	⌒	⌒
Transpose	⊔	⊔
Close up	linking ⁀ characters	⁀
Insert or substitute space between characters or words	/ through character or ⅄ where required	Y
Reduce space between characters or words	| between characters or words affected	↑

USING eANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

WILEY-BLACKWELL

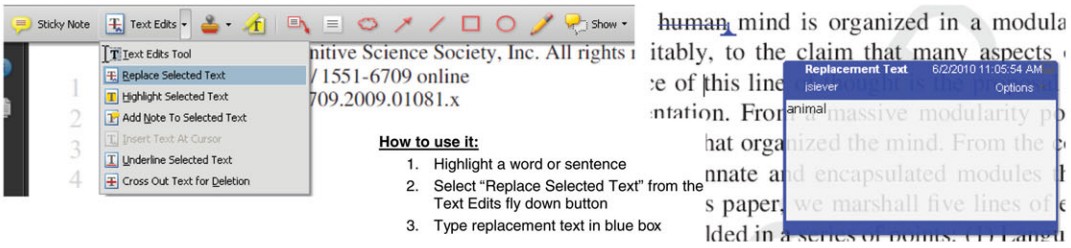
Required software to eAnnotate PDFs: Adobe Acrobat Professional or Acrobat Reader (version 8.0 or above). (Note that this document uses screenshots from Acrobat Reader 9. For screenshots from Acrobat Reader X, a separate document is available on the journal e-proofing site.)
The latest version of Acrobat Reader can be downloaded for free at: http://get.adobe.com/reader/

Once you have Acrobat Reader 8, or higher, open on your PC you should see the Commenting Toolbar:

****(If the above toolbar does not appear automatically go to *Tools>Comment & Markup>Show Comment & Markup Toolbar*)****

1. Replacement Text Tool — For replacing text.

Strikes a line through text and opens up a replacement text box.

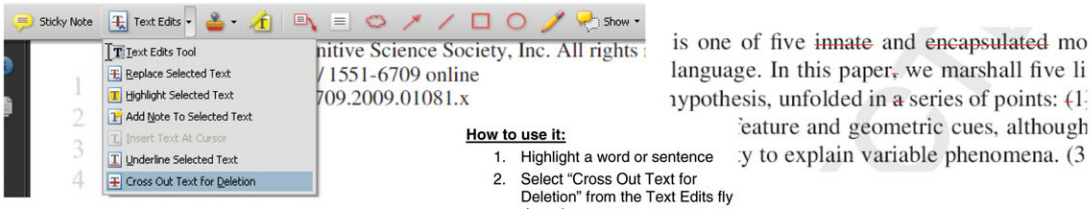


How to use it:

1. Highlight a word or sentence
2. Select "Replace Selected Text" from the Text Edits fly down button
3. Type replacement text in blue box

2. Cross-out Text Tool — For deleting text.

Strikes a red line through selected text.

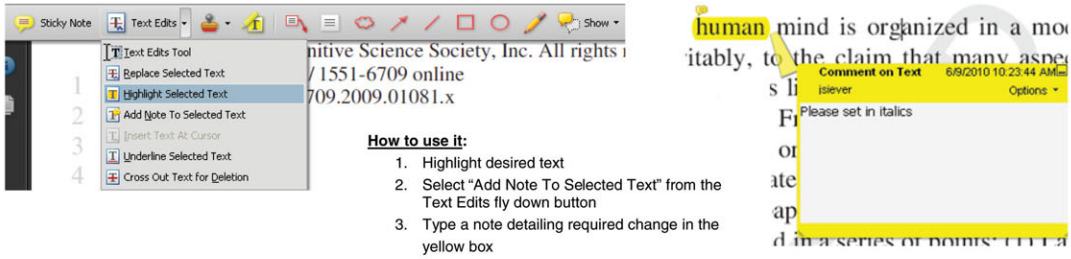


How to use it:

1. Highlight a word or sentence
2. Select "Cross Out Text for Deletion" from the Text Edits fly down button

3. Highlight Tool — For highlighting a selection to be changed to bold or italic.

Highlights text in yellow and opens up a text box.

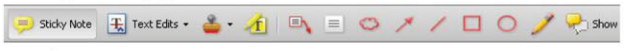


How to use it:

1. Highlight desired text
2. Select "Add Note To Selected Text" from the Text Edits fly down button
3. Type a note detailing required change in the yellow box

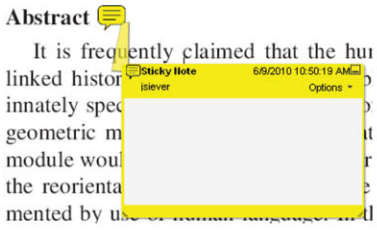
4. Note Tool — For making notes at specific points in the text

Marks a point on the paper where a note or question needs to be addressed.



How to use it:

1. Select the Sticky Note icon from the commenting toolbar
2. Click where the yellow speech bubble symbol needs to appear and a yellow text box will appear
3. Type comment into the yellow text box

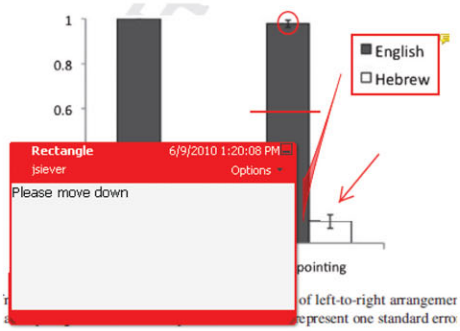


5. Drawing Markup Tools — For circling parts of figures or spaces that require changes

These tools allow you to draw circles, lines and comment on these marks.

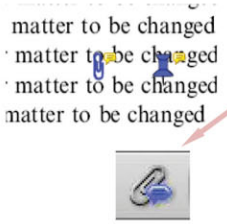
How to use it:

1. Click on one of shape icons in the Commenting Toolbar
2. Draw the selected shape with the cursor
3. Once finished, move the cursor over the shape until an arrowhead appears and double click
4. Type the details of the required change in the red box



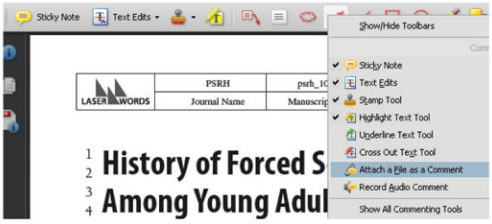
6. Attach File Tool — For inserting large amounts of text or replacement figures as a files.

Inserts symbol and speech bubble where a file has been inserted.

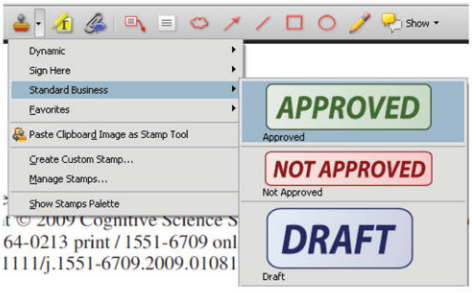


How to use it:

1. Right click on the Commenting Toolbar
2. Select "Attach a File as a Comment"
3. Click on paperclip icon that appears in the Commenting Toolbar
4. Click where you want to insert the attachment
5. Select the saved file from your PC or network
6. Select type of icon to appear (paperclip, graph, attachment or tag) and close

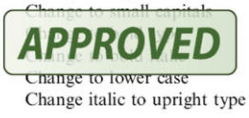


7. Approved Tool (Stamp) — For approving a proof if no corrections are required.



How to use it:

1. Click on the Stamp Tool in the toolbar
2. Select the Approved rubber stamp from the 'standard business' selection
3. Click on the text where you want to rubber stamp to appear (usually first page)



Change to lower case
Change italic to upright type

Help

For further information on how to annotate proofs click on the Help button to activate a list of instructions: