A New Panel Data Treatment for Heterogeneity in Time Trends*

Alois Kneip
Department of Economics
University of Bonn
Bonn, Germany

Robin C. Sickles†
Department of Economics
Rice University
Houston, Texas

Wonho Song‡
Department of Economics
Chung-Ang University
Seoul, Korea

February 9, 2011

*Earlier versions of this paper under the title "On Estimating the Mixed Effects Model" were presented at North American Productivity Workshop, Toronto, June 2004, COMPSTAT 2004, Prague, August 2004, and the European Workshop of Efficiency and Productivity IX, Brussels, July 2005. The paper was also given at the 2005 Econometric Society World Congress, University College, London and Michigan State, Rice University, and Syracuse University Econometrics Workshops. The authors thank participants at those conferences and workshops, particularly Peter Schmidt, Mahmoud El-Gamal, Yoosoon Chang, Joon Park, and Leopold Simar for constructive criticisms and insights. We would also like to thank three anonymous referees and Editor Guido Kuersteiner for their insightful comments and criticisms that substantially strenghened the paper. We would also like to thank Levent Kutlu for his very crucial research assistance. The usual caveat applies.

†Corresponding author: Department of Economics - MS 22, Rice University, 6100 S. Main Street, Houston, TX 77005-1892, Tel: 713-348-3322, Fax: 713-348-5278, E-mail: rsickles@rice.edu.

‡Song gratefully acknowledges financial support from Chung-Ang University through Research Grants in 2011.

Running Title: Panel Heterogeneity in Time Trends

Corresponding Author:

Robin C. Sickles

Department of Economics MS-22

Rice University

6100 South Main Street

Houston, Texas 77005

rsickles@rice.edu

Abstract

Our paper introduces a new estimation method for arbitrary temporal heterogeneity in panel data models. The paper provides a semiparametric method for estimating general patterns of cross-sectional specific time trends. The methods proposed in the paper are related to principal component analysis and estimate the time-varying trend effects using a small number of common functions calculated from the data. An important application for the new estimator is in the estimation of time-varying technical efficiency considered in the stochastic frontier literature. Finite sample performance of the estimators is examined via Monte Carlo simulations. We apply our methods to the analysis of productivity trends in the U.S. banking industry.

JEL Classification: C13, C14, C22, C23, D24, G21.

Key words and phrases: Time trends, panel models, principal component analysis, smoothing splines, banking efficiency.

1 Introduction

Substantial research interest has focused on controlling for unobserved heterogeneity in panel models. Work by Park and Simar and Park, Sickles, and Simar (1994, 1998, 2003, 2007), and Sickles (2005) has focused on semi-parametric efficient panel data estimators for the standard fixed and random effects models with various specifications, including autoregressive errors and dynamic models. As the specifications of unobserved heterogeneity become more and more general, in particular allowing for temporal variation in the unobserved effects, and as trend stationarity of individual cross-sections comes under closer scrutiny, the proper specification of time effects becomes no less important than the specification of a difference or trend stationary time series (Nelson and Plosser, 1982; Maddala and Kim, 1998; Kao and Chiang, 2000; Baltagi, Egger, and Pfaffermayr, 2003; Mark and Sul, 2003, Chang, 2004).

In this paper, we extend the random and fixed effects model in such a way that we do not impose any explicit restrictions on the temporal pattern of individual effects. They are considered as random functions of time, representing a sample of smooth individual time trends. Detailed modelling and analysis of the general structure of these trends are the central points of our methodology. This goal is particularly important in our application to stochastic frontier analysis, where individual effects allow to access time-varying technical efficiencies of banks in the U. S. banking system.

The basic qualitative assumption is a fairly smooth, slowly varying local behavior of trends, although they may have pronounced effects on temporal patterns on the long-run. We formalize this idea and show that our model can be used for virtually any smooth pattern of temporal and cross-sectional changes in unobserved heterogeneity (time trends) and allows for the possibility that parameter heterogeneity is due to variables other than the constant term. This generality is accomplished by approximating the effect terms nonparametrically. The approach is based on a factor model, where time-varying individual effects are represented by linear combinations of a small number of unknown basis functions, with coefficients varying across cross-sectional units. Fixed effects, basis functions and corresponding coefficients are estimated from the data using methods related to principal component analysis coupled with smoothing spline techniques. Asymptotic distributions of the new estimators are derived, and rank tests are applied to determine the dimensionality of the factor model. Furthermore, goodness-of-fit tests of pre-specified parametric models are elaborated. Simulation experiments indicate that in finite samples our method works much better than other well known time-varying effects estimators. As an illustration, the effects are interpreted in the context of a stochastic frontier production function (Aigner, Lovell, and Schmidt, 1977) and our method is applied to the analysis of time-varying technical efficiency in the U.S. banking industry.

Factor models related to our setup have already been extensively studied in the econometric literature. Among others, important contributions are given by the work of Forni and Lippi (1997), Forni and Reichlin (1998), Stock and Watson (2002), Forni et al. (2000), Bai and Ng (2002), Bernanke and Boivin (2003) or Ahn, Lee, and Schmidt (2005). Bai (2003, 2009) provides a general inferential theory. Our approach fully integrates the panel and factor models. It allows us to simultaneously estimate fixed effects, common factors (basis functions), and individual factor scores under a wide variety of conditions, including the possible existence of dynamic effects and/or correlations between individual effects and explanatory variables. Different from existing work, the asymptotic theory also covers situations where dynamic effects follow non-stationary time series models, as for example random walks.

Another related branch of research is given by the statistical literature on "functional data analysis" which deals with the analysis of multiple smooth curves. For an overview one may consult the book by Ramsay and Silverman (1997). Explicit factor models and corresponding inferential results based on "functional principal component analysis" are given, for example, by Kneip (1994), Ferré (1995), or Kneip and Utikal (2001) for different applications. An essential feature of our approach, taken from this literature, is the use of nonparametric smoothing techniques as an inherent part of the estimation procedure. The asymptotic theory of Section 3.2 indicates that econometric factor models in other contexts may also benefit from incorporating smoothing procedures, since compared to standard results one may then achieve dramatically improved rates of convergence when estimating common factors.

The remainder of the paper is organized as follows. The basic setup is described in Section 2. Section 3 introduces our new estimator for arbitrary time-varying effects, derives its asymptotic distribution, and provides other analytical results for optimal choice for the number of principal components and smoothing parameters. The finite sample performance of our new estimator is evaluated using Monte Carlo simulations in section 4. In Section 5 we use the new estimator to analyze the technical efficiency of banks in the U. S. banking system. Concluding remarks follow in Section 6. The mathematical proofs are collected in the Appendix.

2 Model

2.1 Basic Setup

Panel studies in econometrics provide data from a sample of individual units where each unit is observed repeatedly over time . Econometric analysis then aims to model the variation of some response variable Y in dependence of a vector of explanatory variables $X \in \mathbb{R}^p$.

We will assume panel data based on a balanced design with T equally spaced repeated measurements per individual. The resulting observations of n individuals can then be represented in the form (Y_{it}, X_{it}), $t = 1, \dots T$, $i = 1, \dots, n$, where the index i denotes individual units (e.g. firms, households, etc.) and the index t denotes time periods. We consider the model

$$Y_{it} = \beta_0(t) + \sum_{j=1}^{p} \beta_j X_{itj} + v_i(t) + \epsilon_{it}, \quad i = 1, \dots, n, t = 1, \dots, T, \tag{1}$$

where $\beta_0(t)$ denotes a general average function, and $v_i(t)$ are non-constant individual effects. In order to ensure identifiability we require that $\sum_{i=1}^{n} v_i(t) = 0$ for all t. We are mainly interested in analyzing β and $v_i(t)$. The influence of $\beta_0(t)$ can be eliminated by using centered variables $Y_{it} - \bar{Y}_t$, $X_{ijt} - \bar{X}_{tj}$, where $\bar{Y}_t = \frac{1}{n}\sum_{i=1}^{n} Y_{it}$ and $\bar{X}_{tj} = \frac{1}{n}\sum_{i=1}^{n} X_{itj}$. Then

$$Y_{it} - \bar{Y}_t = \sum_{j=1}^{p} \beta_j (X_{itj} - \bar{X}_{tj}) + v_i(t) + \epsilon_{it} - \bar{\epsilon}_i, \quad i = 1, \dots, n, t = 1, \dots, T, \tag{2}$$

with $\bar{\epsilon}_t = \frac{1}{n}\sum_{i=1}^{n} \epsilon_{it}$. Note that after having estimated β and $v_i(t)$, the average function $\beta_0(t)$ may be estimated in a final step of our analysis (see Section 3).

In this approach "individual effects" $v_i(t)$ necessarily play a more important role than in textbook panel data models. Identifiability of (1) requires that all variables X_{itj}, $j = 1, \dots, p$ possess considerable variation over t. All individual differences are captured by $v_i(t)$, and this includes the effects of additional variables, like e.g. socioeconomic attributes, which characterize individuals but do not change over time. For example, suppose that there are q additional explanatory variables $X_{i,p+1}, \dots, X_{i,p+q}$ which do not change over time. The traditional framework then leads to the model $Y_{it} = \beta_0 + \sum_{j=1}^{p} \beta_j X_{itj} + \sum_{j=p+1}^{p+q} \beta_j X_{ij} + \tau_i + \epsilon_{it}$ with constant individual coefficients τ_i. In model (1), $v_i(t)$ then is a constant function with $v_i(t) \equiv \sum_{j=p+1}^{p+q} \beta_j X_{ij} + \tau_i$.

Our focus lies on estimating and analyzing $v_i(t)$, $t = 1, \dots, T$. This is of course motivated by our application in the field of stochastic frontier analysis, where individual effects determine technical efficiencies and are the main quantity of interest. We will additionally rely on the following structural assumption:

Assumption 1. *For some fixed $L \in \{0, 1, 2, \dots\}$, $L < T$, there exists an L-dimensional subspace $\mathcal{L}_T$ of $\mathbb{R}^T$ such that $v_i \in \mathcal{L}_T$ holds with probability 1.*

The space $\mathcal{L}_T$ as well as its dimension L are unknown. But the assumption implies that v_i can be parametrized in terms of suitable basis functions (common factors) $g_1, \dots, g_L$ with $\mathcal{L}_T := span\{g_1, \dots, g_L\}$ and corresponding individual coefficients:

$$v_i(t) = \sum_{r=1}^{L} \theta_{ir} g_r(t). \tag{3}$$

The centered model (2) can then be rewritten in the form

$$Y_{it} - \bar{Y}_t = \sum_{j=1}^{p} \beta_j (X_{itj} - \bar{X}_{tj}) + \sum_{r=1}^{L} \theta_{ir} g_r(t) + \epsilon_{it} - \bar{\epsilon}_t, \quad i = 1, \dots, n, t = 1, \dots, T \tag{4}$$

Our approach consists in using the data in order to *estimate* L as well as basis functions $g_1, \dots, g_L$ and corresponding coefficients θ_{ir}.

Parametric mixed effects models of the form (4) are widely used in applications and assume that individual effects can be modeled by linear combinations of *pre-specified* basis function (e.g. polynomials). For example, in the context of production frontier analysis Cornwell, Schmidt, and Sickles (1990) assume that the v_i can be modeled by quadratic polynomials. In our notation, then $L = 3$ and $\mathcal{L}_T$ is the space of all quadratic polynomials. Obviously, our approach is much more flexible and avoids misspecifications by using the data to determine the structure of basis functions.

Indeed, it does not seem to be very restrictive to require that (3) holds for *some* L. Formally, (3) is always fulfilled if for all sufficiently large n, T the empirical covariance matrix $\Sigma_{n,T}$ of the vectors $(v_i(1), \dots, v_i(T))'$, $i = 1, \dots, n$, possesses rank L. This corresponds to the setup of factor models as considered by Bai (2003, 2009) or Ahn et al. (2005). Recall, however, that different from the cited papers our focus lies upon analyzing non-stationary, smooth time trends.

There are several advantages of (3) compared to a completely nonparametric analysis of $v_1, \dots, v_n$. An important point is more efficient estimation. The basis functions $g_1, \dots, g_L$ represent a common functional structure and can thus be determined by combining information from *all* individual curves. They can thus be estimated with much faster rates of convergence than an individual v_i. Under some additional assumptions, the coefficients θ_{ri} are then obtained with the same rate of convergence as if $g_1, \ldots, g_L$ were known. Roughly speaking, (3) dramatically improves accuracy of estimates and allows to determine $v_1, \ldots, v_n$ with parametric rates of convergence.

Furthermore, (3) is well-suited for economic interpretation and further econometric analysis. By $g_1, \ldots, g_L$ we denote general functional components whose structure provide information about the *common* functional structure of *all* individual $v_1, \ldots, v_n$. There may exist a substantial interpretation in terms of general economic developments.

All *differences* between individuals are captured by the coefficients θ_{ir}. For example, a standard panel model as discussed above leads to $L = 1$, $g_1(t) \equiv 1$, and $\theta_{i1} = \sum_{j=p+1}^{p+q} \beta_j X_{ij} + \tau_i$. When having estimated θ_{i1}, estimates of $\beta_{p+1}, \ldots, \beta_{p+q}$ can then be obtained from a linear regression of θ_{i1} on $X_{i,p+1}, \ldots, X_{i,p+q}$. This generalizes to more interesting situations with $L \geq 1$ and non-constant functions $g_r(t)$. Effects of socioeconomic or demographic variables which do not change over time may be quantified by regressing the scores θ_{ir} on $X_{i,p+1}, \ldots, X_{i,p+q}$. In many applications such regressions will constitute an important step in econometric analysis.

2.2 Identifiability and standardization

An intrinsic problem of factor models is non-uniqueness of common factors. Given some basis $g_1, \ldots, g_L$, for every regular $L \times L$ matrix A the linear transformation

$$(\mathbf{g}_1(t), \ldots, \mathbf{g}_L(t))' := A \cdot (g_1(t), \ldots, g_L(t))', \ (\vartheta_{1i}, \ldots, \vartheta_{Li})' := A^{-1} \cdot (\theta_{1i}, \ldots, \theta_{Li})', \qquad (5)$$

leads to a parametrization with alternative basis functions and coefficients such that

$$v_i(t) = \sum_{r=1}^{L} \theta_{ir} g_r(t) = \sum_{r=1}^{L} \vartheta_{ir} \mathbf{g}_r(t)$$

holds with probability 1, and $\mathcal{L}_T := span\{g_1, \ldots, g_L\} = span\{\mathbf{g}_1, \ldots, \mathbf{g}_L\}$. Only $\mathcal{L}_T$ is uniquely determined but not a particular basis. If for example $L = 2$ and $\mathcal{L}_T$ is the space of all linear functions, then two equivalent parameterization are given by $v_i(t) = \vartheta_{i1}(2t - 5) + \vartheta_{i2}(t + 5)$ and $v_i(t) = \theta_{i1} + \theta_{i2} t$, where $\theta_{i1} = 5(\vartheta_{i2} - \vartheta_{i1})$ and $\theta_{i2} = 2\vartheta_{i1} + \vartheta_{i2}$.

Any underlying, "generic" basis is thus only identifiable up to linear transformations of the form (5). In order to specify a well-defined estimation problem we will rely on the following standardization which identifies a suitable parametric representation out of the equivalence class given by (5):

(a) $\frac{1}{n}\sum_{i=1}^{n}\theta_{i1}^2 \geq \frac{1}{n}\sum_{i=1}^{n}\theta_{i2}^2 \geq \cdots \geq \frac{1}{n}\sum_{i=1}^{n}\theta_{iL}^2 > 0$.

(b) $\frac{1}{n}\sum_{i=1}^{n}\theta_{ir} = 0$ and $\frac{1}{n}\sum_{i=1}^{n}\theta_{ir}\theta_{is} = 0$ for all $r, s \in \{1, \ldots, L\}$, $r \neq s$.

(c) $\frac{1}{T}\sum_{t=1}^{T} g_r(t)^2 = 1$ and $\sum_{t=1}^{T} g_r(t)g_s(t) = 0$ for all $r, s \in \{1, \ldots, L\}$, $r \neq s$.

Provided that $n > L$, $T > L$, Conditions (a) - (c) do not impose any restriction, and they introduce a suitable normalization which ensures identifiability of the components up to sign changes (instead of θ_{ir}, g_r one may also use $-\theta_{ir}, -g_r$). Note that (a) - (c) lead to orthogonal vectors g_r as well as empirically uncorrelated coefficients θ_{ir}. [1]

In a textbook panel model we have $L = 1$ and $\mathcal{L}_T$ is the space of all constant functions. Our normalization then leads to $g_1(t) \equiv 1$. The model by Battese and Coelli (1992) corresponds to $L = 1$ and $g_1(t) = \exp(-\eta(t-T))/\sqrt{\frac{1}{T}\sum_{s=1}^{T}\exp(-\eta(s-T))^2}$. For $L > 1$, the specific structure of g_r will usually depend on n and T ($g_r \equiv g_{r;n,T}$). But the real objects of interest are the structure of the factor space $\mathcal{L}_T$ and the distribution of v_i within $\mathcal{L}_T$. If there exists a "true" basis $\mathbf{g}_1, \ldots, \mathbf{g}_L$ generating v_i, then it will necessarily be connected with $g_1, \ldots, g_L$ by a linear transformation (5) for some (unidentifiable) matrix A, and there is a unique space $\mathcal{L}_T = \{v | v = \sum_{r=1}^{L}\theta_r g_r,\ \theta_1, \ldots, \theta_L \in \mathbb{R}\} = \{v | v = \sum_{r=1}^{L}\vartheta_r \mathbf{g}_r,\ \vartheta_1, \ldots, \vartheta_L \in \mathbb{R}\}$ (as e.g. a linear space, a space of quadratic polynomials, etc.). Relation (5) also implies that there exists a corresponding one-to-one relation between the coefficients ϑ_{ir} and θ_{ir} for *any possible realization* v_i, and the distribution of $(\theta_{1i}, \ldots, \theta_{iL})$ reflects all aspects of the distribution of $v_i(t)$. In this sense conditions (a) - (c) define a specific set of orthogonal basis functions which can be estimated with a particularly high degree of accuracy (see Subsection 3.3). Of course, suitable rotations of estimated g_r may be applied in subsequent analysis.

Our estimation procedure will be based on the fact that under the above normalization $g_1, g_2, \ldots$ are to be obtained as principal components of the sample $v_1 = (v_1(1), \ldots, v_1(T))', \ldots, v_n = (v_n(1), \ldots, v_n(T))'$. More precisely, let

$$\Sigma_{n,T} = \frac{1}{n}\sum_{i=1}^{n} v_i {v_i}' \tag{6}$$

denote the empirical covariance matrix of $v_1, \ldots, v_n$ (recall that $\sum_{i=1}^{n} v_i = 0$). We use $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_T$ as well as $\gamma_1, \gamma_2, \ldots, \gamma_T$ to denote the resulting eigenvalues and orthonormal eigenvectors of $\Sigma_{n,T}$. Some simple algebra [compare, e.g., with Rao (1958)] then shows that

$$g_r(t) = \sqrt{T} \cdot \gamma_{rt} \quad \text{for all} \quad r = 1, \ldots, L,\ t = 1, \ldots, T, \tag{7}$$

$$\theta_{ir} = \frac{1}{T}\sum_{t=1}^{T} v_i(t) g_r(t) \quad \text{for all} \quad r = 1, 2, \ldots, L,\ i = 1, \ldots, n, \tag{8}$$

$$\lambda_r = \frac{T}{n}\sum_{i=1}^{n} \theta_{ir}^2 \quad \text{for all} \quad r = 1, 2, \ldots, L \tag{9}$$

Furthermore, for all $l = 1, 2, \ldots$

$$\sum_{r=l+1}^{T} \lambda_r = \frac{1}{n}\sum_{i,t}(v_i(t) - \sum_{r=1}^{l} \theta_{ir} g_r(t))^2 = \frac{1}{n}\min_{\tilde{g}_1,\ldots,\tilde{g}_l}\sum_{i=1}^{n}\min_{\vartheta_{i1},\ldots,\vartheta_{il}}\sum_{t=1}^{T}(v_i(t) - \sum_{r=1}^{l}\vartheta_{ir}\tilde{g}_r(t))^2 \tag{10}$$

One can infer from relation (10) that $v_i \approx \sum_{r=1}^{l} \theta_{ir} g_r(t)$ provides the best possible approximation of the functions v_i in terms of an l-dimensional linear model. If $n > L$, $T > L$, Model (3) holds for some dimension L if and only if $rank(\Sigma_{n,T}) = L$.

Let us consider the structure of possible spaces $\mathcal{L}_T$ more closely. In the context of mixed effects models $\mathcal{L}_T$ will be a deterministic space of smooth functions. As discussed above, examples are spaces of linear functions or quadratic polynomials. The population analogue of $\Sigma_{n,T}$ is then the covariance matrix Σ_T with $\Sigma_T = \lim_{n\to\infty} \Sigma_{n,T}$ a.s, and for large n the function g_r will be close to the corresponding principal component of Σ_T.

A basic motivation of our paper is to develop a method which is capable to deal with any smooth pattern of temporal changes in individual effects. However, from a time series point of view "smooth" trends are often described by discrete time stochastic processes. In this case, basis functions are generated by an underlying random mechanism, and consequently $\mathcal{L}_T$ is a **random space**. For example, let us study the case that all v_i are generated by linear combinations of L independent random walks. More precisely, suppose that

$$\mathcal{L}_T = span\{\mathbf{g}_1, \ldots, \mathbf{g}_L\}, \quad \text{where } \mathbf{g}_r(t+1) = \mathbf{g}_r(t) + \delta_{r,t}, \quad r = 1, \ldots, L \tag{11}$$

for some fixed $\mathbf{g}_1(1), \ldots, \mathbf{g}_L(1)$ and i.i.d. innovations $\delta_{r,1}, \delta_{r,2}, \ldots$ with $\mathbf{E}(\delta_{r,t}) = 0$, $var(\delta_{r,t}) = \sigma^2_{\delta,r}$. Moreover, $\delta_{r,t}$ is independent of $\delta_{s,t}$ for $r \neq s$. The structure of $\mathcal{L}_T$ then depends on the realizations of $\delta_{r,t}$ and thus is random. The particular basis $\mathbf{g}_1, \ldots, \mathbf{g}_L$ will of course not correspond to $g_1, \ldots, g_L$, but recall that necessarily $\mathcal{L}_T = span\{g_1, \ldots, g_L\}$ if n, T are sufficiently large.

By definition, $v \in \mathcal{L}_T$ means that there are parameters $\vartheta_1, \ldots, \vartheta_L$ such that $v(t) = \sum_{r=1}^{L} \vartheta_r \mathbf{g}_r(t) = v(t-1) + \sum_{r=1}^{L} \vartheta_r \delta_{r,t}$. Each v in $\mathcal{L}_T$ is thus a random walk with independent innovations $\delta_v = \sum_{r=1}^{L} \vartheta_r \delta_{r,t}$. This of course carries over to our sample functions

$v_i = \sum_{r=1}^{L} \vartheta_{ir}\mathbf{g}_r(t)$. We assume that each statistical unit of the population possesses an individual, fixed set of coefficients.

3 Estimation and theoretical results

3.1 Estimation

In practice, $v_1, \dots, v_n$ are unknown and all components of model (4) thus have to be estimated from the data. The idea of our estimation procedure can be described as follows: Recall that individual effects are supposed to represent "smooth" trends. The first step of our procedure relies on the use of an auxiliary functional variable ν_i defined on the interval $[1, T]$ which interpolates the T different values of v_i. Estimates $\hat{\beta}$ and functional approximations $\hat{\nu}_i$ are determined by least squares, where smoothness of $\hat{\nu}_i$ is controlled by a roughness penalty. Then an estimate $\hat{v}_i(t)$ of $v_i(t)$ is defined as $\hat{v}_i(t) := \hat{\nu}_i(t)$, $t = 1, \dots, T$. This corresponds to a penalized least squares approach similar to methods proposed, for example, by Engle et al. (1986), Speckman (1988), or Härdle et al. (2000). Two further steps of our procedure then provide estimates $\hat{g}_r$ and $\hat{\theta}_{ir}$ of the components of the factor decomposition. It will be shown in Section 3.2 that $\sum_{r=1}^{L} \hat{\theta}_{ir}\hat{g}_r$ provide much more efficient estimates of v_i than the purely nonparametric estimates $\hat{v}_i$.

Let us first introduce some additional notations. Let $\bar{Y}_t = \frac{1}{n}\sum_{i=1}^{n} Y_{it}$, $\bar{Y} = (\bar{Y}_1, \dots, \bar{Y}_T)'$, $Y_i = (Y_{i1} \dots, Y_{iT})'$ and $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{iT})$. Furthermore, let $X_{ij} = (X_{i1j}, \dots, X_{iTj})'$, $\bar{X}_{tj} = \frac{1}{n}\sum_{i=1}^{n} X_{itj}$, and $\bar{X}_j = (\bar{X}_{1j}, \dots, \bar{X}_{Tj})'$. We will use X_i and $\bar{X}$ to denote the $T \times p$ matrices with elements X_{itj} and $\bar{X}_{tj}$.

Step 1: For a preselected smoothing parameter $\kappa > 0$ determine estimates $\hat{\beta}_1, \dots, \hat{\beta}_p$ and functional approximations $\hat{\nu}_1, \dots, \hat{\nu}_n$ by minimizing

$$\sum_{i=1}^{n} \frac{1}{T} \sum_{t} \left(Y_{it} - \bar{Y}_t - \sum_{j=1}^{p} \beta_j (X_{itj} - \bar{X}_{tj}) - \nu_i(t) \right)^2 + \sum_{i=1}^{n} \kappa \frac{1}{T} \int_1^T (\nu_i^{(m)}(s))^2 ds \qquad (12)$$

over all possible values of β and all m-times continuously differentiable functions $\nu_1, \dots, \nu_n$ on $[1, T]$. Then estimate $v_i(t)$ by $\hat{v}_i(t) := \hat{\nu}_i(t)$, $t = 1, \dots, T$, $i = 1, \dots, n$. Here, $\nu_i^{(m)}$ denotes the m-th derivative of ν_i.

Spline theory implies that any solution $\hat{\nu}_i$, $i = 1, \dots, n$, of (12) possess an expansion $\hat{\nu}_i(t) = \sum_{j=1}^{T} \hat{\zeta}_{ji} z_j(t)$ in terms of a natural spline basis $z_1, \dots, z_T$ of order $2m$ (for a discussion of natural splines and definitions of possible basis functions see, for example, Eubank, 1988). In practice, one will often choose $m = 2$ which leads to cubic smoothing splines.

If Z and A denote $T \times T$ matrices with elements $\{z_j(t)\}_{j,t=1,\dots,T}$ and $\{\int_1^T z_j^{(m)}(s) z_k^{(m)}(s) ds\}_{j,k=1,\dots,T}$, the above minimization problem can be reformulated in matrix notation: Determine $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)'$ and $\hat{\zeta}_i = (\hat{\zeta}_{1i}, \dots, \hat{\zeta}_{Ti})'$ by minimizing

$$\sum_{i=1}^{n} \left(\|Y_i - \bar{Y} - (X_i - \bar{X})\beta - Z\zeta_i\|^2 + \kappa \zeta_i' A \zeta_i \right), \tag{13}$$

where $\|\cdot\|$ denotes the usual Euclidean norm in $\mathbb{R}^T$, $\|a\| = \sqrt{a'a}$.

Note that Z is a regular $T \times T$ matrix. It is then easily seen that with

$$\mathcal{Z}_\kappa = Z(Z'Z + \kappa A)^{-1} Z' = \left(I - \kappa (Z')^{-1} A Z^{-1} \right)^{-1}$$

the solutions are given by

$$\hat{\beta} = \left(\sum_{i=1}^{n} (X_i - \bar{X})'(I - \mathcal{Z}_\kappa)(X_i - \bar{X}) \right)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})'(I - \mathcal{Z}_\kappa)(Y_i - \bar{Y}) \tag{14}$$

as well as

$$\hat{\zeta}_i = (Z'Z + \kappa A)^{-1} Z'(Y_i - \bar{Y} - (X_i - \bar{X})\hat{\beta}).$$

Therefore,

$$\hat{v}_i = Z\hat{\zeta}_i = \mathcal{Z}_\kappa (Y_i - \bar{Y} - (X_i - \bar{X})\hat{\beta}) \tag{15}$$

estimates $v_i = (v_i(1), \dots, v_i(T))'$.

Note that $\mathcal{Z}_\kappa$ is a positive semi-definite, symmetric matrix. All eigenvalues of $\mathcal{Z}_\kappa$ take values between 0 and 1. Moreover, $tr(\mathcal{Z}_\kappa^2) \leq tr(\mathcal{Z}_\kappa) \leq T$.

Step 2: Determine the empirical covariance matrix $\hat{\Sigma}_{n,T}$ of $\hat{v}_1 = (\hat{v}_1(1), \hat{v}_1(2), \dots, \hat{v}_1(T))', \dots, \hat{v}_n = (\hat{v}_n(1), \hat{v}_n(2), \dots, \hat{v}_n(T))'$ by

$$\hat{\Sigma}_{n,T} = \frac{1}{n} \sum_{i=1}^{n} \hat{v}_i \hat{v}_i'$$

and calculate its eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \hat{\lambda}_T$ and the corresponding eigenvectors $\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_T$.

Step 3: Set $\hat{g}_r(t) = \sqrt{T} \cdot \hat{\gamma}_{rt}$, $r = 1, 2, \dots, L$, $t = 1, \dots, T$, where $\hat{\gamma}_{rt}$ is the $t-$th element of the eigenvector $\hat{\gamma}_r$. For all $i = 1, \dots, n$ then determine $\hat{\theta}_{1i}, \dots, \hat{\theta}_{Li}$ by minimizing

$$\sum_{t=1}^{T} \left(Y_{it} - \bar{Y}_t - \sum_{j=1}^{p} \hat{\beta}_j (X_{itj} - \bar{X}_{tj}) - \sum_{r=1}^{L} \theta_{ri} \hat{g}_r(t) \right)^2 \tag{16}$$

with respect to $\theta_{1i}, \dots, \theta_{Li}$. [2]

Remarks:

1) In spite of the use of an auxiliary functional variable in Step 1 of our procedure, the required "smoothness" of $v_i(t)$ has to be interpreted in a very general sense. In Section 3.2 we will show that the estimators adopt fast rates of convergence if all v_i are *sufficiently smooth* in the sense that $\frac{1}{T}\sum_{t=2}^{T-1}(v_i(t-1) - 2v_i(t) + v_i(t+1))^2$ is small compared to $\frac{1}{T}\sum_{t=1}^{T} v_i(t)^2$.

2) An obvious problem is the choice of κ. A possible approach based on cross-validation will be discussed at the end of Subsection 3.2.

3.2 Asymptotic Theory

We now consider properties of our estimators. It is assumed that individual units are drawn by independent random sampling from the underlying population. We then analyze the asymptotic behavior as $n, T \to \infty$. We do not impose any condition on the magnitude of the quotient T/n. Our analysis will be based on the use of cubic smoothing splines with $m = 2$. We will require that Assumption 1 holds with a fixed dimension L for all n, T.

The following additional assumptions now provide the basis of our theoretical analysis. We will write $\lambda_{min}(A)$ and $\lambda_{max}(A)$ to denote the minimal and maximal eigenvalues of a symmetric matrix A, and g_r will be used to represent the vector $(g_r(1), \ldots, g_r(T))'$.

Assumption 2. *There exists a* ***nondecreasing*** *function $c(T)$ of T such that for all $r, s = 1, \ldots, L$, $r \neq s$*

- $\mathbf{E}(\frac{1}{T}\sum_{t=1}^{T} v_i(t)^2) = O(c(T))$,
- $\frac{1}{n}\sum_{i=1}^{n}\theta_{ir}^2 = O_P(c(T))$, $\quad \frac{1}{n}\sum_{i=1}^{n}\theta_{ir}^4 = O_P(c(T)^2)$,
- $c(T) = O_P(\frac{1}{n}\sum_{i=1}^{n}\theta_{ir}^2)$, $\quad c(T) = O_P(|\frac{1}{n}\sum_{i=1}^{n}\theta_{ir}^2 - \frac{1}{n}\sum_{i=1}^{n}\theta_{is}^2|)$

Note that by (9) and (10) we have $\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} v_i(t)^2 = \sum_{r=1}^{L}\frac{1}{n}\sum_{i=1}^{n}\theta_{ir}^2$, as well as

$$\sum_{r=l}^{L}\frac{\lambda_r}{T} = \sum_{r=l}^{L}\frac{1}{n}\sum_{i=1^n}\theta_{ir}^2 = \frac{1}{nT}\min_{\tilde{g}_1,\ldots,\tilde{g}_l}\sum_{i=1}^{n}\min_{\vartheta_{i1},\ldots,\vartheta_{il}}\sum_{t=1}^{T}(v_i(t) - \sum_{r=1}^{l}\vartheta_{ir}\tilde{g}_r(t))^2$$

for all $l = 1, \ldots, L-1$. By requiring that $\frac{1}{n}\sum_{i=1}^{n}\theta_{ir}^2 = O_P(c(T))$ as well as $c(T) = O_P(\frac{1}{n}\sum_{i=1}^{n}\theta_{ir}^2)$ we assume that each component $\frac{1}{n}\sum_{i=1}^{n}\theta_{ir}^2$ increases *exactly with rate* $c(T)$. This is obviously equivalent to saying that for any $l < L$ the error in approximating v_i by *the best possible* model with only l components increases exactly with rate $c(T)$. Constants have to be different, for example $\sum_{i=1}^{n}\theta_{i1}^2$ may be equal to $c(T)/2$, $\sum_{i=1}^{n}\theta_{i2}^2$ to $c(T)/10$, etc.

Assumption 3. *There exists a **nonincreasing** function $b(T)$ of T such that as $n, T \to \infty$ the second order differences of $v_i(t)$ satisfy*

$$\mathbf{E}\left(\frac{1}{T}\sum_{t=2}^{T-1}(v_i(t-1) - 2v_i(t) + v_i(t+1))^2\right) = O(b(T)) \tag{17}$$

By Proposition 1 in the appendix the value of $b(T)$ determines the bias of a smoothing spline estimator for all values of κ and may serve as a measure of smoothness. An even more interesting quantity is $b(T)/c(T)$. It will be shown in Theorem 1 that the smaller $b(T)/c(T)$ the faster the corresponding rate of convergence for suitable choice of κ. Of course, by Assumption 1) smoothness of v_i reflects the degree of smoothness of the underlying basis functions.

In order to clarify the impact of the above assumption, let us study some illustrative scenarios.

Example 1: Traditional smoothness. We first consider the typical setup of nonparametric mixed effects models where $\mathcal{L}_T$ is a deterministic space generated by smooth, at least twice continuously differentiable basis functions. The corresponding asymptotics can be formalized by assuming that there are i.i.d. non-zero random functions $\nu_1, \ldots, \nu_n$ on $L^2[0,1]$ such that $\nu_i(\frac{t}{T}) = v_i(t)$ for $t = 1, \ldots, T$. Then $c(T) = 1$. The functions $\nu_1, \ldots, \nu_n$ are a.s. twice continuously differentiable with $\mathbf{E}(\int_0^1 \nu_i''(t)^2 dt) < \infty$ and $0 < \mathbf{E}(\int_0^1 \nu_i(t)^2 dt) < \infty$. By Taylor expansions we obtain $\nu_i(\frac{t}{T}) - 2\nu_i(\frac{t-2}{T}) + \nu_i(\frac{t-4}{T}) = \frac{1}{T^2}\nu_i''(t) + o_P(\frac{1}{T^2})$ and therefore $\mathbf{E}\left(\frac{1}{T}\sum_{t=5}^{T}(v_i(t) - 2v_i(t-2) + v_i(t-4))^2\right) = \frac{1}{T^4}\mathbf{E}(\int_0^1 \nu_i''(t)^2 dt) + o(\frac{1}{T^4})$.
The relevant quantities in Assumptions 2 and 3 thus amount to

$$c(T) = 1, \quad b(T) = b(T)/c(T) = 1/T^4. \tag{18}$$

In this context it is of course also possible to deal with less smooth situations. If the ν_i are only piecewise smooth, possessing a finite number of discontinuities, then $b(T) = b(T)/c(T) = 1/T$.

Example 2: Random walks. Recall the discussion in Section 2.2 and assume that v_i are generated by a linear combination of L independent random walks as defined by (11). Then $\mathbf{E}(\frac{1}{T}\sum_{t=1}^{T} v_i(t)^2) = O(T)$. However, the mean squared *second differences* of random walks remain bounded as $T \to \infty$. Therefore, in this situation we can assume that Assumption 2 and 3 holds with

$$c(T) = T, \quad b(T) = 1, \quad b(T)/c(T) = 1/T \tag{19}$$

Note that if $\mathcal{L}_T = span\{\mathbf{g}_1, \dots, \mathbf{g}_L\}$, where $\mathbf{g}_1, \dots, \mathbf{g}_L$ are $I(2)$ processes whose *first differences* are random walks, then $\mathbf{E}(\frac{1}{T}\sum_{t=1}^{T} v_i(t)^2) = O(T^2)$, while the mean squared *second differences* still remain bounded. Then $c(T) = T^2, b(T) = 1, b(T)/c(T) = 1/T^2$.

We also want to emphasize that our approach is also able to deal with non-$I(q)$ processes. Let $\mathcal{L}_T = span\{\mathbf{g}_1, \dots, \mathbf{g}_L\}$ with $\mathbf{g}_r(t) = \sqrt{|\mathbf{g}_r(1)^2 + \delta_{r,1} + \delta_{r,2} + \dots + \delta_{r,t}|}$, where $\mathbf{g}_r(1)$ and $\delta_{r,t}$ satisfy the same conditions as in the above random walk example. The stochastic trend induced by this process cannot be eliminated by differencing, since for any $q = 1, 2, \dots$ the q-th order differences of r_t are *not* stationary. On the other hand, the resulting $v_i(t)$ are reasonably smooth. It is easily checked that then Assumptions 2) and 3) hold with $c(T) = T^{1/2}, b(T) = T^{-1/2}, b(T)/c(T) = 1/T$.

Two final assumptions now concern the behavior of $X_{it,j}$ and of the error term.

Assumption 4. *There exists a nondecreasing function $d(T) \leq c(T)$ of T with $d(T) = o(T)$ such that as $n, T \to \infty$ $\mathbf{E}(\frac{1}{T}\sum_{t=1}^{T} X_{it,j}^2) = O(d(T))$ holds for all $j = 1, \dots, p$ as $n, T \to \infty$. Furthermore, there is a constant $C_0 < \infty$ such that for all $\kappa \geq 1$*

$$\mathbf{E}\left(\lambda_{max}\left([\sum_{i=1}^{n}(X_i - \bar{X})'(I - \mathcal{Z}_\kappa)(X_i - \bar{X})]^{-1}\right)\right) \leq C_0 \frac{1}{nT}, \tag{20}$$

and there exists a constant $C_1 < \infty$ such that for all $j = 1, \dots, p$ and all vectors $a \in \mathbb{R}^T$

$$a'(I - \mathcal{Z}_\kappa) \cdot \mathbf{E}\left((X_{ij} - \bar{X}_j)(X_{ij} - \bar{X}_j)'|\ \mathcal{L}_T\right)(I - \mathcal{Z}_\kappa)a \leq C_1 \cdot \|(I - \mathcal{Z}_\kappa)a\|^2. \tag{21}$$

holds with probability 1 for all sufficiently large n, T.

If $\mathcal{L}_T$ is a deterministic space, then of course $\mathbf{E}(Z|\mathcal{L}_T) = \mathbf{E}(Z)$ for any random variable Z.

Assumption 5. *The error terms ϵ_{it} are i.i.d. with $\mathbf{E}(\epsilon_{it}) = 0$, $var(\epsilon_{it}) = \sigma^2 > 0$, and $\mathbf{E}(\epsilon_{it}^8) < \infty$. Moreover, ϵ_{it} is independent from $v_i(s)$ and $X_{is,k}$ for all t, s, j.*

Although, as shown above, our approach is able to cope with trends which do not fit into the usual $I(q)$ framework, some of our assumptions are restrictive from a time series point of view. Apart from assuming i.i.d. errors in Assumption 5, Assumption 4 contains regularity conditions which impose restrictions on the design matrix. It is essentially required that the time paths $\{X_{itj} - \bar{X}_{ij}\}_t$ are "less smooth" than those of $\{v_i(t)\}_t$. In particular, stationary processes generate non-smooth time paths. Note, however, that some interesting cases, as for example cointegration between Y and X, are excluded. We believe that more general results can be obtained, but part of the methodology may have to be adapted to the specific situation.

However, Assumption 4 does not impose any strong restriction when dealing with stationary processes X_{it} satisfying $d(T) = 1$. To illustrate the point, consider the simplest case $p = 1$ and assume that $X_{it} = \tilde{X}_{it} + \delta_i$, where $\{\tilde{X}_{it}\}_t$ are independent realizations of a zero mean $ARMA(q_1, q_2)$ process and δ_i are independent, zero mean random variables with variance Δ^2. Then

$$\mathbf{E}\left((X_i - \bar{X})(X_i - \bar{X})'\right) = (1 - \frac{1}{n})\Gamma + \Delta^2 \cdot \mathbf{1}\mathbf{1}',$$

where Γ is the autocovariance matrix of the underlying $ARMA(q_1, q_2)$ process. Since $p = 1$ we have $\mathbf{E}[(X_i - \bar{X})'(I - \mathcal{Z}_\kappa)(X_i - \bar{X})]) = \mathbf{E}[\lambda_{max}((X_i - \bar{X})'(I - \mathcal{Z}_\kappa)(X_i - \bar{X}))])$, and it is easily checked that this term is proportional to T for all $\kappa > 1$. Relation (20) is an immediate consequence. Moreover,

$$\mathbf{E}\left((X_i - \bar{X})(X_i - \bar{X})'|\ \mathcal{L}_T\right) = (1 - \frac{1}{n})\Gamma_{|\mathcal{L}_T} + \mathbf{E}(\delta_i^2|\ \mathcal{L}_T) \cdot \mathbf{1}\mathbf{1}',$$

where $\Gamma_{|\mathcal{L}_T}$ denotes the corresponding conditional autocovariance matrix given $\mathcal{L}_T$. Since by construction of $\mathcal{Z}_\kappa$, $\mathcal{Z}_\kappa \mathbf{1} = \mathbf{1}$ for $\mathbf{1} = (1, 1, \ldots, 1)'$, we arrive at

$$a'(I - \mathcal{Z}_\kappa)\mathbf{E}\left((X_{ij} - \bar{X})(X_{ij} - \bar{X})'|\ \mathcal{L}_T\right)(I - \mathcal{Z}_\kappa)a = a'(I - \mathcal{Z}_\kappa)\Gamma_{|\mathcal{L}_T}(I - \mathcal{Z}_\kappa)a.$$

For any stationary $ARMA(q_1, q_2)$ the maximal eigenvalue of Γ remains bounded as $T \to \infty$, and hence (21) is necessarily fulfilled for deterministic spaces $\mathcal{L}_T$ with $\Gamma_{|\mathcal{L}_T} = \Gamma$. If $\mathcal{L}_T$ is generated by stochastic processes, then the structure of the $ARMA(q_1, q_2)$-process characterizing the explanatory variable may be correlated with these processes, but (21) will remain true if $\lambda_{max}(\Gamma_{|\mathcal{L}_T})$ remains stochastically bounded, which does not seem to be a very strong condition.

Our estimator can be seen as a generalization of the LSDV estimator in a standard panel model. Let us focus on the simple situation that $L = 1$, $v_i(t) = \theta_{i1}$ and that there is some correlation between θ_{i1} and X_i. In dynamic panel models as well as some nonlinear models it is well-known that the LSDV estimator of the coefficients β then suffers from an incidental parameter bias (see e.g. Hahn and Newey, 2004, for a possible approach to bias reduction in nonlinear models). However, for a linear panel model with exogeneous regressors and i.i.d. error terms (as considered in our paper) the LSDV estimator $\hat{\beta}$ of β *is unbiased* even in the presence of correlations. In this context an incidental parameter problem only exists for the estimates $\hat{\theta}_{i1}$ of the individual effects $v_i(t) \equiv \theta_{i1}$ (Baltagi, 2001). These estimates $\hat{\theta}_{i1}$ are inconsistent unless $T \to \infty$.

Our results reflect this situation. If a standard panel model with correlated θ_{i1} and X_i our β-estimates remain unbiased, but by Theorem 1 (c) consistency of $\hat{\theta}_{i1}$ requires that $T \to \infty$. Before stating our main theorem we have to introduce some additional notation. Let $\mathcal{S}_T$ denote the linear space of all vectors $\tilde{a} \in \mathbb{R}^T$ which are straight lines, i.e. $\tilde{a}_t = \alpha_0 + \alpha_1 t$ for some $\alpha_0, \alpha_1 \in \mathbb{R}$, $t = 1, \ldots, T$, and let $\mathcal{S}_T^* \subset \mathbb{R}^T$ be the linear space orthogonal to $\mathcal{S}_T$. Any vector $a \in \mathbb{R}^T$ can be written in the form $a = \tilde{a} + a^*$, where $\tilde{a} \in \mathcal{S}_T$ and $a^* \in \mathcal{S}_T^*$ is the nonlinear part of a orthogonal to straight lines. Consequently, $v_i = \tilde{v}_i + v_i^*$ and $X_{ij} = \tilde{X}_{ij} + X_{ij}^*$ can be decomposed into linear and nonlinear parts with $\tilde{v}, \tilde{X}_{ij} \in \mathcal{S}_T$ and $v_i^*, X_{ij}^* \in \mathcal{S}_T^*$.

We will say that v_i and X_i are **"uncorrelated up to linear components" (ulc-uncorrelated)** if $\mathbf{E}(v_i^* v_l^* | X^*, \mathcal{L}_T) = \mathbf{E}(v_i^* v_l^* | \mathcal{L}_T)$ holds for all $i, l \in \{1, \ldots, n\}$, where $X^* = (X_{itj}^*)_{i,t,j}$.

We want to emphasize that v_i and X_i are necessarily ulc-uncorrelated in a standard panel model with constant individual effects and q additional explanatory variables $X_{i,p+1}, \ldots, X_{i,p+q}$ which do not change over time. Then $v_i \equiv \tilde{v}_i$ for the constant function $\tilde{v}_i(t) \equiv \sum_{j=p+1}^{p+q} \beta_j (X_{ij} - \bar{X}_j) + \tau_i - \bar{\tau}$, and hence $v_i^* \equiv 0$ does not depend at all on X.

By Assumption 4 we necessarily have $X^* \neq 0$. The bias of our parameter estimators $\hat{\beta}$ will depend on whether or not v_i and X_i are ulc-correlated. In order to provide some intuition note that a basic property of spline estimators is the fact that for *any straight line* $\tilde{a}$ we have $\mathcal{Z}_\kappa \tilde{a} = \tilde{a}$ and $(I - \mathcal{Z}_\kappa)\tilde{a} = 0$ for all values of κ. When considering the structure of our estimator $\hat{\beta}$ given by (14) it is now easily seen that all linear parts $\tilde{X}_{ij}$ and $\tilde{v}_i$ cancel out and do not at all influence $\hat{\beta}$. Therefore, only correlation between the nonlinear parts v_i^* and X_{ij}^* can create an additional bias.

We will use "$\mathbf{E}_\epsilon$" to denote conditional expectation given v_i and X_i, $i = 1, \ldots n$. Moreover, $\tilde{X}_i = X_i - \bar{X}$. Additionally note that eigenvectors are only unique up to sign changes. In the following we will always assume that the right "versions" are used. This will go without saying.

Recall that we consider theoretical behavior of our estimators as $n, T \to \infty$. Sensible smoothing parameters have to depend on n, T .[3] We will require that parameters $\kappa \equiv \kappa_{n,T} > 0$ are of an appropriate order of magnitude such that $\kappa b(T) \to 0$ as well as $\kappa^{1/4}/T \to 0$ as $n, T \to \infty$.

Theorem 1. Under the above assumptions we obtain as $n, T \to \infty$

(a) $\|\beta - \mathbf{E}_\epsilon(\hat\beta)\| = O_P(\sqrt{b_\beta(n,T,\kappa)})$, where

$$b_\beta(n,T,\kappa) := \begin{cases} \frac{\max\{1,\kappa\}b(T)}{Tn} & \text{if } X_i \text{ and } v_i \text{ are ulc-uncorrelated,} \\ \\ \frac{\max\{1,\kappa\}b(T)}{T} & \text{else,} \end{cases}$$

and $V_{n,T}^{-1/2}(\hat\beta - \mathbf{E}_\epsilon(\hat\beta)) \sim \mathbf{N}(0, I)$, where

$$V_{n,T} = \sigma^2 \left(\sum_{i=1}^n \tilde X_i'(I - \mathcal{Z}_\kappa)\tilde X_i\right)^{-1} \left(\sum_{i=1}^n \tilde X_i'(I - \mathcal{Z}_\kappa)^2 \tilde X_i\right) \left(\sum_{i=1}^n \tilde X_i'(I - \mathcal{Z}_\kappa)\tilde X_i\right)^{-1} = O_P\left(\frac{1}{nT}\right).$$

Therefore, $\|\beta - \hat\beta\| = |\beta - \mathbf{E}_\epsilon(\hat\beta) - (\hat\beta - \mathbf{E}_\epsilon(\hat\beta))\| = O_P(\sqrt{b_\beta(n,T,\kappa) + 1/(nT)})$.

(b) For all $r = 1, \ldots, L$

$$T^{-1/2}\|g_r - \hat g_r\| = O_P\left(\sqrt{\frac{\boldsymbol{\kappa} b(T) + d(T) b_\beta(n,T,\kappa)}{c(T)}} + \sqrt{\frac{1}{nc(T)\max\{1,\kappa^{1/4}\}}} + \frac{1}{T^2 c(T)^2}\right),$$

where $\boldsymbol{\kappa} = \min\{\kappa, \kappa^2\}$.

(c) For all $r = 1, \ldots, L$

$$|\hat\theta_{ri} - \theta_{ri}| = O_P\left(\sqrt{T^{-1} + \boldsymbol{\kappa} b(T) + d(T) b_\beta(n,T,\kappa) + (n\max\{1,\kappa^{1/4}\})^{-1}}\right).$$

Furthermore, if $\boldsymbol{\kappa} b(T) + d(T) b_\beta(n,T,\kappa) + (n\max\{1,\kappa^{1/4}\})^{-1} = o(T^{-1})$, then

$$\sqrt{T}(\hat\theta_{1i} - \theta_{1i}, \ldots, \hat\theta_{Li} - \theta_{Li})' \to_d \mathbf{N}(0, \sigma^2 I), \quad i = 1, \ldots, n. \tag{22}$$

(d)

$$\frac{\sum_{i=1}^n \|v_i - \sum_{r=1}^L \hat\theta_{ri}\hat g_r\|}{\sum_{i=1}^n \|v_i\|} = O_P\left(\sqrt{\frac{T^{-1} + \boldsymbol{\kappa} b(T) + d(T) b_\beta(n,T,\kappa) + (n\max\{1,\kappa^{1/4}\})^{-1}}{c(T)}}\right).$$

(e) If additionally $\frac{T}{n\max\{1,\kappa^{1/4}\}} \to 0$ as well as $Td(T)b_\beta(n,T,\kappa) + \frac{d(T)}{n} + \frac{1}{Tc(T)} = o\left(\frac{T}{n\max\{1,\kappa^{1/4}\}}\right)$, then

$$\frac{n\sum_{r=L+1}^T \hat\lambda_r \; - (n-1)\sigma^2 \cdot tr(\mathcal{Z}_\kappa \hat{\mathcal{P}}_L \mathcal{Z}_\kappa)}{\sigma^2\sqrt{2n \cdot tr((\mathcal{Z}_\kappa \hat{\mathcal{P}}_L \mathcal{Z}_\kappa)^2)}} \to_d \mathbf{N}(0,1), \tag{23}$$

$$\frac{n \cdot tr(\mathcal{P}_L \hat\Sigma_{n,T}) - (n-1)\sigma^2 \cdot tr(\mathcal{Z}_\kappa \mathcal{P}_L \mathcal{Z}_\kappa)}{\sigma^2\sqrt{2n \cdot tr((\mathcal{Z}_\kappa \mathcal{P}_L \mathcal{Z}_\kappa)^2)}} \to_d \mathbf{N}(0,1), \tag{24}$$

where $\hat{\mathcal{P}}_L = I - \frac{1}{T}\sum_{r=1}^L \hat g_r \hat g_r'$, and $\mathcal{P}_L$ is the projection matrix projecting into the $n - L$ dimensional linear space orthogonal to $span\{\mathcal{Z}_\kappa g_1, \ldots, \mathcal{Z}_\kappa g_L\}$.

A proof of the theorem can be found in the appendix. Obviously, convergence rates depend on the values of $c(T)$, $b(T)$ and $d(t)$. As an illustration let us evaluate the rates to be obtained for the two examples discussed above.

Example 1: Traditional smoothness (continued). If $d(T) = 1$, then by (18) optimal smoothing parameters $\kappa \equiv \kappa_{n,T}$ for estimating the functional components g_r have to *increase* with the sample size. More precisely, $\kappa \sim (nT)^{-4/5} \cdot T^4$ if $n = o(T^4)$ and $T = o(n^4)$, which leads to $\kappa b(T) \sim (nT)^{-4/5}$. Then necessarily $b_\beta(n, T, \kappa) = o(1/\sqrt{nT})$, and the theorem implies that

$$T^{-1/2}\|g_r - \hat{g}_r\| = O_P((nT)^{-2/5}), \quad \frac{\sum_{i=1}^n \|v_i - \sum_{r=1}^L \hat{\theta}_{ri}\hat{g}_r\|}{\sum_{i=1}^n \|v_i\|} = O_P(T^{-1/2})$$

$$V_{n,T}^{-1/2}(\hat{\beta} - \beta) \sim \mathbf{N}(0, I), \text{ and } \sqrt{T}(\hat{\theta}_{1i} - \theta_{1i}, \ldots, \hat{\theta}_{Li} - \theta_{Li})' \rightarrow_d \mathbf{N}(0, \sigma^2 I).$$

It is immediately seen (22) implies that $\hat{\theta}_{ri}$ is estimated as efficiently as in a parametric model with known functions g_r. We want to emphasize that $\kappa \sim (nT)^{-4/5} \cdot T^4$ corresponds to an *undersmoothing* of individual functions. The optimal smoothing parameter for spline estimation of an individual function v_i is of order $\kappa_{ind} \sim T^{-4/5} \cdot T^4$ which results in the usual nonparametric rate of convergence $\sum_{i=1}^n \|v_i - \hat{v}_i\|/(\sum_{i=1}^n \|v_i\|) = O_P(T^{-2/5})$. Based on our factor model it is thus possible to estimate the functions v_i with a parametric rate of convergence $T^{-1/2}$ instead of the nonparametric rate $T^{-2/5}$ characterizing a completely nonparametric approach.

Example 2: Random walks (continued). In addition to (19) assume that, as for example in the case of $ARMA(p, q)$-processes, X_{it} satisfies Assumption 4 with $d(T) = 1$. Suitable smoothing parameters $\kappa \equiv \kappa_{n,T}$ for estimating the functional components g_r have to *decrease* with the sample size. With $\kappa \sim (nT)^{-1/2}$ we have $\kappa b(T) = (nT)^{-1/2}$. The bias in estimating β is of order $\kappa b_\beta(n, T) = O(1/\sqrt{nT})$ if X_i and v_i are ulc-uncorrelated, and $\kappa b_\beta(n, T) = O(1/\sqrt{T})$ else. It will thus not be negligible compared to the standard error. The additional requirements ensuring the distributional results in Theorem 1c) hold if v_i and X_i are ulc-uncorrelated, while 1e) additionally requires that $n > T$. Furthermore,

$$T^{-1/2}\|g_r - \hat{g}_r\| = O_P(T^{-3/2} + (nT)^{-1/2}), \frac{\sum_{i=1}^n \|v_i - \sum_{r=1}^L \hat{\theta}_{ri}\hat{g}_r\|}{\sum_{i=1}^n \|v_i\|} = O_P(T^{-3/2} + (nT)^{-1/2}),$$

which shows that the relative error in approximating v_i by $\sum_{r=1}^L \hat{\theta}_{ri}\hat{g}_r$ is even smaller than in the case of traditional smoothness. Note that when approximating v_i by nonparametric estimates $\hat{v}_i$, then variance of the estimator does not decrease with n, and the convergence rate deteriorates to $\frac{\sum_{i=1}^n \|v_i - \hat{v}_i\|}{\sum_{i=1}^n \|v_i\|} = O_P(T^{-1/2})$.

Remarks:

1) The question arises whether it is possible to determine the best smoothing parameter for estimating $g_1, g_2, \ldots$ directly from the data. A straightforward approach consists in a "leave-one-individual-out" cross-validation. For a fixed L and $i = 1, \ldots, n$ let $\hat{\beta}_{-i}$ and $\hat{g}_{r,-i}$ denote the respective estimates of β and g_r obtained from the data (Y_{kj}, X_{kj}), $k = 1, \ldots, i-1, i+1, \ldots, n$, $j = 1, \ldots, T$, and let $\hat{\theta}_{r,-i}$ denote the corresponding estimates of θ_{ri} to be obtained when using $\hat{\beta}_{-i}$, $\hat{g}_{r,-i}$ instead of $\hat{\beta}$, $\hat{g}_r$ in Step 3 of our estimation procedure. All these estimates depend on κ, and one may approximate an optimal smoothing parameter by minimizing

$$CV(\kappa) := \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(Y_{it} - \bar{Y}_t - (X_{it} - \bar{X}_t)\hat{\beta}_{-i} - \sum_{r=1}^{L}\hat{\theta}_{r,-i}\hat{g}_{r,-i}(t))^2$$

over κ. Note that by (4) and by the independence of $\hat{\beta}_{-i}, \hat{g}_{r,-i}$ from ϵ_{it}

$$\begin{aligned}\mathbf{E}_\epsilon[CV(\kappa)] = {} & \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left((X_{it} - \bar{X}_t)(\beta - \hat{\beta}_{-i}) + v_i - \sum_{r=1}^{L}\hat{\theta}_{r,-i}\hat{g}_{r,-i}(t)\right)^2 + \frac{(n-1)(T-L)}{nT}\sigma^2 \\ & + O_P\left(\frac{1}{n}[\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}((X_{it} - \bar{X}_t)(\beta - \hat{\beta}_{-i}) + v_i - \sum_{r=1}^{L}\hat{\theta}_{r,-i}\hat{g}_{r,-i}(t))^2]^{1/2}\right)\end{aligned}$$

holds for all κ. It therefore seems to be reasonable to expect that this approach "in tendency" selects a κ providing a small mean squared error between true and estimated model. Cross-validation techniques are standard practice in nonparametric regression, but even in the random walk example discussed above it will provide smoothing parameters of the right order of magnitude. Due to bias any sequence $\kappa \equiv \kappa_{n,T}$ with $\kappa \to \infty$ as $n, T \to$ will lead to $P(CV(\kappa) > C) \to 1$ for any constant $C > 0$, while Theorem 1 implies that a monotonically *decreasing* sequence $\kappa \equiv \kappa_{n,T}$ yields $CV(\kappa) \to_P \sigma^2$. A precise theoretical analysis is not in the scope of the present paper.

2) Our theoretical results rest upon the assumption of i.i.d. errors. This is different from Bai (2009) who allows some correlation and heteroskedasticity of ϵ_{ij} in both cross-section and time-series dimension. We expect that results similar to Theorem 1 can be established in this context, although rates of convergence and, in particular, the distributions in (23) and (24) may change in dependence of the correlation structure. A precise analysis is not in the scope of the present paper.

3.3 Dimensionality and model tests

Result (23) of Theorem 1(e) may be used to estimate the dimension L. A prerequisite is of course the availability of a reasonable estimator of σ^2. We propose to use

$$\hat{\sigma}^2 := \frac{1}{(n-1)\cdot tr(I-\mathcal{Z}_\kappa)^2}\sum_{i=1}^n \|(I-\mathcal{Z}_\kappa)(Y_i-\bar{Y}-(X_i-\bar{X})\hat{\beta})\|^2. \tag{25}$$

We want to emphasize that this estimator may have a tendency to overestimate σ^2, but it is suitable for dimension selection (see proof of Theorem 2). Once L has been determined, a better estimator is $\tilde{\sigma}^2 = \frac{1}{(n-1)T}\sum_{i=1}^n \|Y_i-\bar{Y}-(X_i-\bar{X})\hat{\beta}-\sum_{r=1}^L \hat{\theta}_{ir}\hat{g}_r\|^2$. It follows from the results of Theorem 1 that $\tilde{\sigma}^2$ is consistent and may be used in the context of model tests (see below). We now use the following procedure to determine an estimate $\hat{L}$ of L:

First select an $\alpha > 0$ (e.g., $\alpha = 1\%$). For $l = 1, 2, \ldots$ determine

$$\Delta(l) := \frac{n\sum_{r=l+1}^T \hat{\lambda}_r \ - (n-1)\hat{\sigma}^2\cdot tr(\mathcal{Z}_\kappa\hat{\mathcal{P}}_l\mathcal{Z}_\kappa)}{\hat{\sigma}^2\sqrt{2n\cdot tr((\mathcal{Z}_\kappa\hat{\mathcal{P}}_l\mathcal{Z}_\kappa)^2)}}. \tag{26}$$

Choose $\hat{L}$ as the smallest $l = 1, 2, \ldots$ such that $\Delta(l) \leq z_{1-\alpha}$, where $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution.

The following theorem provides a theoretical justification of this procedure. A proof is given in the appendix.

Theorem 2. In addition to the assumptions of Theorem 1 assume that $\frac{T}{n\max\{1,\kappa^{1/4}\}} \to 0$ as well as $Td(T)b_\beta(n,T,\kappa) + \frac{d(T)}{n} + \frac{1}{Tc(T)} = o\left(\frac{T}{n\max\{1,\kappa^{1/4}\}}\right)$. Then, $\liminf_{n,T\to\infty}\mathbf{P}(\hat{L} = L) \geq 1-\alpha$ for fixed $\alpha > 0$. Moreover, if $\alpha \equiv \alpha_{n,T}$ is such that $\alpha_{n,T} \to 0$ and $z_{1-\alpha_{n,T}} \leq \log(\min\{n,T\})$ as $n,T\to\infty$, then $\lim_{n,T\to\infty}\mathbf{P}(\hat{L} = L) = 1$.

There are of course alternative ways for estimating L. Bai and Ng (2002) propose six related criteria for determining the dimension of a factor model: PC_{p1} - PC_{p3} and IC_{p1} - IC_{p3}. For example, in our context PC_{p2} consists in minimizing

$$\frac{1}{nT}\sum_{i=1}^n \|Y_i-\bar{Y}-(X_i-\bar{X})\hat{\beta}-\sum_{r=1}^l \hat{\theta}_{ir}\hat{g}_r\|^2 + l\hat{\sigma}^{2*}\frac{n+T}{nT}\log(\min\{n,T\})$$

over $l = 0, 1, \ldots, L_{max}$, where $L_{max} > L$ is a maximal possible number of factors, and where $\hat{\sigma}^{2*}$ is some consistent estimate of σ^2. IC_{p1} - IC_{p3} use slightly different penalty functions, thus $\frac{1}{nT}\sum_{i=1}^n \|Y_i-\bar{Y}-(X_i-\bar{X})\hat{\beta}-\sum_{r=1}^l \hat{\theta}_{ir}\hat{g}_r\|^2$ is replaced by $\log(\frac{1}{nT}\sum_{i=1}^n \|Y_i-\bar{Y}-(X_i-\bar{X})\hat{\beta}-\sum_{r=1}^l \hat{\theta}_{ir}\hat{g}_r\|^2)$ in IC_{p1} - IC_{p3}. It is now easily seen that under the assumptions of Theorem 2 these criteria will lead to consistent estimates of L.[4] Different from (26) they may still be applicable if error terms are correlated. However, we want to emphasize that for i.i.d. errors our test-based selection method is much more specifically adapted to the underlying asymptotic distribution of noise components.

Relation (24) may serve to test the validity of a pre-specified parametric model of the form $v_i(t) = \sum_{j=1}^{L} \vartheta_{ri} g_r(t)$ for some known basis functions g_r. If $P_{\mathbf{g},L}$ denotes the projection matrix projecting into the $n-L$ dimensional linear space orthogonal to $span\{Z_\kappa g_1, \ldots, Z_\kappa g_L\}$, then the null hypothesis: $H_0 : v_i(t) = \sum_{j=1}^{L} \vartheta_{ri} g_r(t)$ is rejected if

$$\frac{n \cdot tr(\mathcal{P}_{\mathbf{g},L}\hat{\Sigma}_{n,T}) \ - (n-1)\tilde{\sigma}^2 \cdot tr(\mathcal{Z}_\kappa \mathcal{P}_{\mathbf{g},L} \mathcal{Z}_\kappa)}{\tilde{\sigma}^2 \sqrt{2n \cdot tr((\mathcal{Z}_\kappa \mathcal{P}_{\mathbf{g},L} \mathcal{Z}_\kappa)^2)}} > z_{1-\alpha} \tag{27}$$

Obviously, under H_0 we have $P_{\mathbf{g},L} = P_L$, and by (24) the test possesses an asymptotically correct size. But the derivation of (24) is based on the fact that $tr(P_L \Sigma_{n,T}) = 0$ and hence $tr(P_L \hat{\Sigma}_{n,T}) = tr(P_L(\hat{\Sigma}_{n,T} - \Sigma_{n,T}))$. If H_0 is false, then generally $tr(P_{\mathbf{g},L}\Sigma_{n,T}) = O_P(Tc(T))$, and therefore $tr(P_{\mathbf{g},L}\hat{\Sigma}_{n,T}) = tr(P_{\mathbf{g},L}\Sigma_{n,T}) + tr(P_{\mathbf{g},L}(\hat{\Sigma}_{n,T} - \Sigma_{n,T}))$ will in tendency be too large.

This test can of course be particularly applied to verify the validity of a standard panel model with constant individual effects. Then $L = 1$, $P_{\mathbf{g},L} = I - \frac{1}{T}11'$ with $1 = (1, \ldots, 1)'$, $c(T) = 1$, and $b_v(\kappa) = b_w(\kappa^*) = 0$ for all possible choices of κ, κ^*.

4 Simulations

In this section, we investigate the finite sample performances of the new estimator described in Section 2 (hereafter we will call it KSS estimator) through Monte Carlo experiments. A competing time-varying individual effects estimator is based on the Cornwell, Schmidt, and Sickles fixed effects estimator (CSSW, 1990). We specify the time-varying individual effects as a second-order polynomial in time using this estimator, as the authors did in their empirical illustration. We also consider the classical time-invariant fixed and the random effects estimators (Baltagi, 2005). These estimators also have been used extensively in the stochastic frontier productivity literature wherein the firm effects are interpreted as measures of relative technical efficiencies.

We consider the panel data model (1):

$$Y_{it} = \beta_0(t) + \sum_{j=1}^{p} \beta_j X_{itj} + v_i(t) + \epsilon_{it}$$

We simulate samples of size $n = 30, 100, 300$ with $T = 12, 30$ in a model with $p = 2$ regressors. The error process ϵ_{it} is drawn randomly from i.i.d. $N(0, 1)$. The values of true β are set equal to $(0.5, 0.5)$. In each Monte Carlo sample, the regressors are generated according to a bivariate VAR model as in Park, Sickles, and Simar (2003, 2007):

$$X_{it} = RX_{i,t-1} + \eta_{it}, \text{ where } \eta_{it} \sim \mathbf{N}(0, I_2), \; R = \begin{pmatrix} 0.4 & 0.05 \\ 0.05 & 0.4 \end{pmatrix} \tag{28}$$

To initialize the simulation, we choose $X_{i1} \sim N(0, (I_2 - R^2)^{-1})$ and generate the samples using (28) for $t \geq 2$. Then, the obtained values of X_{it} are shifted around three different means to obtain three balanced groups of firms from small to large. We fix each group at $\mu_1 = (5, 5)'$, $\mu_2 = (7.5, 7.5)'$, and $\mu_3 = (10, 10)'$. The idea is to generate a reasonable cloud of points for X. In all of our data generating processes (DGP's) we set the mean function $\beta_0(t) = 0$.

We generate time-varying individual effects in the following ways:

$$\begin{aligned}
\text{DGP1 \& DGP2} \quad &: \quad v_i(t) = \theta_{i0} + \theta_{i1}\frac{t}{T} + \theta_{i2}\left(\frac{t}{T}\right)^2 \\
\text{DGP3 \& DGP4} \quad &: \quad v_i(t) = \phi_i r_t \\
\text{DGP5 \& DGP6} \quad &: \quad v_i(t) = \upsilon_{i1} g_{1t} + \upsilon_{i2} g_{2t} \\
\text{DGP7 \& DGP8} \quad &: \quad v_i(t) = \xi_i \\
\text{DGP9 \& DGP10} \quad &: \quad v_i(t) = \varphi_{i0} + \varphi_{i1}\frac{t}{T} + \varphi_{i2}\left(\frac{t}{T}\right)^2 + \phi_i r_t + \upsilon_{i1} g_{1t} + \upsilon_{i2} g_{2t}
\end{aligned}$$

where θ_{ij} $(j = 0, 1, 2)$ ∼i.i.d. 5N(0,1), φ_{ij} $(j = 0, 1, 2)$ ∼i.i.d. 3N(0,1), $r_{t+1} = r_t + \delta_t$, $\delta_t, \phi_i, \xi_i, \upsilon_{ij} (j = 1, 2)$ ∼i.i.d. N(0,1), $g_{1t} = \sin(\pi t/4)$ and $g_{2t} = \cos(\pi t/4)$. The odd numbered DGPs are those with exogenous regressors and the even numbered DGPs are those with endogenous regressors. The correlation between the effects and the second regressor is chosen to be 0.5.[5] DGP1 and DGP2 utilize time varying effects that follow a second order polynomial in time that varies from cross-section to cross-section and possess 3 common factors. DGP3 and DGP4 specify the effects as random walk processes and have 1 common factor. DGP5 and DGP6 are considered in order to model effects with large temporal variations and have 2 common factors. DGP7 and DGP8 are the usual constant effects models with symmetric effects and of course have 1 common factor. We consider DGP9 and DGP10 in order to provide some evidence on the usefulness of our estimator in speaking to the ongoing debate on the number of factors displayed by stock returns (estimates range between 2 and 10) and macroeconomic time series (estimates range between 2 and 7) (Stock and Watson, 2005). DGP9 and DGP10 generate effects with 6 common factors.

For the KSS estimator, cubic smoothing splines were used to approximate $v_i(t)$ in step 1, and the smoothing parameter κ was selected by using 'leave-one-individual-out' cross-validation.[6] The coefficient parameter β is updated using $\hat{g}_r(t)$ obtained in step 3 of (16), which is found to generate substantial efficiency gains. However, the updated estimates $\hat{\beta}^{(1)}$ are not plugged into step 2 again because there is no efficiency gain observed for $\hat{g}_r(t)$. Simulation experiments were repeated 1,000 times, except for the DGP's with n=300. For those the number of simulations is 500 times.

We now present the simulation results. Because of space limitations, we can not display all of the Monte Carlo results.[7] We do, however, present results for DGP 1, 3, and 9 and discuss results from the other experiments. We present a variety of performance metrics for the competing estimators based on DGP1-10. We calculate normalized mean squared error (MSE), bias, variance, and empirical size (based on a nominal type I error of 0.05) for the coefficients. The normalized mean squared error is :

$$\mathbf{R}(\widehat{v}, \mathbf{v}) = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \left(\widehat{v}_i(t) - v_i(t)\right)^2}{\sum_{i=1}^{n} \sum_{t=1}^{T} v_i^2(t)}.$$

We also calculate the MSE of the estimated effects as well as the average optimal dimensions, L, chosen by $\Delta(l)$ criterion we outlined in the previous section. We note that the optimal dimension, L, is correctly chosen on average for the KSS estimator in all DGPs. Thus, we can verify the validity of the dimension test $\Delta(l)$ discussed in Section 3.

<Insert Tables 1,2,3 about here>

We examine the results for exogenous regressors first and those for endogenous regressors later. DGP1 is consistent with the assumptions for the time-varying effects of the CSSW estimator. Hence, we may expect that the CSSW estimator performs reasonably well, which is confirmed in Table 1. Table 1 also shows that the performances of the KSS estimator are comparable to those of the CSSW estimator. This implies that the KSS estimator is quite general and efficient in estimating time-varying effects of the forms of smooth curves such as the second order polynomials. As such, it is not surprising that the results of the KSS estimator is much better than those of Within and GLS estimators by any standards.

This is true even when the data is as small as n = 30 and T = 12. In particular, the KSS estimator outperforms these estimators in terms of MSE of effects.

DGP3 is considered to evaluate the performance of the estimators for the arbitrary form of individual effects generated by a random walk. Hence, estimators based on a relatively simple function of time such as we used for the CSS within estimator is not sufficient for this type of DGP. However, the KSS estimator does not impose any specific forms on the temporal pattern of effects, and thus it can approximate any shape of time varying effects. We may then expect good performances of the KSS estimator even in this situation, and the results confirm such a belief. The KSS estimator dominantly outperforms the other estimators. It is particularly conspicuous in terms of MSE of effects. CSSW performs reasonably well for the effects, but it is no better than the others for other criteria.

DGP5 generates effects with large temporal variations. As T increases, the variations become large. The other estimators assume pre-specified and simple functional forms, thus they are expected to perform less satisfactorily for this DGP. The KSS estimator allows arbitrary functional forms as well as multiple individual effects. Hence, it is expected to perform well even under this DGP. Indeed, the results show that the KSS estimator performs very well, especially for large T, with the correct number of L chosen on average.[8] On the other hand, the other estimators suffer from severe distortions in the estimates of the effects, although coefficient estimates look reasonably good.

DGP7 represents the reverse situation so that there is no temporal variation in the effects. The Within and GLS estimators work very well. However, the performance of the KSS estimator is fairly good and comparable to those of the Within and GLS estimators. These results indicate that the KSS estimator may be safely used even when temporal variation is not evident. DGP9 is based on a 6 factor model for the effects. The KSS estimator dominates the other treatments for heterogeneity as the number of cross sections and times series increase. In all experiments, the KSS estimator also has better size characteristics than the competing treatments. It also delivers on properly identifying the number of common factors, with an average value of $L = 6$.

Results from the even-numbered experiments correspond to data generating processes which extend the preceding odd-numbered experiment to a setting in which there exists correlation between the effects and the second regressor. We can see that the treatments in which such potential correlations are explicitly addressed via the within transformation (Within, CSSW, KSS) dominate the other estimators in most situations when the temporal patterns of the effects are either consistent with the particular estimator's assumptions or when they are nested within the estimator's general treatment of time varying effects. However, as a general statement, across all experiments only the KSS estimator stands out as the favored estimator. This is because the misspecification of the temporal pattern of the effects appears to be as important as the added complication that the effects are correlated with the second regressors. This issue does not appear to have been given the attention in the panel data literature that it deserves. Also, because the generation of the X_{it} via the VAR specifies a correlated set of regressors, coefficient biases and resulting distortions in estimated variances and empirical size are not localized in the second coefficient but impact the first coefficient as well. As n and T increase the KSS estimator again dominates the other treatments for unobserved heterogeneity.

5 Efficiency Analysis of Banking Industry

5.1 Empirical Model

We next compare the various estimators in an empirical illustration of efficiency changes in the US banking industry after a series of deregulatory initiatives in the early 1980's. We model the multiple output/multiple input banking technology using the output distance function (Adams, Berger, and Sickles, 1999). The output distance function, $D(Y, X) \leq 1$, provides a radial measure of technical efficiency by specifying the fraction of aggregated outputs (Y) produced by given aggregated inputs (X). An m-output, n-input deterministic distance function can be approximated by

$$\frac{\prod_j^m Y_j^{\gamma_j}}{\prod_k^q X_k^{\delta_k}} \leq 1, \tag{29}$$

for $j = 1, \ldots, m$ and $k = 1, \ldots, q$ where the index j denotes outputs, the index k denotes inputs, and the $\gamma_j's$ and the $\delta_k's$ are weights of outputs and inputs, respectively, describing the technology of a firm. If it is not possible to increase the index of total output without either decreasing an output or increasing an input, the firm is producing efficiently or the value of the distance function equals 1.

The Cobb-Douglas stochastic distance frontier that we utilize below in our empirical illustration can be derived by multiplying (29) through by the denominator, approximating the terms using natural logarithms of outputs and inputs, and adding a disturbance term ϵ_{it} to account for statistical noise. We also specify a nonnegative stochastic term $u_i(t)$ for the firm specific level of radial technical inefficiency, with variations in time allowed. We then normalize the outputs with respect to the first output and rearrange to get

$$\ln y_{J,it} = \sum_{j=1}^{m} \gamma_j(-\ln \widehat{y}_{j,it}) - \sum_{k=1}^{q} \delta_k(-\ln x_{k,it}) - u_i(t) + \epsilon_{it},$$

where y_J is the normalizing output and $\widehat{y}_{j,it} = y_{j,it}/y_{J,it}$, $j = 1,\ldots,m$, $j \neq J$. To streamline notations, let $Y_{it} = \ln y_{J,it}$, and define $p = m - 1 + q$ vectors X_{it} with elements $-\ln \widehat{y}_{j,it}$, , $j \neq J$, and $-\ln x_{k,it}$. Furthermore, set $\beta = (\gamma', \delta')$, and $v_i(t) = -u_i(t) - \beta_0(t)$, where $\beta_0(t) := \frac{1}{n}\sum_{i=1}^{n} -u_i(t)$. We can then write the stochastic distance frontier as

$$Y_{it} = \beta_0(t) + X_{it}'\beta + v_i(t) + \epsilon_{it}. \tag{30}$$

This model can be viewed as a generic panel data model we introduced in equation (1) above in which the effects are interpreted as time-varying firm efficiencies, and fits into the class of frontier models developed and extended by Aigner, Lovell, and Schmidt (1977), Meeusen and van den Broeck (1977), Schmidt and Sickles (1984), and Cornwell, Schmidt, and Sickles (1990)[9]. Once the individual effects $v_i(t)$ are estimated, technical efficiency for a particular firm i at time t is calculated as $TE_i(t) = \exp\{v_i(t) - \max_{j=1,\ldots,n}(v_j(t))\}$ for the CSSW (Cornwell, Schmidt, and Sickles, 1990) and the KSS estimators. Technical efficiency is calculated similarly for the standard time-invariant fixed effects and random effects estimators following Schmidt and Sickles (1984). We also consider the Battese and Coelli (BC, 1992) estimator which is a likelihood-based random effects estimator wherein the likelihood function is derived from a mixture of normal noise and an independent one-sided efficiency error, usually specified as a half-normal. In the BC estimator, effect levels are allowed to differ across cross-sectional units but their temporal pattern is fixed across cross-sectional units and are specified as technical efficiencies $TE_i(t) = \exp(-\eta(t - T))\xi_i$ where ξ_i are independent half normal random effects and η parameterizes the temporal pattern in the firms' efficiencies.

5.2 Data

We use panel data from 1984 through 1995 for U.S. commercial banks in limited branching regulatory environment. The data are taken from the Report of Condition and Income (Call Report) and the FDIC Summary of Deposits[10]. The data set include 667 banks or 8,004 total observations.

The variables used to estimate the Cobb-Douglas stochastic distance frontier are $Y = \ln(\text{real estate loans})$; $X = -\ln(\text{certificate of deposit}), -\ln(\text{demand deposit}), -\ln(\text{retail time and savings deposit}), -\ln(\text{labor}), -\ln(\text{capital})$, and $-\ln(\text{purchased funds})$; $Y^* = -\ln(\text{commercial and industrial loans/real estate loans})$, and $-\ln(\text{installment loans/real estate loans})$. For a complete discussion of the approach used in this paper, see Adams, Berger, and Sickles (1999).

5.3 Empirical Results

The Hausman-Wu test, which tests the correlation assumptions for regressors and individual effects, was performed. The test statistic is 203.58, and the null hypothesis of no correlation is rejected at the 1% significance level. Thus there is strong evidence against the exogeneity assumption underlying the random effects GLS estimator. Consequently, in the following analysis we do not report the results from the random effects GLS estimator. The assumption is also fatal to the consistency of the random effects BC estimator. However, we will provide estimation results for the BC estimator as well to compare them with those from the other estimators (Within, CSSW, and KSS) which are robust to the existence of correlation between regressors and effects.

We test the dimensionality using $\Delta(l)$ test. The dimension L is chosen according to the rule described in Section 3 with the maximum dimension set to 8. Using the 1% significance level, the critical value is 2.33. With $L = 7$ the test statistic is 1.36 which is below the critical value. The optimal choice of dimensionality is thus 7[11].

< Insert Table 4 and Figure 1 about here >

Table 4 presents parameter estimates from within, BC, CSSW, and KSS[12]. We have also calculated Spearman rank correlations of estimated effects between the three estimators. They show relatively close correspondences, ranging from a low of 0.7937 between KSS and BC to a high of 0.8974 between KSS and CSSW. Average technical efficiencies for Within, BC, CSSW, and KSS are 0.4553, 0.6111, 0.6220, 0.6056, respectively[13]. One may expect that during the period of deregulation firms tend to become more efficient due to increased competitive pressures in the industry. Figure 1 displays the temporal pattern of changes in average efficiency for time-variant efficiency estimators. We can also construct an estimate of efficiency change over the sample period based on a pooled estimator that combines estimates from each of the time-varying measures. These results indicate a consensus growth of about 0.8% per year in efficiency during the sample period. Were these rates of cost diminution applied to the US banking industry the implied savings based on 1995 revenues and costs (Klee and Natalucci, 2005) would be on the order of $30 billion-our estimated measure of the benefits from deregulation of this key service industry.

6 Conclusion

In this paper we have introduced a new approach to estimating temporal heterogeneity in panel data models. We estimate the effects using the procedure combining smoothing spline techniques with principal component analysis. In this way, we can approximate virtually any shapes of time-varying effects. As we have pointed out, these methods can be transparently ported to the time series literature to address the issues of proper detrending filters in time series models.

Simulation experiments show that previous estimators, which do not allow for general temporal variations in effects terms or which misspecify the temporal pattern of variations, may suffer from serious distortions. On the other hand, our new estimator performs very well regardless of the assumption on the temporal pattern of individual effects. We have used this estimator to analyze the technical efficiency of U.S. banks in the limited branching regulatory environment for relatively small banks for the period of 1984-1995, and discovered that the relatively small banks in our sample have became more efficient over the years. The implied savings to the banking industry by 1995, were all banks to have enjoyed a similar efficiency gain as did our sample banks, is on the order of $30b.

Of course there are extensions of our work that may be pursued. For example, relaxing covariate exogeneity by framing our model in a multivariate system would appear to be feasible. Our approach can also be used to address possible nonstationarities in univariate and multivariate panel systems. Another extension that we are pursuing (Bada, Kneip, Sickles, 2011) and which holds promise involves extensions of our methods to examine general panel model approaches using our factor model specification when the disturbance term exhibits various forms of weak and strong dependencies. Research on robust methods to control for general forms of unobserved heterogeneiety while consistently estimating important covariate effects is quite dynamic and holds great promise for development of many new and improved estimation methods and approaches.

7 Appendix: Proof of Theorems

The proof of our theorems relies on the following proposition which derives some basic properties of cubic spline estimators ($m = 2$). We want to note, however, that our setup is slightly different from usual spline theory which considers smoothing over the fixed interval $[0, 1]$.

Proposition 1. For all $T \geq 3$

$$\frac{1}{T}\|(I - \mathcal{Z}_\kappa)v\|^2 \leq 4\frac{\kappa}{T}\sum_{t=2}^{T-1}(v(t-1) - 2v(t) + v(t+1))^2 \tag{A.1}$$

holds for all possible $v = (v(1), \ldots, v(T))'$ and all $\kappa > 0$. Furthermore, there exist constants $D_0, D_1, D_2 < \infty$ such that for all sufficiently large T

$$tr(\mathcal{Z}_\kappa^2) \leq D_0 \frac{T}{\max\{1, \kappa^{1/4}\}},$$

and if $\kappa < 1$, then $\frac{1}{T}\|(I - Z_\kappa)v\|^2 \leq \frac{D_1\kappa^2}{T}\|(I - Z_1)v\|^2$ for all $v = (v(1), \ldots, v(T))'$.

Proof: We first analyze properties of Z_κ. Obviously for every $\kappa > 0$ all eigenvalues of Z_κ and of $I - Z_\kappa$ are between 0 and 1. Further properties of this matrix are well studied for spline estimators defined on the fixed interval $[0, 1]$. But we have $z_j(t) = z_j^*(t/T)$, where $z_1, \ldots, z_T$ is the natural spline basis used to construct our estimator in Section 3.1, while $z_1^*, \ldots, z_T^*$ is a basis for all natural splines defined on $[0, 1]$ with knots $1/T, 2/T, \ldots, 1$. Obviously, $z_j'' = z_j^{*''}/T^2$. Defining the matrices Z^* and $A^* = \{\int_{1/T}^1 z_j^{*(m)}(s) z_k^{*(m)}(s)ds\}_{j,k=1,\ldots,T}$ similar to Z, A in Section 3.1, some straightforward arguments show that $Z_\kappa = (I + \kappa(Z')^{-1}AZ^{-1})^{-1} = (I + \frac{\kappa}{T^4}T(Z^{*\prime})^{-1}A^*(Z^*)^{-1})^{-1}$. Let $\psi_1 < \psi_2 < \ldots$ denote the eigenvalues of $T(Z^{*\prime})^{-1}A^*(Z^*)^{-1}$. Since we consider cubic smoothing splines, we have $\psi_1 = \psi_2 = 0$, and the results of Utreras (1983) imply that there exist constants $0 < Q_0, Q_1 < \infty$ such that $Q_0 \leq \psi_j \cdot (\pi j)^{-4} \leq Q_1$ for all $j = 3, \ldots, T$ and all sufficiently large T. Obviously, the eigenvalues of Z_κ^2, $I - Z_\kappa$ and $(I - Z_\kappa)^2$ then are $\frac{1}{(1+\kappa T^{-4}\psi_j)^2}$, $\frac{\kappa T^{-4}\psi_j}{1+\kappa T^{-4}\psi_j}$ and $\frac{\kappa^2 T^{-8}\psi_j^2}{(1+\kappa T^{-4}\psi_j)^2}$. We can conclude that there exist constants $D_1, D_2, D_2^*, D_3, D_3^*$ such that $tr(Z_\kappa^2) \leq D_2 T/\max\{1, \kappa^4\}$ and such that for all possible vectors v with $(I - Z_1)v \neq 0$

$$D_1^*\kappa^2 \leq \frac{v'(I - \mathcal{Z}_\kappa)^2 v}{v'(I - \mathcal{Z}_1)^2 v} \leq D_1\kappa^2, \quad D_3^*\kappa \leq \frac{v'(I - \mathcal{Z}_\kappa)v}{v'(I - \mathcal{Z}_1)v} \leq D_3\kappa, \quad \text{if } 0 < \kappa < 1 \tag{A.2}$$

Let us now analyze bias for $\kappa \geq 1$. By definition, the vector $Z_\kappa v$ is obtained by $Z_\kappa v = (\nu(1), \ldots, \nu(T))'$, where ν minimizes $\frac{1}{T}\sum_{t=1}^T (v(t) - \nu(t))^2 + \kappa\frac{1}{T}\int_1^T |\nu''(t)|^2 dt$ with respect to all cubic natural spline functions defined on the knot sequence $1, \ldots, T$. Let s_v denote the cubic spline interpolant of v, i.e. s_v is the (unique) natural spline function satisfying $s_v(t) = v(t)$ for all $t = 1,\dots,T$. Since $\frac{1}{T}\sum_{t=1}^T (v(t) - s_v(t))^2 = 0$, we can conclude that $\frac{1}{T}\|(I - Z_\kappa)v\|^2 = \frac{1}{T}\|v - Z_\kappa v\|^2 \leq \kappa \frac{1}{T}\int_1^T s_v''(\tau)^2 d\tau$.

The well-known properties of cubic spline interpolants (see for example de Boor, 1978) imply that $s_v''(1) = s_v''(T) = 0$, and $s_v''(s) = s_v''(t+1)[s-t] + s_v''(t)[t+1-s]$ for $s \in [t, t+1]$. Therefore, $\int_1^T s_v''(\tau)^2 d\tau = \sum_{t=1}^{T-1} \frac{1}{3}(s_v''(t)^2 + s_v''(t+1)^2 + s_v''(t)s_v''(t+1)) \leq \sum_{t=2}^{T-1} s_v''(t)^2$. Furthermore, the second derivatives of s_v at $t = 2,\dots,T-1$ are to be computed by the system of equations $s_v''(t-1) + 4s_v''(t) + s_v''(t+1) = 6(v(t-1) - 2v(t) + v(t+1)$. Hence, if B denotes the $(T-2) \times (T-2)$ matrix with $B_{ij} = 4$ if $i = j$, $B_{ij} = 1$ if $|i-j| = 1$, and $B_{ij} = 0$ if $|i-j| > 1$, $i, j = 1,\dots,T-2$, we obtain

$$\begin{pmatrix} s_v''(2) \\ \vdots \\ s_v''(T-1) \end{pmatrix} = 6B^{-1} \begin{pmatrix} v(1) - 2v(2) + v(3) \\ \vdots \\ v(T-2) - 2v(T-1) + v(T) \end{pmatrix}.$$

But B is a diagonal dominant matrix and by Gershgorin's circle theorem its smallest eigenvalue is larger or equal to 3. It follows that $\sum_{t=2}^{T-1} s_v''(t)^2 \leq 4\sum_{t=2}^{T-1}(v(t-1) - 2v(t) + v(t+1))^2$. Relation (A.1) is an immediate consequence.

Proof of Theorem 1: To simplify notation let $\tilde{X}_i = X_i - \bar{X}$, $\tilde{X}_{ij} = X_{ij} - \bar{X}_j$, and let $\kappa := \min\{\kappa, \kappa^2\}$. We obtain

$$\begin{aligned}\hat{\beta} &= (\sum_{i=1}^n \tilde{X}_i'(I - \mathcal{Z}_\kappa)\tilde{X}_i)^{-1} \sum_{i=1}^n \tilde{X}_i'(I - \mathcal{Z}_\kappa)(Y_i - \bar{Y}) \\ &= \beta + (\sum_{i=1}^n \tilde{X}_i'(I - \mathcal{Z}_\kappa)\tilde{X}_i)^{-1} \sum_{i=1}^n \tilde{X}_i'(I - \mathcal{Z}_\kappa)v_i + (\sum_{i=1}^n \tilde{X}_i'(I - \mathcal{Z}_\kappa)\tilde{X}_i)^{-1} \sum_{i=1}^n \tilde{X}_i'(I - \mathcal{Z}_\kappa)(\epsilon_i - \bar{\epsilon}).\end{aligned}$$

Consequently, $E_\epsilon(\hat{\beta}) - \beta = (\sum_{i=1}^n \tilde{X}_i'(I - Z_\kappa)\tilde{X}_i)^{-1} \sum_{i=1}^n \tilde{X}_i'(I - Z_\kappa)v_i$.

Let g_r be defined by (7) when replacing there γ_{rt} by the eigenvectors γ_{rt} of $\Sigma_T := E(v_i v_i' | \mathcal{L}_T)$. Then $L_T := span\{g_1,\dots,g_L\}$, and $v_i = \sum_{r=1}^L \vartheta_{ir} b_r$ with $E(\vartheta_{ir}) = 0$. Furthermore, Assumption 3) and Proposition 1 imply that $E(\vartheta_{ir}^2 \frac{1}{T}\|(I - Z_\kappa)g_r\|^2) = O(\kappa b(T))$ for all $r = 1,\dots,L$. Let X_{ij} denote the T-vectors with elements X_{itj}, $t = 1,\dots,T$, and recall that by the Markov inequality we have $P(|Z_{n,T}| \geq \delta) \leq E(|Z_{n,T}|^r)/\delta^r$ for all possible sequences of random variables $|Z_{n,T}|$ with $E(|Z_{n,T}|^r) < \infty$ and all $\delta > 0$. We thus necessarily have $Z_{n,T} = O_P(E(|Z_{n,T}|^r)^{1/r})$. This generalizes to conditional expectations.

In the general case, the $j = 1,\dots,p$ elements of the vectors $\sum_{i=1}^n \tilde{X}_i'(I - Z_\kappa)v_i$ can thus be bounded by

$$|\sum_{i=1}^{n} \tilde{X}_{ij}'(I - \mathcal{Z}_\kappa)v_i| \leq n \sum_{r=1}^{L} \sqrt{|\frac{1}{n}\sum_{i=1}^{n} \vartheta_{ir}^2| \cdot |\mathbf{g}_r'(I - \mathcal{Z}_\kappa)(\frac{1}{n}\sum_{i=1}^{n} \tilde{X}_{ij}\tilde{X}_{ij}')(I - \mathcal{Z}_\kappa)\mathbf{g}_r|}$$

We have $|g_r'(I-Z_\kappa)(\frac{1}{n}\sum_i \tilde{X}_{ij}\tilde{X}_{ij}')(I-Z_\kappa)g_r| = O_P(|g_r'(I-Z_\kappa)E[\tilde{X}_{ij}\tilde{X}_{ij}'|\, L_T](I-Z_\kappa)g_r|)$, and Assumption 4 thus leads to $|g_r'(I - Z_\kappa)(\frac{1}{n}\sum_i \tilde{X}_{ij}\tilde{X}_{ij}')(I - Z_\kappa)g_r| = O_P(C_1\|(I - Z_\kappa)g_r\|^2)$. Furthermore, $\frac{1}{n}\sum_i \vartheta_{ir}^2 = O_P(E[\vartheta_{ir}^2|L_T]$, and for any random variables Z_1, Z_2, V_1, V_2, the relations $Z_1 = O_P(V_1), Z_2 = O_P(V_2)$ imply that $Z_1Z_2 = O_P(V_1V_2)$. Together with Assumption 3) and Proposition 1 we can thus conclude that

$$\begin{aligned}
|\sum_{i=1}^{n} \tilde{X}_{ij}'(I - \mathcal{Z}_\kappa)v_i| &= O_P\left(n\sum_{r=1}^{L} \sqrt{\mathbf{E}[\vartheta_{ir}^2|\mathcal{L}_T] \cdot C_1 \cdot \|(I - \mathcal{Z}_\kappa)\mathbf{g}_r\|^2}\right) \\
&= O_P\left(n\sum_{r=1}^{L} \sqrt{C_1\mathbf{E}[\vartheta_{ir}^2\|(I - \mathcal{Z}_\kappa)\mathbf{g}_r\|^2]}\right) = O_P(n\sqrt{T\boldsymbol{\kappa}b(T)}).
\end{aligned}$$

It follows from (20) as well as (A.2) that $(\sum_i \tilde{X}_i'(I - Z_\kappa)\tilde{X}_i)^{-1} = O_P(\frac{1}{\max\{1,\kappa\}nT})$. When combining these arguments we arrive at $\|E_\epsilon(\hat{\beta}) - \beta\| = O_P((\frac{\max\{1,\kappa\}b(T)}{T})^{1/2})$.

Note that $Z_\kappa z = z$ and $(I-Z_\kappa)z = (I-Z_\kappa)^{1/2}z = 0$ for all κ, if $z = (z(1), \ldots, z(T))'$ is a linear function. If v_i and X_i are ulc-uncorrelated, then in the notation used in the definition of ulc-uncorrelatedness $\tilde{X}_{ij}'(I - Z_\kappa)^{1/2} = \tilde{X}_{ij}^{*\prime}(I - Z_\kappa)^{1/2}$, $(I - Z_\kappa)^{1/2}v_i = (I - Z_\kappa)^{1/2}v_i^*$, and therefore

$$\begin{aligned}
&\mathbf{E}[\tilde{X}_{ij}'(I - \mathcal{Z}_\kappa)v_i)^2|\mathcal{L}_T] \\
&\quad = tr\left(\mathbf{E}[(I - \mathcal{Z}_\kappa)^{1/2}\tilde{X}_{ij}\tilde{X}_{ij}'(I - \mathcal{Z}_\kappa)^{1/2}|\mathcal{L}_T] \cdot \mathbf{E}[(I - \mathcal{Z}_\kappa)^{1/2}v_iv_i'(I - \mathcal{Z}_\kappa)^{1/2}|\mathcal{L}_T]\right) \\
&\quad = \sum_{r=1}^{L} \mathbf{E}[\vartheta_{ir}^2|\mathcal{L}_T] \cdot \mathbf{g}_r'(I - \mathcal{Z}_\kappa)\mathbf{E}[\tilde{X}_{ij}\tilde{X}_{ij}'|\mathcal{L}_T](I - \mathcal{Z}_\kappa)\mathbf{g}_r = O_P(T \cdot \boldsymbol{\kappa}b(T))
\end{aligned}$$

Since due to our normalization $E(v_i(t)v_l(t)) = O(E(v_i(t)^2/n)$, it can be shown by similar arguments that $E[\tilde{X}_{ij}'(I - Z_\kappa)v_i)(\tilde{X}_{lj}'(I - Z_\kappa)v_l|L_T] = O_P(T \cdot \kappa b(T)/n)$ for $i \neq l$. Therefore, $E[\sum_i \tilde{X}_{ij}'(I - Z_\kappa)v_i)^2|L_T] = O_P(nT \cdot \kappa b(T))$, and $|\sum_i \tilde{X}_{ij}'(I - Z_\kappa)v_i| = O_P(\sqrt{nT \cdot \boldsymbol{\kappa}b(T)})$, which leads to $\|E_\epsilon(\hat{\beta}) - \beta\| = O_P(\sqrt{\max\{1,\kappa\}b(T)/(nT)})$. By Assumptions 4) and 5) as well as (A.2) the assertion on $\hat{\beta} - E_\epsilon(\hat{\beta}) = (\sum_i \tilde{X}_i'(I - Z_\kappa)\tilde{X}_i)^{-1}\sum_i \tilde{X}_i'(I - Z_\kappa)(\epsilon_i - \bar{\epsilon}) = (\sum_i \tilde{X}_i'(I - Z_\kappa)\tilde{X}_i)^{-1}\sum_i \tilde{X}_i'(I - Z_\kappa)\epsilon_i$ follows from standard arguments.

In order to prove Assertion (b) first note that

$$\hat{v}_i = v_i + r_i, \quad \text{with } r_i = -(I - \mathcal{Z}_\kappa)v_i + \mathcal{Z}_\kappa(\epsilon_i - \bar{\epsilon}) + \mathcal{Z}_\kappa\tilde{X}_i(\beta - \hat{\beta}).$$

Therefore,

$$\hat{\Sigma}_{n,T} = \Sigma_{n,T} + B, \quad B = \frac{1}{n}\sum_{i=1}^{n}(v_i r_i' + r_i v_i' + r_i r_i'). \tag{A.3}$$

$\Sigma_{n,T}$ possesses exactly L nonzero eigenvalues $\lambda_1 > \dots \lambda_L$. Assertion (b) of Lemma A.1 of Kneip and Utikal (2001) implies that for all $r = 1, \dots, L$

$$\gamma_r - \hat{\gamma}_r = S_r B \gamma_r + R, \quad \text{with} \ \ \|R\| \leq \frac{6 \sup_{\|a\|=1} a'B'Ba}{\min_s |\lambda_r - \lambda_s|^2} \tag{A.4}$$

and with $S_r = \sum_{s\neq r} \frac{1}{\lambda_s - \lambda_r} P_s - \frac{1}{\lambda_r} P_{L+1}$, where P_s denotes the projection matrix projecting into the eigenspace corresponding to the eigenvalue λ_s of $\Sigma_{n,T}$, while $P_{L+1} = I - \sum_{r=1}^{L} \gamma_r \gamma_r'$.

In order to evaluate the above expression we first have to analyze the stochastic order of magnitude of the different elements of B. Consider the terms appearing in $\frac{1}{n}\sum_i (v_i r_i' + r_i v_i')$. Using Assumptions 1) - 4) together with Proposition 1 some straightforward arguments now lead to

$$\sup_{\|a\|=1} \|\frac{1}{n}\sum_{i=1}^{n}(I - \mathcal{Z}_\kappa) v_i v_i' a\| \leq \frac{1}{n}\sum_{i=1}^{n}\sup_{\|a\|=1} |v_i' a| \sqrt{v_i'(I - \mathcal{Z}_\kappa)(I - \mathcal{Z}_\kappa) v_i} = O_P(T\sqrt{c(T)\boldsymbol{\kappa} b(T)}), \tag{A.5}$$

$$\sup_{\|a\|=1} \|\frac{1}{n}\sum_{i=1}^{n} v_i v_i'(I - \mathcal{Z}_\kappa) a\| \leq \sup_{\|a\|=1} \frac{1}{n}\sum_{i=1}^{n} \sqrt{v_i' v_i}\, |v_i'(I - \mathcal{Z}_\kappa) a| = O_P(T\sqrt{c(T)\boldsymbol{\kappa} b(T)}), \tag{A.6}$$

$$\begin{aligned}\sup_{\|a\|=1} \|\frac{1}{n}\sum_{i=1}^{n}(\mathcal{Z}_\kappa \tilde{X}_i(\beta - \hat{\beta})) v_i' a\| &\leq \frac{1}{n}\sum_{i=1}^{n} |v_i' a| \sqrt{(\beta - \hat{\beta})' \tilde{X}_i' \mathcal{Z}_\kappa^2 \tilde{X}_i (\beta - \hat{\beta})} \\ &= O_P\left(T\sqrt{c(T) d(T)(b_\beta(n,T,\kappa) + 1/(nT))}\right).\end{aligned} \tag{A.7}$$

By similar arguments

$$\sup_{\|a\|=1} \|\frac{1}{n}\sum_{i=1}^{n} v_i (\mathcal{Z}_\kappa \tilde{X}_i(\beta - \hat{\beta}))' a\| = O_P\left(T\sqrt{c(T) d(T)(b_\beta(n,T,\kappa) + 1/(nT))}\right) \tag{A.8}$$

Recall that $tr(Z_\kappa^2) = O(T/\max\{1, \kappa^{1/4}\})$. Obviously, $E(tr((\frac{1}{n}\sum_i v_i \epsilon_i' Z_\kappa)\cdot(\frac{1}{n}\sum_i Z_\kappa \epsilon_i v_i'))) = O_P(\frac{Tc(T)\cdot tr(\mathcal{Z}_\kappa^2)}{n}) = O_P(T^2 c(T)/(\max\{1, \kappa^{1/4}\} n))$, and $\frac{1}{n}\sum_i v_i \bar{\epsilon}' Z_\kappa = 0$. Therefore

$$\sup_{\|a\|=1} \|\frac{1}{n}\sum_{i=1}^{n} \mathcal{Z}_\kappa (\epsilon_i - \bar{\epsilon}) v_i' a\| \leq [tr((\frac{1}{n}\sum_{i=1}^{n} v_i \epsilon_i' \mathcal{Z}_\kappa)\cdot(\frac{1}{n}\sum_{i=1}^{n} \mathcal{Z}_\kappa \epsilon_i v_i'))]^{\frac{1}{2}} = O_P\left(T\sqrt{\frac{c(T)}{\max\{1, \kappa^{1/4}\} n}}\right), \tag{A.9}$$

Similarly,

$$\sup_{\|a\|=1} \|\frac{1}{n}\sum_{i=1}^{n} v_i (\epsilon_i - \bar{\epsilon})' \mathcal{Z}_\kappa a\| = O_P\left(T\sqrt{\frac{c(T)}{\max\{1, \kappa^{1/4}\} n}}\right). \tag{A.10}$$

For the leading terms appearing in $\frac{1}{n}\sum_i r_i r_i'$ we obtain

$$\sup_{\|a\|=1} \|\frac{1}{n}\sum_{i=1}^{n}(I-\mathcal{Z}_\kappa)v_i v_i'(I-\mathcal{Z}_\kappa)a\| = O_p(T\cdot \boldsymbol{\kappa} b(T)), \tag{A.11}$$

$$\sup_{\|a\|=1} \|\frac{1}{n}\sum_{i=1}^{n}(\mathcal{Z}_\kappa \tilde{X}_i(\beta-\hat{\beta}))(\mathcal{Z}_\kappa \tilde{X}_i(\beta-\hat{\beta}))'a\| = O_P\left(Td(T)\cdot (b_\beta(n,T,\kappa)+1/(nT))\right). \tag{A.12}$$

Since all eigenvalues of Z_κ take values between 0 and 1, we have $tr(Z_\kappa)^4 \leq tr(Z_\kappa^2) = O(T/(\max\{1,\kappa^{1/4}\}n))$, and thus $E(tr[(\frac{1}{n}\sum_i Z_\kappa\epsilon_i\epsilon_i' Z_\kappa - \sigma^2 Z_\kappa^2)\cdot(\frac{1}{n}\sum_i Z_\kappa\epsilon_i\epsilon_i' Z_\kappa - \sigma^2 Z_\kappa^2)]) = \frac{1}{n}E(tr[Z_\kappa\epsilon_i\epsilon_i' Z_\kappa Z_\kappa\epsilon_i\epsilon_i' Z_\kappa - \sigma^4 Z_\kappa^4]) = O_P(tr(Z_\kappa^4)/n) = O_P(T/(\max\{1,\kappa^{1/4}\}n))$. Therefore,

$$\sup_{\|a\|=1} \|\frac{1}{n}\sum_{i=1}^{n}(\mathcal{Z}_\kappa(\epsilon_i-\bar{\epsilon})(\epsilon_i-\bar{\epsilon})'\mathcal{Z}_\kappa - \sigma^2\mathcal{Z}_\kappa^2)a\| = O_P\left(\sqrt{\frac{T}{\max\{1,\kappa^{1/4}\}n}}\right) \tag{A.13}$$

Assumption 2) additionally implies that $\frac{1}{\lambda_r} = O_P(\frac{1}{T\cdot c(T)})$ as well as $\frac{1}{\min_s|\lambda_r-\lambda_s|} = O_P(\frac{1}{T\cdot c(T)})$. When combining (A.4) with (A.5) - (A.13) we thus obtain

$$\begin{aligned}\|S_r B\gamma_r\| &\leq \|\sigma^2 S_r\mathcal{Z}_\kappa^2\gamma_r\| + \frac{1}{\min_s|\lambda_r-\lambda_s|}\|(B-\sigma^2\mathcal{Z}_\kappa^2)\gamma_r\| \\ &= \|\sigma^2 S_r\mathcal{Z}_\kappa^2\gamma_r\| + O_P\left(\sqrt{\frac{\boldsymbol{\kappa} b(T)+d(T)b_\beta(n,T,\kappa)}{c(T)}} + \sqrt{\frac{1}{nc(T)\max\{1,\kappa^{1/4}\}}}\right)\end{aligned} \tag{A.14}$$

By definition of S_r we have $S_r\gamma_r = 0$. Furthermore, Assumption 3 implies that $\|(I-\mathcal{Z}_\kappa)\gamma_r\| = O_P((\frac{\boldsymbol{\kappa} b(T)}{c(T)})^{1/2})$. Hence,

$$\|\sigma^2 S_r\mathcal{Z}_\kappa^2\gamma_r\| \leq \|\sigma^2 S_r(I-\mathcal{Z}_\kappa)\gamma_r\| + \|\sigma^2 S_r\mathcal{Z}_\kappa(I-\mathcal{Z}_\kappa)\gamma_r\| = O_P(\frac{(\boldsymbol{\kappa} b(T))^{1/2}}{Tc(T)^{3/2}}), \tag{A.15}$$

Let us now consider the remainder term R in (A.4). Note that all eigenvalues of Z_κ are less or equal to 1, and thus $\sup_{\|a\|=1} a'Z_\kappa^4 a \leq 1$. Relations (A.5) - (A.13) then imply

$$\begin{aligned}\frac{\sup_{\|a\|=1} a'B'Ba}{\min_s|\lambda_r-\lambda_s|^2} &\leq 2\frac{\sup_{\|a\|=1} a'(B-\sigma^2\mathcal{Z}_\kappa^2)'(B-\sigma^2\mathcal{Z}_\kappa^2)a}{\min_s|\lambda_r-\lambda_s|^2} + 2\frac{\sup_{\|a\|=1} a'\mathcal{Z}_\kappa^4 a}{\min_s|\lambda_r-\lambda_s|^2} \\ &= O_P\left(\frac{\boldsymbol{\kappa} b(T)+d(T)b_\beta(n,T,\kappa)}{c(T)} + \frac{1}{T^2c(T)^2} + \frac{1}{nc(T)\max\{1,\kappa^{1/4}\}})\right)\end{aligned} \tag{A.16}$$

By (A.4), (A.14), (A.15) and (A.16) the asserted rate of convergence follows from

$$T^{-1/2}\|g_r-\hat{g}_r\| = \|\gamma_r-\hat{\gamma}_r\| = O_P\left(\sqrt{\frac{\boldsymbol{\kappa} b(T)+d(T)b_\beta(n,T,\kappa)}{c(T)}} + \frac{1}{T^2c(T)^2} + \sqrt{\frac{1}{nc(T)\max\{1,\kappa^{1/4}\}}}\right).$$

Let us switch to Assertion (c). Definition of $\hat{\theta}_{ir}$ as well as Assertions a) and b) imply

$$\begin{aligned}\hat{\theta}_{ri} &= \frac{1}{T}\hat{g}_r'(Y_i - \bar{Y} - \tilde{X}_i\hat{\beta}) = \frac{1}{T}g_r'(Y_i - \bar{Y} - \tilde{X}_i\hat{\beta}) + \frac{1}{T}(\hat{g}_r - g_r)'(Y_i - \bar{Y} - \tilde{X}_i\hat{\beta}) \\ &= \theta_{ri} + \frac{1}{T}g_r'(\epsilon_i - \bar{\epsilon}) + O_P\left(\sqrt{\boldsymbol{\kappa} b(T) + d(T)b_\beta(n,T,\kappa) + (n\max\{1,\kappa^{1/4}\})^{-1}}\right).\end{aligned}$$

Note that $\sqrt{T}\frac{1}{T}g_r'(\epsilon_i - \bar{\epsilon}) = \sqrt{T}\cdot\frac{1}{T}g_r'\epsilon_i + o_P(1)$. Since $\frac{1}{T}g_r'g_r = 1$ we immediately obtain $\sqrt{T}\cdot\frac{1}{T}g_r'\epsilon_i \rightarrow_d N(0,\sigma^2)$. The asserted rate of convergence is an immediate consequence. Note that due to $g_r'g_s = 0$ the random variables $g_r'\epsilon_i$ and $g_s'\epsilon_i$ are uncorrelated for $r \neq s$. Hence, if additionally $\kappa b(T) + d(T)b_\beta(n,T,\kappa) + (n\max\{1,\kappa^{1/4}\})^{-1} = o(T^{-1})$, the assertion on the multivariate distribution of $\sqrt{T}(\hat{\theta}_{1i} - \theta_{1i}, \ldots, \hat{\theta}_{Li} - \theta_{Li})'$ follows from standard arguments. Since obviously

$$\|v_i - \sum_{r=1}^{L}\hat{\theta}_{ri}\hat{g}_r\| = \|\sum_{r=1}^{L}(\theta_{ri} - \hat{\theta}_{ri})g_r + \sum_{r=1}^{L}\hat{\theta}_{ri}(g_r - \hat{g}_r)\|,$$

Assertion d) is a straightforward consequence of Assumption 2) as well as Assertions b) and c). It remains to prove assertion (e). First note that

$$\hat{v}_i = \mathcal{Z}_\kappa v_i + \tilde{r}_i, \quad \text{with } \tilde{r}_i = \mathcal{Z}_\kappa(\epsilon_i - \bar{\epsilon}) + \mathcal{Z}_\kappa\tilde{X}_i(\beta - \hat{\beta}).$$

Consequently, with $\tilde{\Sigma}_n = Z_\kappa(\frac{1}{n}\sum_i v_i v_i')Z_\kappa$ we obtain

$$\hat{\Sigma}_n = \tilde{\Sigma}_n + \tilde{B}, \quad \tilde{B} = \frac{1}{n}\sum_i(\mathcal{Z}_\kappa v_i\tilde{r}_i' + \tilde{r}_i v_i'\mathcal{Z}_\kappa + \tilde{r}_i\tilde{r}_i').$$

$\tilde{\Sigma}_n$ possesses only L nonzero eigenvalues $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_L$ with corresponding eigenvectors $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_L$. Our assumptions and arguments similar to (A.4) - (A.16) then show that $\tilde{\lambda}_r = O(Tc(T))$, $\frac{1}{\min_s|\tilde{\lambda}_r - \tilde{\lambda}_s|} = O_P(\frac{1}{T\cdot c(T)})$, $\|\gamma_r - \tilde{\gamma}_r\| = O_P(\sqrt{\frac{\boldsymbol{\kappa} b(T)}{c(T)}})$, and

$$\|\hat{\gamma}_r - \tilde{\gamma}_r\| = O_P\left(\sqrt{\frac{d(T)b_\beta(n,T,\kappa)}{c(T)}} + \frac{1}{T^2c(T)^2} + \sqrt{\frac{1}{nc(T)\max\{1,\kappa^{1/4}\}}}\right) \tag{A.17}$$

for all $r, s = 1, \ldots, L$, $r \neq s$.

Assertion (a) of Lemma A.1. of Kneip and Utikal (2001) implies that

$$\sum_{r=L+1}^{T}\hat{\lambda}_r = tr(\mathcal{P}_L\tilde{B}) + R^*, \quad \text{with } R^* \leq \frac{6L\sup_{\|a\|=1}a'\tilde{B}'\tilde{B}a}{\min_s|\tilde{\lambda}_r - \tilde{\lambda}_s|} \tag{A.18}$$

where $P_L = I - \sum_{r=1}^{L}\tilde{\gamma}_r\tilde{\gamma}_r'$. Using again arguments similar to the proof of Assertion (c) it is easily seen that

$$\frac{6L\sup_{\|a\|=1}a'\tilde{B}'\tilde{B}a}{\min_s|\tilde{\lambda}_r - \tilde{\lambda}_s|} = O_P\left(Td(T)b_\beta(n,T,\kappa)^2 + \frac{1}{Tc(T)} + \frac{T}{n\max\{1,\kappa^{1/4}\}}\right), \tag{A.19}$$

$$tr(\mathcal{P}_L\tilde{B}) = tr\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{P}_L\mathcal{Z}_\kappa\tilde{X}_i(\beta-\hat{\beta})(\beta-\hat{\beta})'\tilde{X}_i'\mathcal{Z}_\kappa\right)+tr\left(\mathcal{P}_L\mathcal{Z}_\kappa(\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i-\bar{\epsilon})(\epsilon_i-\bar{\epsilon})')\mathcal{Z}_\kappa\right).$$

Some straightforward computations lead to

$$\mathbf{E}\left(tr(\mathcal{P}_L\mathcal{Z}_\kappa(\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i-\bar{\epsilon})(\epsilon_i-\bar{\epsilon})')\mathcal{Z}_\kappa)\right) = \sigma^2(1-\frac{1}{n})tr(\mathcal{Z}_\kappa\mathcal{P}_L\mathcal{Z}_\kappa),$$

$$\text{Var}\left(tr(\mathcal{P}_L\mathcal{Z}_\kappa(\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i-\bar{\epsilon})(\epsilon_i-\bar{\epsilon})')\mathcal{Z}_\kappa)\right) = \frac{2\sigma^4}{n}\cdot tr((\mathcal{Z}_\kappa\hat{P}_L\mathcal{Z}_\kappa)^2)\cdot(1+o_P(1)) = O_P\left(\frac{tr(\mathcal{Z}_\kappa^4)}{n}\right)$$

Moreover, $tr(Z_\kappa^4/n) = O(T/(n\max\{1,\kappa^{1/4}\}))$. Since $tr(\frac{1}{n}\sum_i P_L Z_\kappa\tilde{X}_i(\beta-\hat{\beta})(\beta-\hat{\beta})'\tilde{X}_i'Z_\kappa P_L) = O_P\left(Td(T)b_\beta(n,T,\kappa)+\frac{d(T)}{n}\right)$ and since by assumption $Td(T)b_\beta(n,T,\kappa)+\frac{d(T)}{n} = o\left(\sqrt{T/(n\max\{1,\kappa^{1/4}\}}\right)$ one may invoke standard arguments to show that

$$\frac{tr(\mathcal{P}_L\tilde{B}) \ -\sigma^2\left(1-\frac{1}{n}\right)tr(\mathcal{Z}_\kappa\mathcal{P}_L\mathcal{Z}_\kappa)}{\sqrt{\frac{2\sigma^4}{n}\cdot tr((\mathcal{Z}_\kappa\mathcal{P}_L\mathcal{Z}_\kappa)^2)}} \rightarrow_d \mathbf{N}(0,1). \tag{A.20}$$

Since $tr(P_L\tilde{B}) = tr(P_L\hat{\Sigma}_n)$, (24) is an immediate consequence. By (A.17)- (A.19) , Relation (A.20) remains valid when $tr(P_L\tilde{B})$ is replaced by $\sum_{r=L+1}^{T}\hat{\lambda}_r$ as well as P_L by $\hat{P}_L$. This proves (23) and hence completes the proof of the theorem. ■

Proof of Theorem 2: It follows from arguments similar to those used in the proof of Theorem 1 that

$$\hat{\sigma}^2 = \frac{1}{(n-1)\cdot tr((I-\mathcal{Z}_\kappa)^2)}\sum_{i=1}^{n}(\epsilon_i-\bar{\epsilon})'(I-\mathcal{Z}_\kappa)^2(\epsilon_i-\bar{\epsilon})$$
$$+\frac{1}{(n-1)\cdot tr((I-\mathcal{Z}_\kappa)^2)}\sum_{i=1}^{n}v_i'(I-\mathcal{Z}_\kappa)^2 v_i + O_P\left(d(T)^{1/2}\kappa b(T)\cdot(\kappa b_\beta(n,T)+\frac{1}{\sqrt{nT}})\right).$$

Obviously, $E\left(\frac{1}{(n-1)\cdot tr((I-\mathcal{Z}_\kappa)^2)}\sum_i(\epsilon_i-\bar{\epsilon})'(I-\mathcal{Z}_\kappa)^2(\epsilon_i-\bar{\epsilon})\right) = \sigma^2$and the properties of Z_κ imply that the variance of this term converges to 0 in probability. Consequently, with

$$0 \le R_{n,T} = \frac{1}{(n-1)\cdot tr((I-\mathcal{Z}_\kappa)^2)}\sum_{i=1}^{n}v_i'(I-\mathcal{Z}_\kappa)^2 v_i = O_p(\max\{1,\kappa\}b(T)) \tag{A.21}$$

we obtain

$$\hat{\sigma}^2 = \sigma^2 + R_{n,T} + o_p\left(1\right). \tag{A.22}$$

Let us now consider the behavior of $\Delta(l)$ for $l < L$. We can infer from (A.22) that

$$\Delta(l) = \left[\frac{n \sum_{r=l+1}^{L} \hat{\lambda}_r \;\; - (n-1)(\sigma^2 + R_{n,T}) \cdot tr(\mathcal{Z}_\kappa(\hat{\mathcal{P}}_l - \hat{\mathcal{P}}_L)\mathcal{Z}_\kappa) - (n-1)R_{n,T} \cdot tr(\mathcal{Z}_\kappa \hat{\mathcal{P}}_l \mathcal{Z}_\kappa)}{\hat{\sigma}^2 \sqrt{2n \cdot tr((\mathcal{Z}_\kappa \hat{\mathcal{P}}_l \mathcal{Z}_\kappa)^2)}} \right.$$
$$\left. + \frac{n \sum_{r=L+1}^{T} \hat{\lambda}_r \;\; - (n-1)\sigma^2 \cdot tr(\mathcal{Z}_\kappa \hat{\mathcal{P}}_L \mathcal{Z}_\kappa)}{\hat{\sigma}^2 \sqrt{2n \cdot tr((\mathcal{Z}_\kappa \hat{\mathcal{P}}_l \mathcal{Z}_\kappa)^2)}} \right] (1 + o_P(1)). \quad \text{(A.23)}$$

By Assumption 2) and Theorem 1d) $n \sum_{r=l+1}^{L} \hat{\lambda}_r = \sum_{r=l+1}^{L} T \sum_i \hat{\theta}_{ir}^2$ is of order $nTc(T)$, while $(n-1)(\sigma^2 + R_{n,T}) \cdot tr(Z_\kappa(\hat{P}_l - \hat{P}_L)Z_\kappa) = O_P(n)$, $(n-1)R_{n,T} \cdot tr(Z_\kappa \hat{P}_l Z_\kappa) = o_P(nTc(T))$, and $\sqrt{2n\hat{\sigma}^4 \cdot tr((\mathcal{Z}_\kappa \hat{\mathcal{P}}_l \mathcal{Z}_\kappa)^2)} = O_P((nT)^{1/2})$. Consequently, the first term on the right hand side of (A.23) increases as $n, T \to \infty$, while the second term is still bounded in probability. We can thus infer that for any $l < L$

$$\mathbf{P}(\Delta(l) > z_{1-\alpha}) \to 1, \mathbf{P}(\Delta(l) > \log(\min\{n, T\}) \to 1, \quad \text{and therefore } \mathbf{P}(\hat{L} \neq l) \to 1 \quad \text{(A.24)}$$

as $n, T \to \infty$. Since $R_{n,T} \geq 0$, Theorem 1(e) implies that for fixed $\alpha > 0$

$$\limsup_{n,T\to\infty} \mathbf{P}(\Delta(L) \geq z_{1-\alpha}) \leq \alpha, \quad \text{while } \lim_{n,T\to\infty} \mathbf{P}(\Delta(L) \geq z_{1-\alpha_{n,T}}) = 0 \text{ if } \alpha_{n,T} \to 0 \quad \text{(A.25)}$$

The assertions of the theorem are now immediate consequences of (A.24) and (A.25). ■

Notes

[1]In our conditions (a) - (c) sample averages could be replaced by expectations in (a) and (b) (for example $E(\theta_{ir}\theta_{is}) = 0$ or, more generally, $E(\theta_{ir}\theta_{is}|\mathcal{L}_T) = 0$ in case $\mathcal{L}_T$ is a random space). We would then have another standardization which would lead to different basis functions, let us call them g_r^*, which could be determined from the eigenvectors of the (conditional) covariance matrix Σ_T. Bai and others exactly use this standardization. In this case g_r^* still depends on T, but not on n. For this case, however, the g_r^* do not provide any additional information compared to our n-dependent g_r. The reason is that $v_i(t) = \sum_{r=1}^L \theta_{ir}^* g_r^* = \sum_{r=1}^L \theta_{ir} g_r$ for any possible realization v_i. Thus the $g_1^*, \ldots, g_L^*$ and $g_1 \ldots, g_L$ simply define different possible parametrizations of v_i. Nevertheless, we could use $g_1^*, \ldots, g_L^*$ instead of $g_1 \ldots, g_L$ to derive theoretical results. There are, however, disadvantages. Additional notation would be necessary resulting in a longer paper with little obvious value added. Furthermore, the difference between g_r^* and g_r is of order $n^{-1/2}$. Consequently, when considering the differences $\|g_r^* - \hat{g}_r\|$ and $\|\theta_{ir}^* - \hat{\theta}_{ir}\|$ there will exist an additional error of order $n^{-1/2}$, and rates of convergence deteriorate. This introduces some quite "artificial" bias since it only reflects the difference of standardization and not a true difference in describing and modeling v_i.

[2]After having estimated the components of (4), one may additionally be interested in estimating the mean function $\beta_0(t)$ in (1). When assuming that β_0 also adopts an expansion of the form $\beta_0(t) = \sum_{r=1}^L \bar{\theta}_r g_r(t)$, estimates of the coefficients $\bar{\theta}_r$ may be determined by minimizing $\sum_{t=1}^T (\bar{Y}_t - \sum_{j=1}^p \hat{\beta}_j \bar{X}_{tj} - \sum_{r=1}^L \vartheta_r \hat{g}_r(t))^2$ over $\vartheta_1, \ldots, \vartheta_L$. A more general approach consists in a nonparametric estimation similar to Step 1. Convergence rates can be obtained in a way similar to Theorem 1 below.

[3]The choice of the smoothing parameter affects the behavior of our estimators. It is known that one can characterize behavior of the effect of the smoothing parameter on the estimated functional form as κ tends to infinity. For example, when the penalty is the integrated quadratic of the higher derivative, the asymptotic form is a polynomial trend of order related to the derivative order, as shown in the Appendix of Phillips (2010).

[4]Note that our estimator $\hat{\beta}$ of β does not depend on L. Arguments similar to those used in the proof of Theorem 2 imply that for any $l < L$ there exists some $a_l > 0$ such that $P(\frac{1}{nT}\sum_{i=1}^n \|Y_i - \bar{Y} - (X_i - \bar{X})\hat{\beta} - \sum_{r=1}^l \hat{\theta}_{ir}\hat{g}_r\|^2 \geq \sigma^2 + a_l) \to 1$ as $n, T \to \infty$. For $L \leq l \leq L_{max}$ a generalization of the arguments of Bai and Ng (2002) leads to $|\frac{1}{nT}\sum_{i=1}^n \|Y_i - \bar{Y} - (X_i - \bar{X})\hat{\beta} - \sum_{r=1}^l \hat{\theta}_{ir}\hat{g}_r\|^2 - \sigma^2| = O_P(\min\{n, T\}^{-1})$. Consistency of the Bai and Ng criteria is an immediate consequence.

[5]Let $X_{it,2}$ be the endogenous part of the regressors X_{it} generated by (27). In order to generate a regressor that is correlated with $v_i(t)$, we define a variable, W_{it}, such that $W_{it} = \rho v_i(t) + \sigma_v\sqrt{1-\rho^2}\varepsilon_{it}$ where σ_v is the standard deviation of $v_i(t)$ and $\varepsilon_{it} \sim \mathbf{N}(0,1)$. Then, we see that $Corr(W_{it}, v_i(t)) = \rho = 0.5$. With this W_{it}, we generate $\tilde{X}_{it,2} = X_{it,2} + W_{it}$, which is used as the endogenous regressor. Here, $X_{it,2} + \sigma_v\sqrt{1-\rho^2}\varepsilon_{it}$ and $\rho v_i(t)$ constitute the exogenous part and endogenous part, respectively. Note that, in generating W_{it}, the effects $v_i(t)$ is multiplied by 10 to balance with the magnitude of $X_{it,2}$.

[6]We let $\kappa = (1-p)/p$ and choose p among a selected grid of 9 equally spaced values between 0.1 and 0.9.

[7]The full set of Monte Carlo results can be found at 'http://www.ruf.rice.edu/~rsickles/working%20papers /Sickles_Tables%201-12.pdf'.

[8]A referee asked for a comparison with the Bai and Ng (2002) criteria for the selection of the number

of factors. We ran a number of comparable Monte Carlo experiments that are available on request. We used the same DGP1-10 and tested for the number of factors for n=30, 100, 300 and for T=12, 30. For DGP1-8, the maximum dimension of factors is set to 5 and for DGP9-10 it is set at 8. In our simulation experiments, we estimated the number of factors using all six criteria proposed by Bai and Ng (2002). As noted in their paper, the criteria are inadequate for small n or T and we verify these findings when T=12 or n=30. We also find that the IC criterion tends to underparametrize, while the PC criterion tends to overparametrize. Particularly, for DGP9 and 10 where there are 6 different types of factors and the factors are correlated with regressors, the performances of Bai and Ng's criteria are very poor and unstable across different n and T. Indeed Bai and Ng mention in their 2002 paper (page 203) that their methods work well only when min{n,T} is 40 or larger. However, our simulation setup is for T=12, 30 which are quite small numbers for Bai and Ng's method. Our simulation results show exactly what is expected, that is, the IC criterion tends to underparameterize (for DGP1, 2) and the PC tends to overparameterize (for DGP3˜8).

[9]In keeping with the stochastic frontier paradigm we allow the technical efficiency to be correlated with the potentially distorted relative output allocations $-\ln \widehat{y}_{j,it}$.

[10]For a more detailed discussion of data, see the Appendix in Jayasiriya (2000).

[11]When we assume $L = 1$ and test the null hypothesis that the individual effect is constant, the test statistic (27) is 73.91. Thus the null hypothesis of linear individual effect is strongly rejected.

[12]We report results with ray returns to scale set to one. No significant ray scale economies appear to exist using these treatments and in other analysis conducted by the authors with these data. Moreover, the equivalence of input and output oriented technical efficiency is preserved when scale economies are unity, thus avoiding difficutlies in interpretation that have been pointed out often in the productivity literature.

[13]To calculate efficiency scores from the effects estimators, the effects estimates are trimmed at the top and bottom 5% level (see Berger, 1993). This does not apply to the BC estimator because it directly calculates efficiencies. For the time-varying effects estimators, the firms which enter the top and bottom 5% range of effects in any time periods were excluded in calculating average efficiencies. Therefore, in this sense, it is not fair to directly compare the efficiencies from the Within or BC estimators with those from the CSS and KSS estimators.

References

Adams, R.M., A.N. Berger & R.C. Sickles (1999) Semiparametric approaches to stochastic panel frontiers with applications in the banking industry. *Journal of Business and Economic Statistics* 17, 349-358.

Ahn, S. C., Y. Lee & P. J. Schmidt (2007) Stochastic frontier models with multiple time-varying individual effects. *Journal of Productivity Analysis* 27, 1-12.

Aigner, D.J., C.A.K. Lovell & P. Schmidt (1977) Formulation and estimation of stochastic frontier models. *Journal of Econometrics* 6, 21-37.

Bada, O., A. Kneip & R. C. Sickles (2011) Panel models with temporal heterogeneity in effects and correlated errors. Mimeo. Rice University.

Bai, J. (2003) Inferential theory for factor models of large dimensions. *Econometrica* 71, 135-171.

Bai, J. (2009) Panel data models with interactive fixed effects. *Econometrica* 77, 1229-1279.

Bai, J. & S. Ng (2002) Determining the number of factors in approximate factor models. *Econometrica* 70, 191-221.

Baltagi, B., P. Egger & M. Pfaffermayr (2003) A generalized design for bilateral trade flow models. *Economics Letters* 80, 391-397.

Baltagi, B. (2005) *Econometric Analysis of Panel Data.* John Wiley & Sons.

Battese, G.E. & T.J. Coelli (1992) Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India. *Journal of Productivity Analysis* 3, 153-169.

Berger, A.N. (1993) “Distribution-free” estimates of efficiency in U.S. banking industry and tests of the standard distributional assumption. *Journal of Productivity Analysis* 4, 261-292.

Bernanke, B. S. & J. Boivin (2003) Monetary policy in a data-rich environment. *Journal of Monetary Economics* 50, 525-546.

de Boor, C. (1978) *A Practical Guide to Splines.* Springer-Verlag.

Chang, Y. (2004) Bootstrap unit root tests in panels with cross-sectional dependency. *Journal of Econometrics* 120, 263-293.

Cornwell, C., P. Schmidt & R.C. Sickles (1990) Production frontiers with cross-sectional and time-series variation in efficiency levels. *Journal of Econometrics* 46, 185-200.

Engle, R., C. Granger, J. Rice, & A. Weiss (1986) Nonparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* 81, 310-320.

Eubank, R.L. (1988) *Nonparametric Regression and Spline Smoothing.* Marcel Dekker.

Ferré, L. (1995) Improvement of some multivariate estimates by reduction of dimensionality. *Journal of Multivariate Analysis* 54, 147-162.

Forni, M. & M. Lippi (1997) *Aggregation and the Microfoundations of Dynamic Macroeconomics.* Oxford University Press.

Forni, M. & L. Reichlin (1998) Let's get real: a factor analytic approach to disaggregated business cycle dynamics. *Review of Economic Studies* 65, 653-473.

Forni, M., M. Hallin], M. Lippi, & L. Reichlin (2000) The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics* 82, 540-554.

Hahn, J. & W. Newey (2004) Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* 72, 1295-1319.

Härdle, W., H. Liang, & J. Gao (2000) *Partially Linear Models.* Physica-Verlag.

Jayasiriya, R. (2000) *Essays on Structural Modeling Using Nonparametric and Parametric Methods with Applications in the U.S. Banking Industry.* Unpublished Ph.D. dissertation. Rice University.

Kao, C. & M. H. Chiang (2000) On the estimation and inference of a cointegrated regression in panel data. *Advances in Econometrics* 15, 179-222.

Klee, E. C. & F. M. Natalucci (2005) Profits and balance sheet developments at U. S. commercial banks in 2004. *Federal Reserve Bulletin* Spring.

Kneip, A. (1994) Nonparametric estimation of common regressors for similar curve data. *Annals of Statistics* 22, 1386-1427.

Kneip, A. & K.J. Utikal (2001) Inference for density families using functional principal component analysis. *Journal of the American Statistical Association* 96, 519-532.

Maddala, G. S. & I. M. Kim (1998) *Unit Roots, Cointegration and Structural Change*. Cambridge University Press.

Mark, N. C. & D. Sul (2003) Cointegration vector estimation by panel DOLS and long-run money demand. *Oxford Bulletin of Economics and Statistics* 65, 655-680.

Meeusen, W. & J. van den Broeck (1977) Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* 18, 435-444.

Nelson, C. R. & C. I. Plosser (1982) Trends and random walks in macroeconomics time series: some evidence and implications. *Journal of Monetary Economics* 10, 139-162.

Park, B. U. & L. Simar (1994) Efficient semiparametric estimation in stochastic frontier models. *Journal of the American Statistical Association* 89, 929-936.

Park, B.U., R.C. Sickles & L. Simar (1998) Stochastic frontiers: a semiparametric approach. *Journal of Econometrics* 84, 273-301.

Park, B.U., R.C. Sickles & L. Simar (2003) Semiparametric efficient estimation of AR(1) panel data models. *Journal of Econometrics* 117, 279-309.

Park, B.U., R.C. Sickles & L. Simar (2007) Semiparametric efficient estimation of dynamic panel data models. *Journal of Econometrics* 136, 281-301.

Phillips, P.C.B. (2010) Two New Zealand pioneer econometricians. *New Zealand Economic Papers* 44, 1-26.

Ramsay, J. & B. Silverman (1997) *Functional Data Analysis*. Springer-Verlag.

Rao, C.R. (1958) Some statistical methods for the comparison of growth curves. *Biometrics* 14, 1-17

Schmidt, P. & R.C. Sickles (1984) Production frontiers and panel data. *Journal of Business and Economic Statistics* 2, 367-374.

Sickles, R. C. (2005) Panel estimators and the identification of firm-specific efficiency levels in parametric, semiparametric and nonparametric settings. *Journal of Econometrics* 50, 126, 305-334.

Speckman, P. (1988) Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B* 50, 413-436.

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

Stock, J. H. & M. W. Watson (2005) Implications of dynamic factor models for VAR analysis. National Bureau of Economic Research Working Paper 11467.

Utreras, F. (1983) Natural spline functions, their associated eigenvalue problem. *Numerical Mathematics* 42, 107-117.

Table 1. Monte Carlo Simulation Results for DGP1

MSE of Effects

n	T	Within	GLS	CSSW	KSS	L
30	12	0.1770	0.1745	0.0091	0.0091	2.405
	30	0.1666	0.1663	0.0036	0.0043	2.804
100	12	0.1285	0.1280	0.0072	0.0073	2.963
	30	0.1240	0.1240	0.0029	0.0030	3.010
300	12	0.1025	0.1025	0.0059	0.0060	3.002
	30	0.1001	0.1001	0.0024	0.0025	3.006

MSE, Bias, Variance, and Size for Coefficients

	T=12				T=30			
n=30	Within	GLS	CSSW	KSS	Within	GLS	CSSW	KSS
MSE	0.0725	0.0637	0.0087	0.0087	0.0247	0.0235	0.0024	0.0026
BIAS1	0.0004	0.0306	0.0015	0.0002	0.0055	0.0169	-0.0008	-0.0012
BIAS2	0.0014	0.0315	0.0012	0.0007	-0.0026	0.0087	0.0003	-0.0003
VAR1	0.0355	0.0302	0.0043	0.0044	0.0122	0.0115	0.0012	0.0013
VAR2	0.0370	0.0316	0.0044	0.0044	0.0124	0.0117	0.0012	0.0012
SIZE1	0.143	0.137	0.083	0.095	0.177	0.168	0.059	0.081
SIZE2	0.171	0.155	0.075	0.087	0.153	0.156	0.048	0.060
n=100	Within	GLS	CSSW	KSS	Within	GLS	CSSW	KSS
MSE	0.0186	0.0164	0.0027	0.0027	0.0068	0.0065	0.0007	0.0007
BIAS1	-0.0023	-0.0060	0.0004	0.0004	-0.0013	-0.0024	-0.0002	-0.0003
BIAS2	0.0031	-0.0007	0.0015	0.0015	-0.0008	-0.0019	0.0000	0.0001
VAR1	0.0095	0.0083	0.0012	0.0013	0.0033	0.0031	0.0004	0.0004
VAR2	0.0091	0.0080	0.0014	0.0015	0.0035	0.0034	0.0004	0.0004
SIZE1	0.163	0.149	0.068	0.072	0.164	0.163	0.058	0.072
SIZE2	0.154	0.140	0.096	0.096	0.171	0.162	0.067	0.068
n=300	Within	GLS	CSSW	KSS	Within	GLS	CSSW	KSS
MSE	0.0061	0.0061	0.0009	0.0009	0.0021	0.0021	0.0002	0.0002
BIAS1	-0.0021	-0.0178	-0.0004	-0.0002	-0.0014	-0.0071	-0.0001	0.0000
BIAS2	-0.0005	-0.0162	0.0016	0.0015	0.0031	-0.0027	0.0004	0.0004
VAR1	0.0032	0.0029	0.0005	0.0005	0.0011	0.0011	0.0001	0.0001
VAR2	0.0029	0.0026	0.0004	0.0004	0.0010	0.0009	0.0001	0.0001
SIZE1	0.154	0.176	0.092	0.094	0.186	0.180	0.070	0.076
SIZE2	0.160	0.150	0.060	0.062	0.146	0.132	0.060	0.060

Table 2. Monte Carlo Simulation Results for DGP3

MSE of Effects

n	T	Within	GLS	CSSW	KSS	L
30	12	0.1655	0.1630	0.0601	0.0170	1.005
	30	0.0976	0.0975	0.0692	0.0100	1.000
100	12	0.1554	0.1547	0.0491	0.0117	1.000
	30	0.0890	0.0890	0.0624	0.0074	1.000
300	12	0.1480	0.1484	0.0450	0.0103	1.000
	30	0.0860	0.0861	0.0597	0.0065	1.000

MSE, Bias, Variance, and Size for Coefficients

	T=12				T=30			
n=30	Within	GLS	CSSW	KSS	Within	GLS	CSSW	KSS
MSE	0.0241	0.0208	0.0137	0.0047	0.0070	0.0067	0.0066	0.0019
BIAS1	0.0003	0.0139	0.0011	-0.0024	0.0055	0.0091	0.0061	0.0003
BIAS2	0.0043	0.0180	0.0037	0.0000	-0.0003	0.0033	-0.0006	-0.0017
VAR1	0.0120	0.0101	0.0063	0.0022	0.0035	0.0034	0.0034	0.0009
VAR2	0.0121	0.0102	0.0074	0.0025	0.0034	0.0033	0.0032	0.0010
SIZE1	0.100	0.100	0.079	0.055	0.111	0.114	0.120	0.046
SIZE2	0.118	0.102	0.090	0.059	0.107	0.103	0.115	0.065
n=100	Within	GLS	CSSW	KSS	Within	GLS	CSSW	KSS
MSE	0.0097	0.0084	0.0049	0.0014	0.0020	0.0020	0.0019	0.0005
BIAS1	-0.0045	-0.0065	-0.0012	-0.0013	0.0001	-0.0028	-0.0002	0.0010
BIAS2	-0.0031	-0.0049	-0.0019	0.0006	0.0003	-0.0025	-0.0005	0.0012
VAR1	0.0045	0.0039	0.0024	0.0007	0.0010	0.0010	0.0010	0.0003
VAR2	0.0052	0.0045	0.0025	0.0007	0.0010	0.0010	0.0010	0.0003
SIZE1	0.098	0.090	0.082	0.045	0.085	0.085	0.107	0.049
SIZE2	0.130	0.115	0.105	0.050	0.099	0.098	0.107	0.044
n=300	Within	GLS	CSSW	KSS	Within	GLS	CSSW	KSS
MSE	0.0034	0.0043	0.0017	0.0005	0.0007	0.0007	0.0006	0.0002
BIAS1	-0.0002	-0.0266	-0.0017	0.0029	-0.0019	-0.0065	-0.0003	0.0020
BIAS2	0.0011	-0.0252	-0.0005	0.0033	0.0002	-0.0044	-0.0005	0.0015
VAR1	0.0018	0.0015	0.0008	0.0002	0.0003	0.0003	0.0003	0.0001
VAR2	0.0016	0.0014	0.0009	0.0002	0.0004	0.0003	0.0003	0.0001
SIZE1	0.114	0.176	0.080	0.054	0.090	0.104	0.098	0.046
SIZE2	0.104	0.152	0.094	0.048	0.082	0.084	0.090	0.046

Table 3. Monte Carlo Simulation Results for DGP9

MSE of Effects

n	T	Within	GLS	CSSW	KSS	L
30	12	0.1792	0.1781	0.0468	0.0013	4.957
	30	0.1654	0.1653	0.0558	0.0004	5.000
100	12	0.1545	0.1540	0.0427	0.0006	5.000
	30	0.1568	0.1567	0.0600	0.0002	6.000
300	12	0.1823	0.1821	0.0578	0.0003	6.000
	30	0.1904	0.1903	0.0746	0.0003	6.000

MSE, Bias, Variance, and Size for Coefficients

	T=12				T=30			
n=30	Within	GLS	CSSW	KSS	Within	GLS	CSSW	KSS
MSE	13.3517	12.4807	5.3708	0.2495	4.3252	4.2655	1.5867	0.0141
BIAS1	-0.1185	-0.7992	0.0748	0.0261	-0.0112	-0.2617	0.0109	0.0027
BIAS2	0.0952	-0.5965	0.0290	0.0088	-0.0467	-0.2961	-0.0004	0.0003
VAR1	6.8048	5.8531	2.6004	0.1262	2.1655	2.0557	0.8162	0.0071
VAR2	6.5238	5.6330	2.7639	0.1225	2.1574	2.0536	0.7704	0.0069
SIZE1	0.165	0.173	0.172	0.196	0.157	0.157	0.152	0.125
SIZE2	0.156	0.147	0.189	0.197	0.168	0.163	0.144	0.120
n=100	Within	GLS	CSSW	KSS	Within	GLS	CSSW	KSS
MSE	3.8630	3.3662	1.6696	0.0342	1.3015	1.2465	0.5392	0.0021
BIAS1	0.1020	0.0316	0.0866	0.0039	-0.0172	-0.0445	-0.0271	0.0021
BIAS2	0.0578	-0.0123	-0.0087	0.0013	0.0058	-0.0221	0.0096	-0.0007
VAR1	1.9732	1.7301	0.8488	0.0179	0.6622	0.6333	0.2860	0.0011
VAR2	1.8761	1.6350	0.8133	0.0163	0.6390	0.6107	0.2523	0.0011
SIZE1	0.153	0.138	0.172	0.140	0.147	0.141	0.149	0.069
SIZE2	0.136	0.121	0.160	0.119	0.156	0.153	0.128	0.070
n=300	Within	GLS	CSSW	KSS	Within	GLS	CSSW	KSS
MSE	1.3521	1.1517	0.6465	0.0065	0.4562	0.4314	0.1850	0.0009
BIAS1	0.0367	-0.0087	0.0234	0.0069	0.0104	0.0025	0.0186	0.0011
BIAS2	0.0382	-0.0045	0.0008	-0.0025	0.0028	-0.0044	-0.0198	-0.0017
VAR1	0.7071	0.6006	0.3303	0.0031	0.2310	0.2186	0.0891	0.0005
VAR2	0.6423	0.5509	0.3156	0.0034	0.2250	0.2128	0.0952	0.0004
SIZE1	0.182	0.154	0.178	0.132	0.162	0.154	0.136	0.060
SIZE2	0.166	0.152	0.170	0.132	0.168	0.166	0.162	0.054

Table 4. Estimation Results

	Within	BC	CSSW	KSS
CD	-0.0357 (0.0047)	-0.0332 (0.0043)	-0.0095 (0.0032)	-0.0008 (0.0019)
DD	-0.0678 (0.0155)	-0.0244 (0.0124)	-0.0908 (0.0134)	-0.0410 (0.0109)
OD	-0.1451 (0.0097)	-0.1433 (0.0091)	-0.1295 (0.0069)	-0.0440 (0.0200)
lab	-0.1517 (0.0165)	-0.1403 (0.0130)	-0.1639 (0.0139)	-0.1254 (0.0093)
cap	-0.0456 (0.0054)	-0.0523 (0.0048)	-0.0461 (0.0054)	-0.0289 (0.0053)
purf	-0.5522 (0.0208)	-0.6065 (0.0151)	-0.5601 (0.0162)	-0.7598 (0.0268)
ciln	0.1583 (0.0045)	0.1596 (0.0042)	0.1468 (0.0037)	0.1202 (0.0031)
inln	0.3745 (0.0061)	0.3639 (0.0054)	0.3512 (0.0056)	0.3237 (0.0050)
time	0.0154 (0.0009)	0.0023 (0.0013)	-	-
Avg TE	0.4553	0.6111	0.6220	0.6056

Figure 1