Author/s:
Ando, T;Bai, J

Title:
Selecting the regularization parameters in high-dimensional panel data models: Consistency and efficiency

Date:
2018-01-01

Citation:
Ando, T. & Bai, J. (2018). Selecting the regularization parameters in high-dimensional panel data models: Consistency and efficiency. Econometric Reviews, 37 (3), pp.183-211. https://doi.org/10.1080/07474938.2015.1092822.

Persistent Link:
https://hdl.handle.net/11343/268149

Selecting the regularization parameters in high-dimensional panel data models: consistency and efficiency [1]

Tomohiro Ando
University of Melbourne

Jushan Bai
Columbia University

Abstract

This paper considers panel data models in the presence of a large number of potential predictors and unobservable common factors. The model is estimated by the regularization method together with the principal components procedure. We propose a panel information criterion for selecting the regularization parameter and the number of common factors under a diverging number of predictors. Under the correct model specification, we show that the proposed criterion consistently identifies the true model. If the model is instead misspecified, the proposed criterion achieves asymptotically efficient model selection. Simulation results confirm these theoretical arguments.

Key words: factor models, information criterion, penalized method, smoothly clipped absolute deviation (SCAD), heterogeneous coefficients, endogeneity

[1]Tomohieo Ando: Melbourne Business School, University of Melbourne, 200 Leicester Street, Carlton, Victoria 3053, Australia E-mail: T.Ando@mbs.edu

Jushan Bai: Department of Economics, Columbia University, 420 West 118 Street, 1022 International Affairs Building, New York, NY, 10027. E-mail: jb3064@columbia.edu.

The research of Ando is partially supported by the Trust Companies Association of Japan. The research of Bai is partially supported by the National Science Foundation (SES0962410 and SES1357598)

1 Introduction

In recent years, high-dimensional sparse panel data models have received considerable attention. In these models, while the number of predictors can be large, only a relatively small number of them are related to the true regression function in each unit. These models also allow for unobservable common factor structures, capturing the dependence among the cross-sectional units. The factor structure also admits multiple individual effects instead of a single effect. Because the number of regressors for each cross-sectional unit can be large and the number of factors are unknown, we consider the regularization method in selecting the relevant regressors and the number of factors.

It is important to note that the model permits correlations between the observable explanatory variables and the unobservable factor structure. This is important for applications in social sciences where many explanatory variables are decision variables that are correlated with the unobservable effects. Existing studies such as Pesaran (2006) and Bai (2009) only consider a small number of regressors, therefore do not consider regularization method, and the issue of selecting regularization parameter does not arise. These studies also assume correct model specifications.

In this paper, we consider the model evaluation of high-dimensional panel data regressions with a large number of cross section units N, and large number of time periods T, and many regressors:

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta}_i + \boldsymbol{u}_i = X_i\boldsymbol{\beta}_i + F\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, N, \tag{1}$$

where

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{pmatrix}, \; X_i = \begin{pmatrix} \boldsymbol{x}_{i1}' \\ \boldsymbol{x}_{i2}' \\ \vdots \\ \boldsymbol{x}_{iT}' \end{pmatrix}, \; F = \begin{pmatrix} \boldsymbol{f}_1' \\ \boldsymbol{f}_2' \\ \vdots \\ \boldsymbol{f}_T' \end{pmatrix}, \; \boldsymbol{u}_i = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{pmatrix}, \; \boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix},$$

where each $\boldsymbol{x}_{it}$ is a $m_i \times 1$ vector of observable regressors, $\boldsymbol{\beta}_i$ is a $m_i \times 1$ vector of unknown slope coefficients, u_{it} is a noise term and has a factor structure such that $u_{it} = \boldsymbol{\lambda}_i'\boldsymbol{f}_t + \varepsilon_{it}$, where $\boldsymbol{\lambda}_i$ is an $r \times 1$ vector of factor loadings, $\boldsymbol{f}_t$ is an $r \times 1$ vector of common factors, and ε_{it} is idiosyncratic error. Further assumptions are discussed in Section 2.

This model is widely applicable. For example, in finance, let y_{it} be the excess return (ER) of asset i in period t. There is considerable evidence that excess returns

are related to firm-level and market/macroeconomic-level characteristics. Observable firm-specific characteristics for $\boldsymbol{x}_{it}$ include, for example, the dividend-to-price ratio, the earnings-to-price ratio, credit ratings, the market value of equity, and industry indicators, etc. Alternatively, the market/macroeconomic-level characteristics include the ER of the market, the book-to-market factor, a size factor, a momentum factor, a short-term reversal factor, a long-term reversal factor, the volatility index, the liquidity factor, the interest rate spread, crude oil returns, and the foreign exchange rate, etc. See, for example, Chen et al. (1986), Fama and French (1993), and others. At present, quantitative analysts in the asset management industry continue the search for a set of variables relating to ERs, trying to identify predictors in large databases that contain tens of thousands of variables. Therefore, the number of candidates for $\boldsymbol{x}_{it}$ in practice can be enormous. As discussed in Bai (2009), $\boldsymbol{f}_t$ functions here as a vector of unobservable factor returns, $\boldsymbol{\lambda}_i$ is the heterogeneous impact of common shocks on asset i, and ε_{it} is the idiosyncratic return. Obviously, we can also find many similar examples of model (1) in bond, foreign exchange, derivative, and commodity markets, etc. High-dimensional panel data models (1) also play an important role in economics, business, and many other disciplines.

In the literature on high-dimensional cross-sectional regression, a regularization term is attached to an objective function. Following the introduction of the Lasso (Tibshirani, 1996), various regularization methods were developed, for e.g., the smoothly clipped absolute deviation (SCAD) approach (Fan and Li, 2001), the adaptive Lasso (Zou, 2006), the Dantzig selector (Candes and Tao, 2007), the minimax concave penalty approach (Zhang, 2010), and many others. See also Fan and Li (2004), Fan and Peng (2004), Zou (2006), Zou et al. (2007), Yuan and Lin (2006, 2007), Zhang and Lu (2007), Huang et al. (2008), Bai and Ng (2010), Fan and Liao (2013), Ando and Bai (2014), Belloni et al. (2012), Lu and Su (2013).

The regularization approach requires the choice of regularization parameters. Moreover, model (1) also needs to select the number of common factors simultaneously. If there are no predictors in the model (1), then it becomes a factor model. To determine the number of factors in pure factor models, we can apply model selection criteria (e.g. Bai and Ng (2002, 2007)), or employ testing procedure (e.g., Onatski (2009), Ahn and Horenstein (2013)).[2] However, in addition to the number of common factors, we must

[2]In this paper, we assume the number of factors is fixed. The case of increasing number of factors is studied by Li et al. (2014).

also determine the value of the regularization parameter.

One may consider using the cross-validation approach. However, it is not easy to apply cross-validation given the factor structure in the model. First, consider the leave-one-unit-out cross-validation, where one unit is deleted sequentially and the model parameters are estimated based on the information for $N-1$ units. Even we are able to estimate the regression coefficients and the factor structures, we cannot obtain the factor loadings of the deleted unit as the factor loadings are unit dependent. Instead, assume that we estimate the model based on the information observed up to time $t-1$, and then forecast the responses of each unit at time t. It is then possible to obtain the estimators of the regression coefficients, and the factor structures (up to time $t-1$) and the corresponding loadings for each unit. It is also natural to evaluate the predictive ability of the estimated model by comparing the difference between the actual realization of responses at t and their forecasts. However, the factor structure at time t is not available. Thus, the cross-validation procedure is not easy to apply. Moreover, cross-validation will require additional computational time. Even if we overcome these difficulties, it is desirable to have an alternative procedure for panel models with unobservable factor structure and a large number of explanatory variables.

Although this paper shares the same model structure with Ando and Bai (2014), the main theme of this paper is different. In contrast to Ando and Bai (2014), where the focus is on the asymptotic distribution for the estimated regression coefficients and the extracted factor structures, here we focus on the model selections. Ando and Bai (2014) studied a C_p type criterion, which is computationally more demanding, and also assumes correct model specification. In this paper, we propose an alternative approach for model selection and study the asymptotic properties of the proposed panel selection criterion. We consider the asymptotic performance of model selection procedures from two important perspectives: (1) consistency, in the sense that the probability that the true panel data model is identified converges to one when a set of candidate models contains the true model, and (2) asymptotic efficiency, in the sense that the panel data model that asymptotically performs the best among the set of candidate models will be selected when the true model is approximated by a family of candidate models. In the context of cross-sectional regression, it is argued that a model selection procedure cannot be both consistent and efficient (Shao, 1997; Yang, 2005). See also the discussion on consistency and asymptotic efficiency in Shibata (1981, 1984), Li (1987), and Shao

(1997). In this paper, we show that this result is also true in the context of panel data models. The proposed criterion is also simple to compute.

As the first contribution of this paper, we obtain the set of regularity conditions needed to achieve consistent model selection in the context of panel data models. In this situation, given a misspecified model, a reasonable goal is to achieve asymptotic efficiency (Flynn et al., 2013). Our second contribution to the literature is that we develop a panel model selection criterion for efficient model selection. In the context of panel data models with interactive fixed effects (Bai, 2009), Lu and Su (2013) investigate the model selection consistency. In contrast to their analysis, we investigate panel data models with heterogeneous regression coefficients. It is also noteworthy that no existing study investigates the efficiency of model selection in the context of high-dimensional panel data models with common factor error structures, as Lu and Su (2013) considered only the consistency of the model selection problem.

Notation. Let $\|A\| = [tr(A'A)]^{1/2}$ be the usual norm of the matrix A, where "tr" denotes the trace of a square matrix. The equation $a_n = O(b_n)$ states that the deterministic sequence a_n is at most of order b_n, $c_n = O_p(d_n)$ states that the random variable c_n is at most of order d_n in terms of probability, and $c_n = o_p(d_n)$ is of a smaller order in terms of probability. We obtain all asymptotic results under a large number of units N and a long time series T.

The remainder of the paper is organized as follows. Section 2 states the assumptions and gives the estimation procedure. Sections 3 and 4 propose the panel information criterion and establish the consistency and asymptotic efficiency, respectively. Section 5 presents numerical studies that explore the finite-sample behavior of the proposed criterion. Section 6 concludes. Proofs are included in the appendix.

2 High-dimensional panel data model

This paper considers a panel data model with a large number of predictors and a set of common factors.

2.1 Assumptions

Here we state the assumptions needed for the asymptotic analysis. Following this, we explain their meaning.

Assumption A1: Common factors

The common factors satisfy $E\|\boldsymbol{f}_t\|^4 < \infty$. Also $T^{-1}\sum_{t=1}^{T} \boldsymbol{f}_t \boldsymbol{f}_t{}' \to \Sigma_F$ as $T \to \infty$, where Σ_F is an $r \times r$ positive definite matrix.

Assumption A2: Factor loadings

The factor-loading matrix for the common factors $\Lambda = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N]'$ satisfies $E\|\boldsymbol{\lambda}_i^4\| < \infty$ and: $\|N^{-1}\Lambda'\Lambda - \Sigma_\Lambda\| \to \mathbf{0}$ as $N \to \infty$, where Σ_Λ is an $r \times r$ positive definite matrix.

Assumption A3: Error terms

(1): $E[\varepsilon_{it}] = 0$, $\text{var}(\varepsilon_{it}) = \sigma_i^2$, and ε_{it} is independent over i and over t.
(2): There exists a positive constant $C < \infty$ such that $E[|\varepsilon_{it}|^8] < C$ for all i and t.
(3): ε_{it} is independent of $\boldsymbol{x}_{js}$, $\boldsymbol{\lambda}_i$, $\boldsymbol{f}_s$ for all i, j, t, s, g.

Assumption A4: Central limit theorem

Let $X_{i,\beta_i^0 \neq 0}$ be the submatrix of X_i, corresponding to the columns of nonzero elements of the true parameter vector $\boldsymbol{\beta}_i^0$. We assume the following matrix

$$\frac{1}{T}\Big[X'_{i,\beta_i^0 \neq 0} M_{F_0} X_{i,\beta_i^0 \neq 0}\Big] \tag{2}$$

is positive definite, where $M_{F_0} = I - F_0(F_0{}'F_0)^{-1}F_0{}'$ and F_0 is the $T \times r$-dimensional matrix of true common factors. We assume the central limit theory

$$\frac{1}{\sqrt{T}} X'_{i,\beta_i^0 \neq 0} M_{F_0} \boldsymbol{\varepsilon}_i \to_d N(\mathbf{0}, J_i^0),$$

where J_i^0 is the probability limit of (as T goes to infinity) $\frac{1}{T}\sigma_i^2 X'_{i,\beta_i^0 \neq 0} M_{F_0} X_{i,\beta_i^0 \neq 0}$.

Assumption A5: Predictors

Let $A_i = \frac{1}{T} X_i' M_F X_i$, $B_i = (\boldsymbol{\lambda}_i \boldsymbol{\lambda}_i') \otimes I_T$, $C_i = \frac{1}{\sqrt{T}} \boldsymbol{\lambda}_i' \otimes (X_i' M_F)$. Let $\mathcal{A}$ be the collection of F such that $\mathcal{A} = \{F : F'F/T = I\}$. We assume

$$\inf_{F \in \mathcal{A}} \Big[\frac{1}{N}\sum_{i=1}^{N} E_i(F)\Big] \quad \text{is positive definite,} \tag{3}$$

where $E_i(F) = B_i - C_i' A_i^- C_i$ and A_i^- is a generalized inverse of A_i.

Remark 1 The full rank assumption of Σ_F and Σ_Λ in Assumptions A1 and A2 is necessary for the number of common factors to be r. Assumption A3 can be relaxed to allows cross-sectional and serial correlations and heteroskedasticities in the idiosyncratic errors ε_{it}. Theorems 1 and 2 in the following sections hold even if we allow such dependence structure. Assumption A4 requires that the matrix in (2) be positive definite. This assumption is needed for consistent estimation of the regression coefficients, even if F_0 were observable. We also require the central limit theory for the asymptotic normality of the estimated $\boldsymbol{\beta}_i$. In Assumption A5, the matrix in (3) is required to be positive definite. Note that for each i, $E_i(F)$ is semi-positive definite. Matrix (3) involves the sum of N such semi-positive definite matrices, thus it is not a strong requirement. This condition is needed to obtain the consistency for the estimated unobserved factor structures as in Ando and Bai (2014).

2.2 Estimation procedure

Recently, Ando and Bai (2014) considered a panel data model with heterogeneous slope coefficients in contrast to the homogeneous regression coefficients in Bai (2009) by allowing a large number of predictors and attempting to select the set of relevant predictors using the SCAD penalty in Fan and Li (2001). Given the number of common factors, the unknown parameters are estimated by minimizing the least-squares objective function with a penalty term:

$$\ell(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N, F, \Lambda | r, \tau) = \sum_{i=1}^{N} \|\boldsymbol{y}_i - X_i\boldsymbol{\beta}_i - F\boldsymbol{\lambda}_i\|^2 + T\sum_{i=1}^{N} p_{\tau,\gamma}\left(|\boldsymbol{\beta}_i|\right) \tag{4}$$

subject to the constraints $F'F/T = I_r$ and $\Lambda'\Lambda$ being diagonal. We require these restrictions to avoid the model identification problem and they are common in the literature (Connor and Korajczyk, 1986; Bai and Ng, 2002; Stock and Watson, 2002).

The SCAD penalty is defined as $p_{\tau,\gamma}(|\boldsymbol{\beta}_i|) = \sum_{j=1}^{m_i} p_{\tau,\gamma}(|\beta_{ij}|)$ with:

$$p_{\tau,\gamma}(|\beta_{ij}|) = \begin{cases} \tau|\beta_{ij}| & (|\beta_{ij}| \leq \tau) \\ \dfrac{\gamma\tau|\beta_{ij}| - 0.5(\beta_{ij}^2 + \tau^2)}{\gamma - 1} & (\tau < |\beta_{ij}| \leq \gamma\tau) \\ \dfrac{\tau^2(\gamma^2 - 1)}{2(\gamma - 1)} & (\gamma\tau < |\beta_{ij}|) \end{cases},$$

for $\tau > 0$ and $\gamma > 2$. Fan and Li (2001) investigate the theoretical property of the SCAD penalty in the context of nonpanel data.

Similarly to Bai (2009) and Ando and Bai (2014), we can estimate $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N, F, \Lambda\}$ using iterative procedures. Given $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N\}$, we can define the matrix $W = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N)$ of dimension $T \times N$ with: $\boldsymbol{w}_i = \boldsymbol{y}_i - X_i\boldsymbol{\beta}_i$. Then, the original model (1) reduces to $\boldsymbol{w}_i = F\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i$, which implies that W has a pure factor structure. The estimated $\hat{F}$ is $\sqrt{T}$ times the eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix WW'. Given $\hat{F}$, the factor-loading matrix can be obtained as $\hat{\Lambda}' = \hat{F}'W/T$. See Bai and Ng (2002: 197–198). Given the factor structure $\{F, \Lambda\}$, the regression coefficients for $\boldsymbol{\beta}_i$ are easily obtained by running SCAD estimation, treating $y_i - F\lambda_i$ as the dependent variables, for $i = 1, 2, ..., N$. The estimates of $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N\}$ and $\{F, \Lambda\}$ depend on each other, the estimators are obtained using the following iterative algorithm.

Estimation algorithm

Step 1. Fix the regularization parameter τ and the number of common factors r. Initialize the unknown regression coefficients $\{\boldsymbol{\beta}_1^{(0)}, \ldots, \boldsymbol{\beta}_N^{(0)}\}$, the common factors, and the corresponding factor-loading matrix $\{F^{(0)}, \Lambda^{(0)}\}$.

Step 2. Given values of $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N\}$, update $\{F, \Lambda\}$.

Step 3. Given values of $\{F, \Lambda\}$, update $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N\}$.

Step 4. Repeat Steps 2 and 3 until convergence.

In practical implementation, we stop the iteration once $\sum_{i=1}^{N} \|\boldsymbol{\beta}_i^{new} - \boldsymbol{\beta}_i^{old}\|^2 < \delta$ is observed for some absolute convergence tolerance δ. In our simulation study, we set $\delta = 10^{-4}$.

Remark 2 It is natural to consider the following two-step procedure. First, ignoring the unobservable factor structure, for each unit i, the regression coefficients $\boldsymbol{\beta}_i$ is estimated by using SCAD penalized least squares, with y_i being the response and X_i being the predictor matrix. Then, the resulting residual vector $\hat{\boldsymbol{u}}_i$ is used to estimate the unobservable common factors. However, this two-step procedure, in general, does not work due to the endogeneity problem (the regressors are correlated with the factors and factor loadings). Under such circumstances, we have to capture the dependency between the regressors and unobservable factor structures simultaneously. Section 5.3 provides a numerical example showing that the proposed iterative method overcomes the endogeneity problem.

In practice, however, we need to select the size of the regularization parameter τ such that the relevant predictors are included. Moreover, the number of common factors is unknown. In the next section, we introduce a new panel information criterion to select these quantities.

3 Main result 1: consistent model selection

Usually, model selection criteria consist of two terms: a fitness term that measures how well the model explains the observed data and a penalty term that penalizes redundant model flexibility. In this paper, we consider the minimization of the following Panel Information Criterion (PIC)

$$\text{PIC}(k,\tau) = \frac{1}{NT}\sum_{i=1}^{N}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_{i,\tau} - \hat{F}\hat{\boldsymbol{\lambda}}_i\right\|^2 + \text{Penalty}(k,\tau). \tag{5}$$

The remaining task is how to construct a proper penalty term.

Here we employ the following penalty term

$$\text{Penalty}(k,\tau) = \frac{1}{N}\sum_{i=1}^{N}\sigma^2 m_{i,\tau} \times \left(\frac{a_{NT}}{T}\right) + \sigma^2 k \times (h_{NT}), \tag{6}$$

where $m_{i,\tau}$ is the number of nonzero components of the parameter estimate $\hat{\boldsymbol{\beta}}_{i,\tau}$, σ^2 is a proper scaling term, and a_{NT} and h_{NT} are positive numbers that control the properties of the variable and factor selection, respectively. The scaling parameter σ^2 is usually chosen to be $\sigma^2 = \lim_{N,T\to\infty}(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}E[\varepsilon_{it}^2]$. In practice, it is estimated by $\hat{\sigma}^2 = (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\varepsilon}_{it}^2$, where $\hat{\varepsilon}_{it}$ are estimated residuals.

When the true model is among some set of candidate models, it is common to consider model selection that enables us to identify the true model consistently. In this section, we explore the conditions for a_{NT} and h_{NT} for consistent model selection. In practice, however, this assumption may not be valid as the data-generating process is likely to be too complex to know exactly (Flynn et al., 2014). This motivated many studies that investigated the asymptotic efficiency (see Shibata (1981) and Li (1987)). Thus, we examine, in Section 4, the asymptotic efficiency when the model is misspecified.

Let $\alpha_{i,0}$ $(i = 1, ..., N)$ denote the set of true predictors for the i-th unit (and thus, the set of irrelevant predictors are not included). Similarly, we define r as the model with the true number of common factors. Let α_i $(i = 1, ..., N)$ be an arbitrary index set

of predictors of i-th unit, and k be an arbitrary number of common factors. Thus, α_i denotes any combination of the set of predictors $\boldsymbol{x}_i$ for the i-th unit. Then, we express an arbitrary model as

$$\alpha = (\alpha_1, ..., \alpha_N)$$

with the set of predictors $\alpha_1, ..., \alpha_N$. Note that each value of the regularization parameter τ leads to a different model specification

$$\alpha_\tau = (\alpha_{1,\tau}, ..., \alpha_{N,\tau})$$

and we emphasize this dependence on τ with the subscript τ. Hereafter, we refer to any candidate model α_τ with either $\alpha_{i,\tau} \nsubseteq \alpha_{i,0}$ for some i or $k < r$ as an underfitted model in the sense that it omits the true predictors or true common factors. Similarly, any candidate model with $\alpha_{i,0} \subset \alpha_{i,\tau}$ $(i = 1, ..., N)$, $r \leq k$ and either $\alpha_{i,0} \neq \alpha_{i,\tau}$ for some i, or $r < k$ is referred to as an overfitted model in the sense that it contains irrelevant predictors or irrelevant common factors.

Based on the above definitions, we partition the regularization parameter interval $\Omega = [0, \tau_{\max}]$ into the underfitted, true, and overfitted subsets, respectively,

$$\begin{aligned}
\Omega_- &= \{\tau \in \Omega : \alpha_{i,0} \nsubseteq \alpha_{i,\tau} \text{ for some } i\} \\
\Omega_0 &= \{\tau \in \Omega : \alpha_{i,0} = \alpha_{i,\tau}, \text{ for } \forall\ i\} \\
\Omega_+ &= \{\tau \in \Omega : \alpha_{i,0} \subset \alpha_{i,\tau} \text{ for } \forall\ i, \text{ and } \alpha_{i,0} \neq \alpha_{i,\tau} \text{ for some } i\},
\end{aligned}$$

where $\alpha_{i,\tau}$ denotes the set of selected predictors under the value of τ. This partition allows us to assess the performance of the regularization parameter selection methods. These notations are similar to those used by Zhang et al. (2010).

To investigate the asymptotic properties of the PIC, we present the penalty conditions

Assumption A6: Expanding speed of the panel

The following holds $\sqrt{T}/N \to 0$ as $N, T \to \infty$.

Assumption A7: Regularization parameter

The regularization parameter satisfies $\tau \to 0$, and $\sqrt{T/m_{\max}} \times \tau \to \infty$, as $T \to \infty$. Here $m_{\max} = \max\{m_1, ..., m_N\}$.

Assumption A6 is a part of the sufficient condition for establishing the asymptotic normality of the regression coefficients. Assumption A7 implies that the size of the regularization parameter depends on the length of the time series T and the diverging order of $m_{\max}$. Note that we allow $m_{\max}$ to go to infinity, but at much slower rate than root-T. In case that $m_{\max}$ is bounded by some fixed constant, Assumption A7 is replaced by $\sqrt{T} \times \tau \to \infty$, as $T \to \infty$.

We show that for any k and τ which cannot identify the true model, the resulting panel information criterion (PIC) $PIC(k, \tau)$ is asymptotically larger than $PIC(r, \tau_0)$ with $\tau_0 \in \Omega_0$ under certain regularity conditions on the penalty term.

Theorem 1 *Suppose that Assumptions A1 $\sim$ A7 hold, and that the model is correctly specified. If*

$$(a)\ a_{NT} \to \infty \ \text{ and } \ a_{NT}/\sqrt{T} \to 0,$$
$$(b)\ h_{NT} \to 0 \ \text{ and } \ \min\{T, N\} \times h_{NT} \to \infty,$$

then the panel information criterion is a consistent model selector.

The appendix provides a derivation of the theorem. The first term on the right-hand side of Penalty(k, τ) in (6) controls the size of the regularization parameter. The second term is relevant to the identification of the true number of common factors. The proposed PIC with the penalty term in Theorem 1 is a consistent model selector.

An example of the functions a_{NT} and h_{NT} that satisfy conditions (a) and (b) are:

$$a_{NT} = \log(T) \ \text{ and } \ h_{NT} = \left(\frac{T+N}{TN}\right) \log(TN),$$

Function $h_{NT} = (T+N)/(TN)\log(TN)$ is considered in Bai (2009). Also, as discussed in Bai and Ng (2002), examples such as $N = \exp(T)$ or $T = \exp(N)$ are rare situations that will violate the conditions. The choice of $a_{NT} = \log(T)$ is often considered in BIC-type criterion in the cross-sectional regression context (See for example, Hong and Preston (2012)).

Then, we have the following PIC:

$$\begin{aligned} PIC_1(k, \tau) &= \frac{1}{NT} \sum_{i=1}^{N} \left\| \boldsymbol{y}_i - X_i \hat{\boldsymbol{\beta}}_{i,\tau} - \hat{F} \hat{\boldsymbol{\lambda}}_i \right\|^2 \\ &\quad + \frac{1}{N} \sum_{i=1}^{N} \sigma^2 m_{i,\tau} \times \left(\frac{\log(T)}{T}\right) + \sigma^2 k \times \left(\frac{T+N}{TN}\right) \log(TN), \end{aligned} \quad (7)$$

where $m_{i,\tau}$ is the number of selected predictors from X_i, i.e., the size of $\alpha_{i,\tau}$. In practice, σ^2 is unknown and is estimated. In applications, we replace σ^2 with $(NT)^{-1}\sum_{i=1}^{N}\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_{i,\tau} - \hat{F}\hat{\boldsymbol{\lambda}}_i\|^2$, which is obtained under the maximum possible dimension of X_i, the maximum possible number of common factors r_{max}.

We choose the number of common factors, k, and the size of the regularization parameter τ, by minimizing the PIC_1 over a specified range of models.

Remark 3 We develop Theorem 1 by assuming that the true model is among the candidate models. If this model is misspecified, we assume that the observed data is generated from $\boldsymbol{y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\mu}_i$ is the true mean structure. Given the situation where the true model is not among the candidate models, the model that minimizes $\lim_{N,T\to\infty}(NT)^{-1}E[\sum_{i=1}^{N}\|\boldsymbol{y}_i - (X_i\boldsymbol{\beta}_i + F\boldsymbol{\lambda}_i)\|^2]$ is regarded as the true model (the pseudo-true model). For each of the specified models (i.e., given α_τ and k), we can also define the pseudo parameter values $\{\boldsymbol{\beta}_i^*, F^*, \Lambda^*\}$ that minimizes $\lim_{N,T\to\infty}(NT)^{-1}\sum_{i=1}^{N}E[\|\boldsymbol{y}_i - (X_i\boldsymbol{\beta}_i + F\boldsymbol{\lambda}_i)\|^2]$. If the model is misspecified, then $\{\boldsymbol{\beta}_i^*, F^*, \Lambda^*\}$ are considered as the pseudo true parameters. Modifying the proof of Theorem 1 in Ando and Bai (2014), we are still able to prove the consistency of the estimated parameters, i.e, $\|\boldsymbol{\beta}_i^* - \hat{\boldsymbol{\beta}}_{i,\tau}\|^2 = o_p(1)$, and $\|F^* - \hat{F}\|^2/T = o_p(1)$.

In this case, we can weaken the restriction on a_{NT}, and only need $a_{NT}/\sqrt{T} \to 0$ and Condition (b) in Theorem 1 for the consistent model selection. Note that, for any specified model, we obtain

$$\frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_{i,\tau} - \hat{F}\hat{\boldsymbol{\lambda}}_i\|^2 = \frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{y}_i - X_i\boldsymbol{\beta}_i^* - F^*\boldsymbol{\lambda}_i^*\|^2 + o_p(1),$$

where the first term on the right-hand-side equation measures the distance between the true mean $\boldsymbol{\mu}_i$ and the infeasible estimate $X_i\boldsymbol{\beta}_i^* + F^*\boldsymbol{\lambda}_i^*$ from the specified model. If the model is misspecified, the term $\frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{y}_i - X_i\boldsymbol{\beta}_i^* - F^*\boldsymbol{\lambda}_i^*\|^2$ does not vanish. More specifically, the model that achieves the minimum value of $\frac{1}{NT}\sum_{i=1}^{N}\|\boldsymbol{y}_i - X_i\boldsymbol{\beta}_i^* - F^*\boldsymbol{\lambda}_i^*\|^2$ is selected because this is the leading term of PIC in (5). In summary, if the true model is not contained in the set of candidate models, then we only need $a_{NT}/\sqrt{T} \to 0$ for selecting the regularization parameter τ.

Remark 4 The penalty term on the number of predictors in (7) looks like the Bayesian information criterion (BIC). We also point out that diverging functions like $a_{NT} = T^\alpha$ $(0 < \alpha < 1/2)$ also satisfy the condition (a) in Theorem 1. In contrast, Theorem 1

implies that the panel information criterion with bounded a_{NT} tends to select overfitted models. However, because of the regularity conditions for the SCAD penalty and the consistency of the regression coefficients, the set of coefficients on the redundant predictors will become zero asymptotically. Therefore, the final model that minimizes the PIC with bounded a_{NT} asymptotically identifies the true model.

4 Main result 2: Efficiency of the model selection criterion

In the previous section, we considered model selection consistency where the model is correctly specified. In practice, however, the true model may not be included in the set of candidate models, or so-called model misspecification. In this section, instead of considering model selection consistency, we establish the asymptotic efficiency of the PIC in the situation where the model is misspecified.

In previous studies, the L_2 norm has been commonly used to assess the efficiency of the model selection procedure (see Shibata (1981) and Li (1987)). To assess the performance of the PIC, we follow these existing studies and define the average squared loss associated with the estimator

$$L(\hat{\boldsymbol{\mu}}_{1,\tau},...,\hat{\boldsymbol{\mu}}_{N,\tau},k)=\frac{1}{NT}\sum_{i=1}^{N}\left\|\boldsymbol{\mu}_i-X_i\hat{\boldsymbol{\beta}}_{i,\tau}-\hat{F}(k)\hat{\boldsymbol{\lambda}}_i(k)\right\|^2\equiv\frac{1}{NT}\sum_{i=1}^{N}\left\|\boldsymbol{\mu}_i-\hat{\boldsymbol{\mu}}_{i,\tau}\right\|^2,$$

where $\boldsymbol{\mu}_i=X_i\boldsymbol{\beta}_0+F_0\boldsymbol{\lambda}_i$ is the true mean structure, $\hat{\boldsymbol{\mu}}_{i,\tau}=X_i\hat{\boldsymbol{\beta}}_{i,\tau}+\hat{F}(k)\hat{\boldsymbol{\lambda}}_i(k)$ $(i=1,...,N)$, τ and k are the values of the regularization parameter and the number of common factors, respectively. As mentioned, a different value of the regularization parameter τ may lead to a different parameter estimate $\hat{\boldsymbol{\beta}}_{i,\tau}$, which is associated with a different model specification α_τ.

Using the average squared loss measure, we further define the asymptotic efficiency. A PIC is said to be asymptotically efficient if

$$\frac{L(\hat{\boldsymbol{\mu}}_{1,\hat{\tau}},...,\hat{\boldsymbol{\mu}}_{N,\hat{\tau}},\hat{r})}{\inf_{k\in[r_{\min},r_{\max}]}\inf_{\tau\in[0,\tau_{\max}]}L(\hat{\boldsymbol{\mu}}_{1,\tau},...,\hat{\boldsymbol{\mu}}_{N,\tau},k)}\to 1$$

in probability, where $\hat{\boldsymbol{\mu}}_{i,\hat{\tau}}=X_i\hat{\boldsymbol{\beta}}_{i,\hat{\tau}}+\hat{F}(\hat{r})\hat{\boldsymbol{\lambda}}_i(\hat{r})$ $(i=1,...,N)$ are associated with the selected value of the regularization parameter $\hat{\tau}$ and the number of common factors $\hat{r}$.

To establish the asymptotic efficiency, we introduce the nonpenalized estimators $\tilde{\boldsymbol{\beta}}_{i,\alpha_\tau}$. For any model $\alpha_\tau=(\alpha_{1,\tau},...,\alpha_{N,\tau})$ with set of predictors $\alpha_{1,\tau},...,\alpha_{N,\tau}$ and number of common factors k, we are able to obtain the nonpenalized parameter estimators

$\tilde{\boldsymbol{\beta}}_{i,\alpha_\tau}$, $(i = 1, ..., N)$ and $\tilde{F}(k)$, $\tilde{\Lambda}(k)$. Here the set of predictors α_τ corresponds to the selected predictors under a fixed value of the regularization parameter τ. Without loss of generality, we assume that the first $m_{i,\tau}$ components of the penalized estimator $\hat{\boldsymbol{\beta}}_{i,\tau}$ and nonpenalized estimator $\tilde{\boldsymbol{\beta}}_{i,\alpha_\tau}$ are nonzero. The nonpenalized estimator here is known as the "postpenalization estimator" in the literature (such as the post-Lasso). Setting the remaining $m_i - m_{i,\tau}$ components of $\tilde{\boldsymbol{\beta}}_{i,\alpha_\tau}$ as zero, the nonpenalized parameter estimators are obtained by minimizing the least-squares objective function:

$$\ell(\boldsymbol{\beta}_{1,\alpha_\tau}, \ldots, \boldsymbol{\beta}_{N,\alpha_\tau}, F(k), \Lambda(k)) = \frac{1}{NT} \sum_{i=1}^{N} \|\boldsymbol{y}_i - X_i \boldsymbol{\beta}_{i,\alpha_\tau} - F(k)\boldsymbol{\lambda}_i(k)\|^2$$

subject to the constraints $F(k)'F(k)/T = I_k$ and $\Lambda(k)'\Lambda(k)$ being diagonal.

Then, we introduce the following technical conditions.

(B1) For all i, the matrix $A_i^0 = X_i' M_{F_0} X_i / T$, satisfies

$$0 < C_1 < \lambda_{\min}\left\{A_i^0\right\} \leq \lambda_{\max}\left\{A_i^0\right\} < C_2 < \infty, \tag{8}$$

where $\lambda_{\min}\{A_i^0\}$ and $\lambda_{\max}\{A_i^0\}$ are the minimum and maximum eigenvalues of A_i^0, and C_1 and C_2 are some positive constants. Also, the true mean structure satisfies, for some $C_3 > 0$

$$T^{-1} \|\boldsymbol{\mu}_i\|^2 < C_3 < \infty \text{ for } \quad i = 1, ..., N, \tag{9}$$

(B2) Let $\boldsymbol{b}_i = (b_{i1}, ..., b_{im_i})'$, where $b_{ij} = p'_\tau(|\hat{\beta}_{ij}|)\text{sgn}(\hat{\beta}_{ij})$ for all j such that $|\hat{\beta}_{ij}| > 0$, and $b_{ij} = 0$ otherwise, and $\hat{\beta}_{ij}$ is the j-th component of the penalized estimator. In addition, let $\tilde{\boldsymbol{\beta}}_{i,\alpha_\tau}$ be the nonpenalized estimator of $\boldsymbol{\beta}_i$ obtained from model α_τ. Then, we assume that in probability,

$$\sup_\tau \frac{\frac{1}{N}\sum_{i=1}^{N} m_i \|\boldsymbol{b}_i\|^2}{R(\tilde{\boldsymbol{\mu}}_{1,\alpha_\tau}, .., \tilde{\boldsymbol{\mu}}_{N,\alpha_\tau}, k)} \to 0, \tag{10}$$

where m_i is the number of columns of X_i, and

$$R(\tilde{\boldsymbol{\mu}}_{1,\alpha_\tau}, .., \tilde{\boldsymbol{\mu}}_{N,\alpha_\tau}, k) = E\left[L(\tilde{\boldsymbol{\mu}}_{1,\alpha_\tau}, .., \tilde{\boldsymbol{\mu}}_{N,\alpha_\tau}, k)\right],$$

where the expectation is taken with respect to the joint distribution of the future, replicate $\{\boldsymbol{z}_1, ..., \boldsymbol{z}_N\}$ with $\boldsymbol{z}_i = X_i\boldsymbol{\beta}_0 + F_0\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i$, conditioned on the true mean structure $\boldsymbol{\mu}_i = X_i\boldsymbol{\beta}_0 + F_0\boldsymbol{\lambda}_i$.

(B3) Define $m_{\max} = \{m_1, ..., m_N\}$. Then,

$$m_{\max}/T^{1/3} \to 0.$$

Condition (B1) is a common assumption. We assume that the matrix of predictors has full column rank in (B1). In a special case of orthogonal design such that $X_i'X_i/T = I_{m_i}$, then because $X_i'M_{F_0}X_i/T \leq X_i'X_i/T$, it follows that the largest eigenvalue of $X_i'M_{F_0}X_i/T$ is bounded. (B1) also requires that the smallest eigenvalue be bounded away from zero.

Condition (B2) ensures that the difference between the estimated mean structure based on the penalized estimate $\hat{\boldsymbol{\mu}}_{i,\tau}$ and the corresponding nonpenalized estimator $\tilde{\boldsymbol{\mu}}_{i,\alpha_\tau}$ is small in comparison with the risk of the nonpenalized estimator. In the context of cross-sectional regression, Zhang, et al (2010) also placed a similar condition. Similar to Zhang, et al. (2010), the following three are sufficient conditions for (B2).
(S1) $\sqrt{T/m_{\max}} \times \tau_{\max} < C$ for all T and for some constant $C > 0$. Here $\tau_{\max}$ (depending on T) is an upper bound for the regularization parameter.
(S2) For any β_{ij}, $p'_{\tau,\gamma}(\beta_{ij}) \leq \tau \times C$ for some constant $C > 0$.
(S3) The average error of the full model α_{full} satisfies $m_{\max}^{-3}T \times \{(TN)^{-1}\sum_{i=1}^{N}\|\boldsymbol{\mu}_i - H_{i,\alpha_{full}}\boldsymbol{\mu}_i\|^2\} \to \infty$, where $\boldsymbol{\mu}_i$ is the true structure and $H_{i,\alpha_full} = X_{i,\alpha_{full}}(X'_{i,\alpha_{full}}X_{i,\alpha_{full}})^{-1}X'_{i,\alpha_{full}}$ is the projection matrix.

Conditions (S1) and (S2) ensure that the penalized estimator $\hat{\boldsymbol{\beta}}_i$ is consistent estimator for pseudo true parameter in Remark 3. If one sets $\tau_{\max} = O(\sqrt{m_{\max}/T})$, then (S1) is satisfied. The SCAD penalty satisfies (S2) automatically. (S3) requires that the converging speed of $(TN)^{-1}\sum_{i=1}^{N}\|\boldsymbol{\mu}_i - H_{i,\alpha_{full}}\boldsymbol{\mu}_i\|^2$ towards zero should be slower than $O(m_{\max}^3/T)$, which is $o(1)$ from (B3). Consider a case where, for some of units $i \in M \subset \{1, ..., N\}$ such that $\frac{1}{N}\sum_{i=1}^{N} I(i \in M) \geq c > 0$, the full model α_{full} does not include the true predictors $x_{i,exc}$ that are not linear combinations of the rest of the variables. In this case, if the corresponding (nonzero) true coefficients are fixed (i.e., $\beta_{i,exc} = O(1)$ for $i \in M$), then the loss is $(TN)^{-1}\sum_{i=1}^{N}\|\boldsymbol{\mu}_i - H_{i,\alpha_{full}}\boldsymbol{\mu}_i\|^2 = O(1)$ (but not $o(1)$). Consequently, (S3) holds. Thus, the sufficient conditions for (B2) are not difficult to satisfy.

Condition (B3) requires that the diverging magnitude of the possible predictors should not be faster than $T^{1/3}$. This condition is similar to the formulation in Li (1987), who examined the efficiency for cross-sectional regression instead of panel models. However, simulations show that m_i can be much larger, even larger than T. Thus the assumption in (B3) is sufficient, not necessary. But to theoretically allow m_i to grow faster necessitates a new kind of technical argument, which is beyond the scope of this paper.

Theorem 2 *Suppose that Assumptions A1 $\sim$ A6 hold, and that the model is not necessary correctly specified. Under Conditions (B1) $\sim$ (B3), the regularization parameter $\hat{\tau}$ selected by minimizing the PIC yields an asymptotically efficient model selection.*

The following panel information criterion satisfies (B1)$\sim$(B3):

$$\begin{aligned} & PIC_2(k,\tau) \\ = \; & \frac{1}{NT}\sum_{i=1}^{N}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_{i,\tau} - \hat{F}\hat{\boldsymbol{\lambda}}_i\right\|^2 + \frac{1}{N}\sum_{i=1}^{N}\sigma^2 m_{i,\tau}\times\left(\frac{2}{T}\right) + \sigma^2 k\times\left(\frac{2}{T}\right), \quad (11) \end{aligned}$$

where again $m_{i,\tau}$ is the number of nonzero components of the parameter estimate $\hat{\boldsymbol{\beta}}_{i,\tau}$.

In practice, σ^2 is often unknown and we replace the σ^2 by a consistent estimator $\hat{\sigma}^2$. The asymptotic loss efficiency of PIC_2 continues to hold when we replace it with $\hat{\sigma}^2$.

Remark 5 The last two terms of (11) can be replaced by the following

$$\text{Penalty}(k,\tau) = \frac{2}{NT}\sum_{i=1}^{N}\text{tr}\{\Omega_i H_{i,\alpha_\tau,\hat{F}}\} + \frac{2}{NT}\sum_{i=1}^{N}\text{tr}\{\Omega_i H_{\hat{F}}\}, \quad (12)$$

which automatically achieves efficient model selection. Here Ω_i is the covariance matrix of the error term $\boldsymbol{\varepsilon}_i$. If the error terms $\varepsilon_{i,t}$ are homoskedastic, $\Omega_i = \sigma^2 I_T$, then the preceding penalty function coincides with the simplified version employed in (11).

5 Numerical results

In this section, we present numerical results from two Monte Carlo simulations and an empirical application. Example 1 considers the setting in which the true panel data model is in a set of candidate models. This setting allows us to examine the finite sample performance of the proposed PIC_1, which is expected to perform well via Theorem 1. In contrast, Example 2 considers a setting in which the true panel data model is not included in a set of candidate models. Thus, the model is misspecified. This setting enables us to assess the finite sample performance of the efficiency of PIC_2. Further simulation study is conducted in Section 5.3. Section 5.4 examines the impacts of the US subprime crisis and the Great East Japan Earthquake (GEJE) on the daily stock returns of financial firms listed on the Tokyo Stock Exchange. We divide the data span into the following periods: (1) October 1, 2008, to March 31, 2009, representing the period after the failure of Lehman Brothers; and (2) March 11, 2011, to September

10, 2011, representing the period after the GEJE. Using the proposed PIC PIC_1, we analyze these two panels.

5.1 Example 1: Consistency

We first describe the data generating processes. The first data-generating model considered is $\boldsymbol{y}_i = X_i\boldsymbol{\beta}_i + F\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i$, the r-dimensional common factor $\boldsymbol{f}_t$ is a vector of $N(0,1)$ variables, each element of the factor loading matrix Λ follows $N(0,1)$. The N-dimensional vector $\boldsymbol{\varepsilon}_t$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix I_N. The number of columns of X_i is set to $m_i = 40$, while the true number of predictors is $m_{i0} = 4$ for $i = 1, \ldots, N$. Fan and Peng (2004) considers $m_i = 16$ and $T = 800$ in the context of cross-sectional regression. Among 16 predictors, they set 5 predictors (more than 30%) as the true predictors. Thus, our setting is challenging as weaker signal strength are considered. Each of the elements of X_i is generated from the uniform distribution over $[-2, 2]$. The nonzero true parameter values of $\boldsymbol{\beta}_i$ are set to be $(-1, 2, -2, 1.5)$. We put these nonzero elements into the first four elements of $\boldsymbol{\beta}_i$ and thus the true parameter vector is $\boldsymbol{\beta}_i = (-1, 2, -2, 1.5, 0, 0, ..., 0)'$ for $i = 1, ..., N$. The true number of common factors is $r = 3$.

We next investigate the case in which the noise term is nonhomoskedastic and cross-sectionally correlated. The second data-generating model considered is $\boldsymbol{y}_i = X_i\boldsymbol{\beta}_i + F\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i$ and $\varepsilon_{it} = e_{it}^1 + \delta_t e_{it}^2$, where $\delta_t = 1$ if t is odd and is zero if t is even, and the N-dimensional vectors $\boldsymbol{e}_t^1 = (e_{1t}^1, \ldots, e_{Nt}^1)'$ and $\boldsymbol{e}_t^2 = (e_{1t}^2, \ldots, e_{Nt}^2)'$ follow multivariate normal distributions with mean $\mathbf{0}$ and covariance matrix $S = (s_{ij})$, with $s_{ij} = 0.3^{|i-j|}$, and $\boldsymbol{e}_t^1$ and $\boldsymbol{e}_t^2$ are independent. The noise terms are not serially correlated. The common factors and the loading matrices, the design matrix X_i and the true parameter vector $\boldsymbol{\beta}_i$ are generated by the same method as before.

As a third example, we investigate the performance of the proposed method when the idiosyncratic errors have some serial and cross-sectional correlations. The model is $\boldsymbol{y}_i = X_i\boldsymbol{\beta}_i + F\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i$ with $\varepsilon_{it} = e_{it} + 0.2\varepsilon_{i,t-1}$, where $t = 1, \ldots, T$, the N-dimensional vector $\boldsymbol{e}_t = (e_{1t}, \ldots, e_{Nt})'$ follow multivariate normal distributions with mean $\mathbf{0}$ and covariance matrix $S = (s_{ij})$, where $s_{ij} = 0.3^{|i-j|}$. The other variables are as defined earlier.

In these simulation settings, we consider various combinations of N and T, including $(N, T) = (75, 75)$, $(75, 100)$, $(100, 100)$, $(100, 400)$, $(100, 800)$, $(100, 100)$, $(100, 400)$, and $(100, 800)$. Even under $(N, T) = (75, 75)$, $\sqrt{T} < 9$ is much smaller than $N = 75$. Thus,

these settings are close to a situation, where $\sqrt{T}/N \to 0$ is satisfied.

We generated 1,000 replicates using each of the three data-generating models. We then applied the proposed model selection criterion, PIC, to select simultaneously the number of common factors and the size of the regularization parameter τ. We set the possible number of common factors in a range from 0 to $k_{\max} = 10$. Possible candidates for the regularization parameter τ are $\tau = 10^{-4\times(9-j)/9}$, $j = 0, 1, ...9$.

To assess the finite sample performance of the proposed methods, we report the percentage of models that are correctly fitted, underfitted, and overfitted by PIC_1 and PIC_3. PIC_1 is defined in (7) and appears as a BIC-type criterion. Here, as a competitor, we consider PIC_3 defined as

$$\begin{aligned} PIC_3(k,\tau) &= \frac{1}{NT}\sum_{i=1}^{N}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_{i,\tau} - \hat{F}\hat{\boldsymbol{\lambda}}_i\right\|^2 \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\sigma^2 m_{i,\tau} \times \left(\frac{2}{T}\right) + \sigma^2 k \times \left(\frac{T+N}{TN}\right)\log(TN). \end{aligned} \tag{13}$$

which is similar to the Akaike information criterion (AIC) in terms of the penalty on the set of predictors. Note that a_{NT} in PIC_3 does not satisfy the conditions in Theorem 1, while a_{NT} in PIC_1 does, unless $N = \exp(T)$ or $T = \exp(N)$. It may be of interest to investigate the behavior of PIC under these settings.

Tables 1 $\sim$ 3 report the percentages of under, correct, and over identification in the model under the three data-generating models. We also evaluate the identification performance for the true predictors X_i. We measured the identification performance using the true positive rate (TPR) and true negative rate (TNR):

$$\begin{aligned} \text{TPR} &= \frac{\sum_{i=1}^{N}\sum_{k} I\{\hat{\beta}_{ik} \neq 0 \text{ and } \beta_{ik} \neq 0\}}{\sum_{i=1}^{N}\sum_{k} I\{\beta_{ik} \neq 0\}} = \frac{\sum_{i=1}^{N}\sum_{k} I\{\hat{\beta}_{ik} \neq 0 \text{ and } \beta_{ik} \neq 0\}}{N \times 4} \\ \text{TNR} &= \frac{\sum_{i=1}^{N}\sum_{k} I\{\hat{\beta}_{ik} = 0 \text{ and } \beta_{ik} = 0\}}{\sum_{i=1}^{N}\sum_{k} I\{\beta_{ik} = 0\}} = \frac{\sum_{i=1}^{N}\sum_{k} I\{\hat{\beta}_{ik} = 0 \text{ and } \beta_{ik} = 0\}}{N \times (m_i - 4)}, \end{aligned}$$

where $I(\cdot)$ is the indicator function, which takes a value of 1 if it is true and 0 otherwise. As we have the TPR and TNR for each of N securities, we take their averages.

To compare model fittings, we further calculate the following error

$$\text{ME} = \frac{1}{NT}\sum_{i=1}^{N}\left\|\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i\right\|^2$$

where the $\boldsymbol{\mu}_i$ is the true mean structure, and $\hat{\boldsymbol{\mu}}_i = X_i\hat{\boldsymbol{\beta}}_{i,\tau} + \hat{F}\hat{\boldsymbol{\lambda}}_i$ is based on the selected model. Then, we report the mean of the relative model error (MRME), where

the relative model error is defined as

$$\text{RME} = \text{ME}/\text{minME} \tag{14}$$

and minME is the model error obtained by fitting the data with the model that provides the minimum value of model error (ME).

As shown in the tables, both PIC_1 and PIC_3 are capable of selecting the set of true number of common factors. This matches Theorem 1 because the penalty on the number of common factors of these two criteria satisfies the required condition of h_{NT}. However, in terms of the identification of true set of predictors, PIC_1 has a higher probability than PIC_3. Thus, PIC_1 has a much higher possibility of correctly identifying the true model. Table 1 also shows that the MRME of PIC_1 is smaller than that of PIC_3. As the sample size increases, the MRME of PIC_1 approaches that of the oracle estimator faster than that of PIC_3.

5.2 Example 2: Efficiency

In practice, it is likely that the set of candidate models does not include some important predictors. In Example 2, we create such a situation and then assess the performance of our panel data model selection procedure, PIC_2.

The data-generating model considered is $\boldsymbol{y}_i = X_i\boldsymbol{\beta}_i + F\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i$, the r-dimensional common factor $\boldsymbol{f}_t$ is a vector of $N(0,1)$ variables, each element of the factor loading matrix Λ follows $N(0,1)$. The N-dimensional vector $\boldsymbol{\varepsilon}_t$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix I_N. The true number of common factors is $r = 3$.

In addition, we partition the predictor as $\boldsymbol{x}_{it} = (\boldsymbol{x}'_{it,full}, \boldsymbol{x}'_{it,exc})'$, where $\boldsymbol{x}_{it,full}$ contains 20 predictors of the full model and $\boldsymbol{x}_{it,exc}$ is the predictor excluded from model fittings. Accordingly, we partition $\boldsymbol{\beta}_i = (\boldsymbol{\beta}'_{i,full}, \boldsymbol{\beta}'_{i,exc})'$, where $\boldsymbol{\beta}_{i,full}$ is a 20×1 vector and $\boldsymbol{\beta}_{i,exc}$ is a scalar. In the context of cross-sectional regression, a similar setting is employed in Zhang et al. (2010). To investigate the performance of the proposed methods under various parameter structures, we let $\boldsymbol{\beta}_{i,full} = \boldsymbol{\beta}_0 + \gamma\boldsymbol{\delta}/\sqrt{T}$, where $\boldsymbol{\beta}_0 = (-1, 2, -2, 1.5, 0, 0, ..., 0)'$, $\boldsymbol{\delta} = (0, 0, 0, 0, 3, 4, -1, 5, 0, ..., 0)'$, γ ranges from 0 to 10. Also, $\boldsymbol{\beta}_{i,exc}$ is set as 3. Because the candidate model is a subset of the full model that contains 20 predictors in $\boldsymbol{x}_{i,full}$, the above model setting ensures that the true panel data model is not included in the set of candidate models.

We simulate 500 data sets with $T = 100, 200, 400$ under $N = 100$. To study the performance of our PIC, we investigate the finite sample's loss efficiency. We recall that a PIC is said to be asymptotically efficient if

$$LE(\hat{k}, \hat{\tau}) = \frac{L(\hat{\boldsymbol{\mu}}_{1,\hat{\tau}}, ..., \hat{\boldsymbol{\mu}}_{N,\hat{\tau}}, \hat{k})}{\inf_{k\in[r_{\min}, r_{\max}]} \inf_{\tau\in[0,\tau_{\max}]} L(\hat{\boldsymbol{\mu}}_{1,\tau}, ..., \hat{\boldsymbol{\mu}}_{N,\tau}, k)},$$

converges to one. Here $\hat{\boldsymbol{\mu}}_{i,\hat{\tau}} = \hat{\boldsymbol{\beta}}_{i,\hat{\tau}} + \hat{F}(\hat{k})\hat{\boldsymbol{\lambda}}_i(\hat{k})$ $(i = 1, ..., N)$ are associated with the selected value of regularization parameter and the number of common factors $\hat{\tau}$ and $\hat{k}$.

Note that $\boldsymbol{\beta}_{i,full}$ depends on the length of time series and γ. As mentioned in Zhang et al. (2010), a sensible approach to evaluate the loss efficiency of the panel selection criterion across different T is to compare the loss efficiency under the same value of $\boldsymbol{\beta}_{i,full}$. Figure 1 (b) shows the $LE(\hat{k}, \hat{\tau})$ of PIC_2, across various γ. Figure 1 clearly indicates that the loss efficiency of PIC_2 converges to 1 regardless of the value of γ, which corroborates the theoretical finding in Theorem 2. On the other hand, the loss efficiency of PIC_1 (in Figure 2(a)) is inferior to that of PIC_2, which is also consistent with the theory.

5.3 Example 3: Additional results

We further investigate the property of our proposed procedures. First, generating the set of predictors from distributions with heavier tails, further simulations are conducted. Moreover, predictors are allowed to have correlation with the unobservable common factors. Secondly, we investigate the effect of simplified penalty versus the penalty given in Remark 6.

As the fourth data generating process, we consider $\boldsymbol{y}_i = X_i\boldsymbol{\beta}_i + F\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i$, the r-dimensional common factor $\boldsymbol{f}_t$ is a vector of $N(0, 1)$ variables, each element of the factor loading matrix Λ follows $N(0, 1)$. The first and third elements of predictors have correlation with $\boldsymbol{f}_t\boldsymbol{\lambda}_i$ as

$$x_{i1,t} = 0.2\boldsymbol{f}_t'\boldsymbol{\lambda}_i + v_{i1,t}, \quad \text{and} \quad x_{i3,t} = 0.05\boldsymbol{f}_t'\boldsymbol{\lambda}_i + v_{i3,t},$$

where $v_{i1,t}$ and $v_{i3,t}$ are generated from Student-t distribution with 5 degrees of freedom. Other elements of X_i also follows Student-t distribution with 5 degrees of freedom. The T-dimensional noise vectors $\boldsymbol{\varepsilon}_i$ follow multivariate normal distributions with mean $\mathbf{0}$ and covariance matrix $S = (s_{ij})$, with $s_{ij} = 0.3^{|i-j|}$. The noise terms are not cross-sectionally correlated. The other variables are the same as in the first data

generating process in Section 5.1. Table 4 reports the performance of PIC. Under various combinations of N and T, model selection is conducted over 1,000 repetitions. As shown in the table, both PIC_1 and PIC_3 still detect the true number of common factors. In terms of the variable selection performance, PIC_1 has a higher probability than PIC_3.

Using the penalty function (12), we can also consider PIC_4 defined as

$$\begin{aligned} PIC_4(k,\tau) &= \frac{1}{NT}\sum_{i=1}^{N}\left\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_{i,\tau} - \hat{F}\hat{\boldsymbol{\lambda}}_i\right\|^2 \\ &\quad \frac{2}{NT}\sum_{i=1}^{N}\mathrm{tr}\{\hat{\Omega}_i H_{i,\alpha_\tau,\hat{F}}\} + \frac{2}{NT}\sum_{i=1}^{N}\mathrm{tr}\{\hat{\Omega}_i H_{\hat{F}}\}, \end{aligned} \tag{15}$$

where $\hat{\Omega}$ is an estimator of Ω_i. Here, we compare the performance of PIC_2 and PIC_4. Data sets are generated from $\boldsymbol{y}_i = X_i\boldsymbol{\beta}_i + F\boldsymbol{\lambda}_i + \boldsymbol{\varepsilon}_i$ and $\varepsilon_{it} = 0.7e_{it,1} + 0.9u_{i2}s_{it,2}$, where $e_{it,1}$ and $e_{it,2}$ follow the normal distributions with mean 0 and standard deviation u_{i1}, and Student-t distribution with 5 degrees of freedom. Here u_{i1} and u_{i2} are generated from uniform $[1,3]$. Thus, the error term exhibits heterogeneity. Ω_i is thus estimated by $\hat{\Omega}_i = \hat{\sigma}_i^2 I$ with $\hat{\sigma}_i^2 = \sum_{t=1}^{T}\|\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}_i - \hat{F}\hat{\boldsymbol{\lambda}}_i\|^2/(T-\hat{p}_i)$ with $\hat{p}_i$ is the number of selected predictors. The other variables are as defined earlier. We consider various combinations of N and T, including $(N,T) = (75,75)$, $(75,100)$, $(100,100)$, $(100,400)$, $(100,800)$, $(100,100)$, $(100,400)$, and $(100,800)$. As a result, we found that the best model selected by PIC_2 and PIC_4 are identical more than 99% cases. Thus, the simple penalty in (6) with the condition (C4) is useful for achieving efficient model selection.

5.4 Analysis of Tokyo Stock Exchange

We employ the daily stock returns on the Tokyo Stock Exchange to compare the impact of the Lehman Brothers' failure, resulting from the US subprime crisis, and the the Great East Japan Earthquake (GEJE) on firms listed on the exchange. We divide the data into two subperiods: (1) October 1, 2008, to March 31, 2009, representing the period after Lehman Brothers' failure ($T = 120, N = 1097$); and (2) March 11, 2011, to September 10, 2011, representing the period after the GEJE ($T = 124, N = 1097$). We removed firms with missing stock returns in each of these periods.

In this application, we would like to capture the unobservable market structure by controlling the predictors. For the choice of $\boldsymbol{x}_{it}$, we use some macroeconomic variables along with some well-known financial factors as predictors. Fama and French (1993)

constructed three financial factors, one based on size (small minus big; SMB), another based on the book-to-market ratio (high minus low; HML), and the third based on ER on the market. It is well known that on a daily basis, the Japanese stock market often moves in the same direction as the US stock market on the previous day. Therefore, it would be useful to include the Standard & Poor's 500 index (SP500). The return on long-term Japanese government bonds (JGB) is a suitable proxy variable for the economic outlook. Crude oil prices (OIL) act as a proxy variable for commodity prices and are a major cost factor in various economic activities in Japan. Currency movements also directly affect the earnings of Japanese companies, so we consider the exchange rate between the Japanese yen and the US dollar (YUSD). In addition, the exchange rate may lead to changes in monetary policy that gives rise to changes in consumer behavior and capital flows to the stock market. To consider trade with Europe, we include the exchange rate between the Japanese yen and the euro (YEURO).

We use the following model to analyze the daily stock returns

$$\begin{aligned} y_{it} &= \beta_{0i} + ER_t \times \beta_{i1} + HML_t \times \beta_{i2} + SMB_t \times \beta_{i3} + SP500_t \times \beta_{i4} + JGB_t \times \beta_{i5} \\ &\quad + OIL_t \times \beta_{i6} + YUSD_t \times \beta_{i7} + YEURO_t \times \beta_{i8} + \boldsymbol{f}_t' \boldsymbol{\lambda}_i + \varepsilon_{i,t}, \end{aligned}$$

where SMB_t, HML_t, ER_t, $SP500_t$, JGB_t, OIL_t, $YUSD_t$, and $YEURO_t$ are their values at time t.

Table 5 details the estimated number of factors and the fraction of predictors that are included in the model for each sample period. In this application, the number of common factors varies up to $r_{\max} = 10$. In addition, the set of candidate values of regularization parameter is set as $\tau = \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. Then the number of common factors and the value of regularization parameter are selected by using our proposed criterion PIC_1. The percentage is calculated as $\sum_{i=1}^{N} I\{\hat{\beta}_{ik} \neq 0\}/N$, where $I(\cdot)$ is the indicator function, which takes a value of 1 if it is true and 0 otherwise, and N is the number of stocks.

Table 5 indicates that the ER on the market has explanatory power for the fluctuations of all stocks during period 1. Thus, ER, as used in the well-known capital asset pricing model is very important for explaining the variation in stock return fluctuations. Although ER is not selected in all individual stock returns, we found that the extracted first factor is strongly related to ER. Thus, ER is important for each of the periods. We can also see that the impact of HML and SMB during period 1 and period 2 appears to differ. However, once again, we found that the extracted

factors are strongly related to these two predictors. Interestingly, the impact of the US stock market (SP500) after the GEJE was smaller. This implies that the market participants considered the US stock market more, even with the failure of Lehman Brothers, whereas they placed more emphasis on the other factors in the aftermath of the GEJE. Overall, the impact of most of the macroeconomic predictors decreased after the earthquake, as the percentage of predictors included in the model decreased.

Table 5 also reports the number of factors estimated using criterion PIC_1. From the table, we can see that the number of factors is zero after the outbreak of the subprime crisis. To interpret the common factors in period 2, we regress these factors on the set of eight predictors. As a result, we found that the first factor is related to ER and SMB at the 5% significance level. Similarly, the second factor is related to JGB, SMB, and HML, the third factor is explained by ER, YEURO, SMB, and HML. Combined with the frequency of the percentage of predictor selection, this empirical evidence indicates that the structure of the Tokyo stock market after the subprime crisis and GEJE is different.

6 Conclusion

In this paper, we proposed PIC to select the values of the regularization parameters for the SCAD penalty and a number of common factors. Furthermore, we considered the theoretical properties of the proposed criterion. If the true model is contained in a set of candidate models, then our PIC_1 criterion identifies the true model with a probability tending to unity. On the other hand, if the true model is approximated by a family of candidate models, the proposed PIC_2 is asymptotically loss efficient. Simulation studies confirmed the finite sample performance of the selection criteria.

References

Ahn, S., Horenstein, A. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* 81:1203–1227.

Ando, T., Bai, J., (2014). Asset pricing with a general multifactor structure *Journal of Financial Econometrics*, forthcoming. doi: 10.1093/jjfinec/nbu026

Bai, J. (2003). Inferential theory for factor models of large dimensions, *Econometrica* 71:135–173.

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

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

Bai, J., Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics* 25:52–60.

Bai, J., Ng, S. (2010). Instrument variable estimation in a data rich environment. *Econometric Theory* 26: 1577-1606

Belloni, A., Chernozhukov, V., Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80:2369–2429.

Candes, E., Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* 35:2313–2404.

Chen, N.-F., Roll, R., Ross, S.A. (1986). Economic forces and the stock market. *Journal of Business* 59:383–403.

Connor, G., Korajczyk, R. (1986). Performance measurement with the arbitrage pricing theory: a new framework for analysis. *Journal of Financial Economics* 15:373–394.

Fama, E. F., French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:3–56.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96:1348–1360.

Fan, J., Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, *Journal of the American Statistical Association* 99:710–723.

Fan, J., Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters, *Annals of Statistics* 32:928–961.

Fan, J., Liao, Y. (2013). Ultra high-dimensional variable selection with endogenous covariates. *Annals of Statistics*, forthcoming.

Flynn, C.J., Hurvich, C.M., Simonoff, J.S. (2013). Efficiency for regularization parameter selection in penalized likelihood estimation of misspecified models. *Journal of the*

American Statistical Association 108:1031–1043.

Hong, H., Preston, B. (2012). Bayesian averaging, prediction and nonnested model selection. *Journal of Econometrics* 167:358–369.

Huang, J., Horowitz, J.L., Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics* 36:587–613.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics* 15:958–975.

Li, Q., Li, H., Shi, Y. (2014). Determining the Number of Factors When the Number of Factors Can Increase with Sample Size. Unpublished manuscript, Department of Economics, Texas A&M University.

Lu, X., Su, L. (2013). Shrinkage estimation of dynamic panel data models with interactive fixed effects. Working paper.

Moon, H., Weidner, M. (2013). Linear regression for panel with unknown number of factors as interactive fixed effects. Working paper, Department of Economics, University College London.

Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models. *Econometrica* 77:1447–1479.

Pesaran, M.H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* 74:967–1012.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* 7:221–264.

Shibata, R. (1981). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8:147–164.

Shibata, R. (1984). Approximation efficiency of a selection procedure for the number of regression variables. *Biometrika* 71:43–49.

Tibshirani, R. (1996). Regression shrinkage and selection via Lasso. *Journal of the Royal Statistical Society* B58: 267–288.

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

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* 10:1040–1053.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92:937–950.

Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society* B68:49–67.

Yuan, M., Lin, Y. (2007). On the non-negative garrote estimator. *Journal of the Royal*

Statistical Society B69:143–161.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 38:894–942

Zhang, H. H., Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* 94:691–703.

Zhang, Y., Li, R., Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* 105:312–323.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* 101:1418–1429.

Zou, H., Hastie, T., Tibshirani, R. (2007). On the degrees of freedom of the LASSO, *Annals of Statistics* 35:2173–2192.

Table 1: Model selection performance of PIC_1 in (7) and PIC_3 in (13) under the first data generating process. Reported are the percentages (%) of under-estimation (U), correct-estimation (C) and over-estimation (O) of the number of factors with 1,000 replications. If the proposed method identifies the true number of common factors 1,000 times out of 1,000 trials, the corresponding three columns with respect to r become 0.00, 1.00, and 0.00 under U, C, and O, respectively. Also reported are the true positive rate (TPR) and true negative rate (TNR) for the predictors. TPR and TNR are averages taken over i ($i = 1, .., N$) for 1,000 replications. MRME is the mean relative model error, defined in (14).

	T	N	U	C	O	TPR	TNR	MRME
PIC_1	75	75	0.00	0.76	0.24	1.00	0.41	3.04
	75	100	0.00	0.92	0.08	1.00	0.43	3.05
	100	100	0.00	1.00	0.00	1.00	0.48	3.20
	100	400	0.00	1.00	0.00	1.00	1.00	1.01
	100	800	0.00	1.00	0.00	1.00	1.00	1.00
	200	100	0.00	1.00	0.00	1.00	0.92	1.12
	200	400	0.00	1.00	0.00	1.00	1.00	1.00
	200	800	0.00	1.00	0.00	1.00	1.00	1.00
PIC_3	75	75	0.00	0.83	0.17	0.99	0.16	3.39
	75	100	0.00	0.97	0.03	1.00	0.18	3.46
	100	100	0.00	1.00	0.00	1.00	0.46	3.35
	100	400	0.00	1.00	0.00	1.00	0.66	1.49
	100	800	0.00	1.00	0.00	1.00	0.83	1.09
	200	100	0.00	1.00	0.00	1.00	0.46	3.68
	200	400	0.00	1.00	0.00	1.00	0.73	1.61
	200	800	0.00	1.00	0.00	1.00	0.97	1.03

Table 2: Model selection performance of PIC_1 in (7) and PIC_3 in (13) under the first data generating process. Reported are the percentages (%) of under-estimation (U), correct-estimation (C) and over-estimation (O) of the number of factors with 1,000 replications. If the proposed method identifies the true number of common factors 1,000 times out of 1,000 trials, the corresponding three columns with respect to r become 0.00, 1.00, and 0.00 under U, C, and O, respectively. Also reported are the true positive rate (TPR) and true negative rate (TNR) for the predictors. TPR and TNR are averages taken over i ($i = 1, .., N$) for 1,000 replications. MRME is the mean relative model error, defined in (14).

	T	N	U	C	O	TPR	TNR	MRME
PIC_1	75	75	0.00	1.00	0.00	1.00	0.43	3.64
	75	100	0.00	1.00	0.00	1.00	0.45	3.72
	100	100	0.00	1.00	0.00	1.00	0.51	3.32
	100	400	0.00	1.00	0.00	1.00	1.00	1.01
	100	800	0.00	1.00	0.00	1.00	1.00	1.00
	200	100	0.00	1.00	0.00	1.00	0.97	1.08
	200	400	0.00	1.00	0.00	1.00	1.00	1.01
	200	800	0.00	1.00	0.00	1.00	1.00	1.00
PIC_3	75	75	0.00	1.00	0.00	1.00	0.41	3.65
	75	100	0.00	1.00	0.00	1.00	0.43	3.73
	100	100	0.00	1.00	0.00	1.00	0.45	3.55
	100	400	0.00	1.00	0.00	1.00	0.69	1.39
	100	800	0.00	1.00	0.00	1.00	0.84	1.11
	200	100	0.00	1.00	0.00	1.00	0.45	3.78
	200	400	0.00	1.00	0.00	1.00	0.70	1.53
	200	800	0.00	1.00	0.00	1.00	0.91	1.08

Table 3: Model selection performance of PIC_1 in (7) and PIC_3 in (13) under the first data generating process. Reported are the percentages (%) of under-estimation (U), correct-estimation (C) and over-estimation (O) of the number of factors with 1,000 replications. If the proposed method identifies the true number of common factors 1,000 times out of 1,000 trials, the corresponding three columns with respect to r become 0.00, 1.00, and 0.00 under U, C, and O, respectively. Also reported are the true positive rate (TPR) and true negative rate (TNR) for the predictors. TPR and TNR are averages taken over i ($i = 1, .., N$) for 1,000 replications. MRME is the mean relative model error, defined in (14).

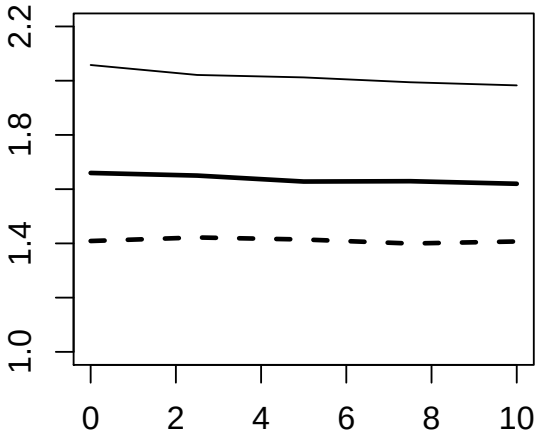
	T	N	U	C	O	TPR	TNR	MRME
PIC_1	75	75	0.00	0.59	0.41	1.00	0.43	2.93
	75	100	0.00	0.83	0.17	1.00	0.44	3.00
	100	100	0.00	1.00	0.00	1.00	0.46	3.37
	100	400	0.00	1.00	0.00	1.00	1.00	1.01
	100	800	0.00	1.00	0.00	1.00	1.00	1.00
	200	100	0.00	1.00	0.00	1.00	0.88	1.25
	200	400	0.00	1.00	0.00	1.00	1.00	1.01
	200	800	0.00	1.00	0.00	1.00	1.00	1.00
PIC_3	75	75	0.00	0.64	0.36	1.00	0.16	3.31
	75	100	0.00	0.95	0.05	1.00	0.22	3.30
	100	100	0.00	1.00	0.00	1.00	0.43	3.45
	100	400	0.00	1.00	0.00	1.00	0.65	1.57
	100	800	0.00	1.00	0.00	1.00	0.84	1.10
	200	100	0.00	1.00	0.00	1.00	0.44	3.56
	200	400	0.00	1.00	0.00	1.00	0.71	1.52
	200	800	0.00	1.00	0.00	1.00	0.96	1.07

Table 4: Model selection performance of PIC_1 in (7) and PIC_3 in (13) under the fourth data generating process. Reported are the percentages (%) of under-estimation (U), correct-estimation (C) and over-estimation (O) of the number of factors with 1,000 replications. If the proposed method identifies the true number of common factors 1,000 times out of 1,000 trials, the corresponding three columns with respect to r become 0.00, 1.00, and 0.00 under U, C, and O, respectively. Also reported are the true positive rate (TPR) and true negative rate (TNR) for the predictors. TPR and TNR are averages taken over i ($i = 1, .., N$) for 1,000 replications. MRME is the mean relative model error, defined in (14).

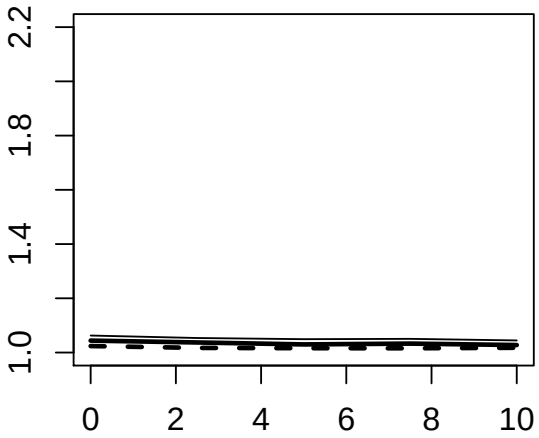
	T	N	U	C	O	TPR	TNR	MRME
PIC_1	75	75	0.00	0.53	0.47	1.00	0.44	3.77
	75	100	0.00	0.91	0.09	1.00	0.43	3.83
	100	100	0.00	1.00	0.00	1.00	0.62	2.51
	100	400	0.00	1.00	0.00	1.00	1.00	1.00
	100	800	0.00	1.00	0.00	1.00	1.00	1.00
	200	100	0.00	1.00	0.00	1.00	1.00	1.00
	200	400	0.00	1.00	0.00	1.00	1.00	1.00
	200	800	0.00	1.00	0.00	1.00	1.00	1.00
PIC_3	75	75	0.00	0.63	0.37	1.00	0.20	4.09
	75	100	0.00	0.94	0.06	1.00	0.39	3.88
	100	100	0.00	1.00	0.00	1.00	0.45	3.46
	100	400	0.00	1.00	0.00	1.00	0.44	4.09
	100	800	0.00	1.00	0.00	1.00	0.44	4.21
	200	100	0.00	1.00	0.00	1.00	0.53	2.40
	200	400	0.00	1.00	0.00	1.00	0.53	3.01
	200	800	0.00	1.00	0.00	1.00	0.53	3.17

Table 5: Application to the Tokyo financial market: the estimated number of factors $\hat{k}$ and the fraction of predictors that are selected for each sample period.

Variables		Period 1	Period 2
Regularization parameter	$\hat{\tau}$	0.1	0.1
Number of common factors	$\hat{k}$	0	3
ER	$N^{-1}\sum_{i=1}^{N} I(\hat{\beta}_{i,1} \neq 0)$	1.000	0.347
SMB	$N^{-1}\sum_{i=1}^{N} I(\hat{\beta}_{i,2} \neq 0)$	0.855	0.666
HML	$N^{-1}\sum_{i=1}^{N} I(\hat{\beta}_{i,3} \neq 0)$	0.851	0.404
SP500	$N^{-1}\sum_{i=1}^{N} I(\hat{\beta}_{i,4} \neq 0)$	0.694	0.511
JGB	$N^{-1}\sum_{i=1}^{N} I(\hat{\beta}_{i,5} \neq 0)$	0.725	0.462
OIL	$N^{-1}\sum_{i=1}^{N} I(\hat{\beta}_{i,6} \neq 0)$	0.654	0.393
YUSD	$N^{-1}\sum_{i=1}^{N} I(\hat{\beta}_{i,7} \neq 0)$	0.655	0.470
YEURO	$N^{-1}\sum_{i=1}^{N} I(\hat{\beta}_{i,8} \neq 0)$	0.689	0.522



(a): The loss efficiency of PIC_1.



(b): The loss efficiency of PIC_2.

Figure 1: The loss efficiency (LE) of PIC_1 and PIC_2 under model misspecification. Horizontal and vertical axises are γ and LE, respectively. Parameter γ is defined in Section 5. Changes in γ imply changes in regression coefficients $\boldsymbol{\beta}_i$. Solid line (—) $T = 50$, Bold line (—) $T = 100$, and Dashed line (- - -) $T = 400$.