Business School / School of Economics

UNSW Business School Working Paper

UNSW Business School Research Paper No. 2017 ECON 12

Constrained Principal Components Estimation of Large Approximate Factor Models

Rachida Ouysse

Constrained principal components estimation of large approximate factor models

Rachida Ouysse*

April 18, 2017

Abstract

Principal components (PC) are fundamentally feasible for the estimation of large factor models because consistency can be achieved for any path of the panel dimensions. The PC method is however inefficient under cross-sectional dependence with unknown structure. The approximate factor model of Chamberlain and Rothschild [1983] imposes a bound on the amount of dependence in the error term. This article proposes a constrained principal components (Cn-PC) estimator that incorporates this restriction as external information in the PC analysis of the data. This estimator is computationally tractable. It doesn't require estimating large covariance matrices, and is obtained as PC of a regularized form of the data covariance matrix. The paper develops a convergence rate for the factor estimates and establishes asymptotic normality. In a Monte Carlo study, we find that the Cn-PC estimators have good small sample properties in terms of estimation and forecasting performances when compared to the regular PC and to the generalized PC method (Choi [2012]).

1 Introduction

Factor models constitute the dominant framework across many disciplines for realistic parsimonious representation of the dynamic behavior of large panels of time series. Principal components estimators (PCEs) of the common factors can be easily computed in panels where the cross-sectional dimension N is large and possibly larger than the sample size T. PCEs are feasible for any path of the panel dimensions and are consistent for both N and T going to infinity (Forni et al. [2009, 2005, 2004]; Bai [2003]; Bai and Ng [2003]; Stock and Watson [2002a,b]). However, principal components are not efficient in the presence of heteroscedasticity or dependence in the error term. Methods based on maximum likelihood (ML) and generalized principal components (GLS) type methods depend on estimating a high-dimensional covariance matrix,

*School of Economics, The University Of New South Wales, Sydney 2052 Australia. Email: rouysse@unsw.edu.au. Preliminary. Please do not circulate.

which is a challenging problem in large systems ($N > T$) when errors are dependent and heteroscedastic.

In this article, we remedy this situation and propose an estimation method that (i) incorporates the dependence features of the data, (ii) doesn't require estimating a large covariance matrix, and (iii) is computationally tractable to ML. The approximate factor model allows dependence in the errors both in the time and cross-section dimensions. This dependence is however bounded by the assumptions in Chamberlain and Rothschild [1983]. The central idea of this article is to incorporate these assumptions as external information into the principal components analysis of the data. This amounts to imposing a bound on the covariances between the errors as a constraint in the principal components framework. The proposed constrianed PC (Cn-PC) estimator can be computed by performing singular value decomposition of a regularised data covariance matrix. This estimator has a dual that can be cast as a penalized PC estimation.

This article is related to a large literature on factor models and a much smaller literature on estimation when N is large and the errors are cross-sectionally dependent. Common factors can be consistently estimated using principal components or maximum likelihood (ML). The fundamental result in the literature is that common factors can be consistently estimated for both N and T going to infinity, with no restrictions on the relative rates of convergence, and under fairly general conditions on the time and cross-sectional dependence of the errors, (Stock and Watson [1998, 2002a,b, 2006]; Bai and Ng [2002]; Bai [2003]; Kapetanios [2010]; Onatski [2010]). The idiosyncratic error component in these studies is homoscedastic and independent. The ML estimation provides a natural framework to account for heteroscedasticity and temporal dependence(Forni et al. [2004]; Forni et al. [2009]). Doz et al. [2012] establish the properties of maximum likelihood estimators for factor models in large panels of time series under heteroscedasticity. Breitung and Tenhofen [2011] propose a two-step generalized least squares estimation that generalized principal components to account for heteroscedasticity and serial correlation in a dynamic factor model with possibly large N. Choi [2012] considers efficient estimation using generalized least squares type PCEs to account for heteroscedasticity and dependence but the framework is only applicable to panels with small N.

We are not aware of substantive econometric research which has looked at the issue of efficiency of PC estimators under cross-sectional dependence when N is large. Does this thin literature reflect a lack of potential and practical relevance for the suggested estimator? The answer is no. First, a decade or so ago, Boivin and Ng [2006] documented through extenssive simulation analysis the potential effects of the presence of dependence on the PC estimators. They find that 'Weighting the data by their properties when constructing the factors also lead to improved forecasts', an that with cross correlated errors the estimated factors may be less useful for forecasting when more series are available. Second, cross-sectional dependence is a likely feature of the data in many applications. Let us consider the example of forecasting inflation and industrial production for the United States. A dataset that is widely used is provided by Stock and Watson [2006] and contains monthly observations from 132 macroeconomic series including real variables such as sectorial industrial production, employment and

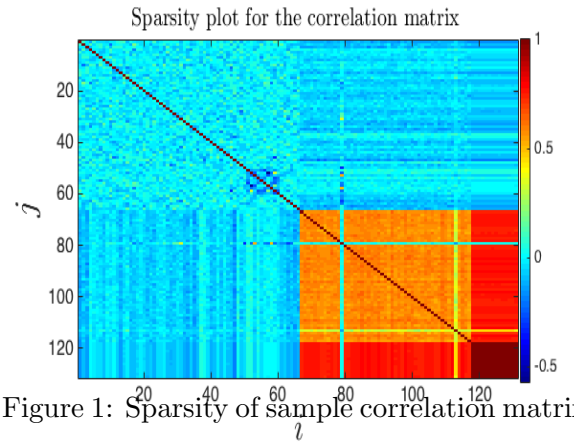


Figure 1: Sparsity of sample correlation matrix

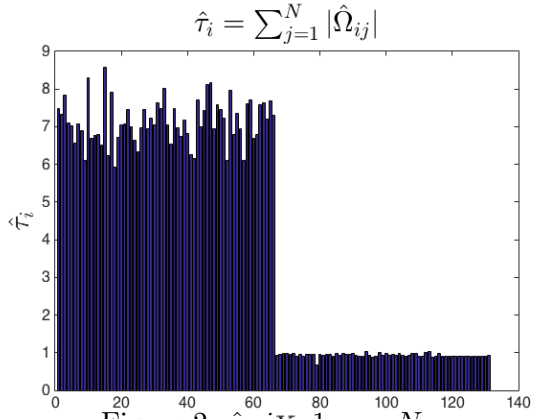


Figure 2: $\hat{\tau}_i, i = 1, \cdots, N$

hours worked; nominal variables such as consumer and price indexes, wages, and money aggregates; and stock prices and exchange rates. The series are tested for unit root and necessary transformations are carried out to achieve stationarity. Assuming a factor structure for the dataset, We estimate a factor model using principal components (see details of estimation in next section) and use Bai and Ng [2003] information criteria to estimate the number of common factors. The residuals from the estimation are then used to compute the sample covariance matrix and sample correlation matrix. Figure (1) is a heat map plot of the sample correlation matrix with each pixel representing the sample correlation coefficient between series i and j for $i, j = 1, \cdots, N(= 123)$. The errors correlation matrix is not diagonal, many of the predictors are highly correlated and the correlations are clustered. The estimated dependence does not have a recognizable pattern or structure. Figure (2) displays sample sums (denoted $\hat{\tau}_i$) of the i^{th} row of the sample covariance matrix, for instance, $\hat{\tau}_i = \sum_{j=1}^{N} |\hat{\Omega}_{ij}|$, where $\hat{\Omega}_{ij}$, is the sample covariance between series i and j, Boivin and Ng [2006]. The amount of cross-sectional dependence varies across groups of predictors and the relative differences of these numbers are large. However, without imposing a structure on the cross-sections dependence (as is the case for time dependence), incorporating these features of the data poses distinct methodological and computational challenges.

To our knowledge, Bai and Liao [2016] is the most relevant study to this article. Bai and Liao [2016] propose ML estimation with penalization of a large covariance sparse matrix. The method produces joint estimates of the factors, the loadings and the covariance matrix and is shown to be more efficient than PC estimators (PCEs) or GLS type PCEs. Bai and Liao [2016] paper is related to a growing literature on estimating large covariance matrices. See among others, Ledoit and Wolf [2004], Ledoit and Wolf [2012] and Lam and Fan [2009a]. Advances in matrix theory have opened a new line of research into consistent estimation of large matrices. Once a consistent estimate of the covariance matrix is achieved, a GLS type estimation or ML can be implemented in a two-step plug-in estimation approach. Bai and Liao [2016] presented results for both a two-step and a joint estimator. The Sparsity assumption of the errors covariance matrix requires many off-diagonal elements to be zero or nearly zero. This assumption is stronger than the weak cross correlation of Chamberlain and Rothschild [1983].

The objective of this article is methodological and practical. This article proposes a novel PC-based estimation of factors in systems with large N (possibly larger than T), and where the errors are cross-sectionally dependent. The suggested estimator solves the PCEs problem under a constraint derived from the assumption of bounded dependence in the sense of Chamberlain and Rothschild [1983]. This constrained system can be solved using the method of principal components. The constrained estimators, denoted Cn-PC estimator, are obtained by performing eigenvalue decomposition to a regularized data covariance matrix. The constrained estimation has a dual problem that can be cast as shrinkage estimation where the regularization is applied to the cross-sectional correlations in the data. The asymptotic properties of the Cn-PC estimators of the factors are derived using the existing techniques of Bai and Ng [2002], Bai [2003], and Choi [2012]. We derive a convergence rate for the Cn-PC estimators of the common factors and show asymptotic normality. The Cn-PC estimator is computationally more attaractive (than ML-based estimators) because the estimation doesn't require, (i) explicit assumption about the structure of sparsity of the covariance matrix, or (ii) estimating and inverting large covariance matrices.

In small samples, Monte Carlo simulations suggest that the Cn-PC has improved accuracy compared to the PCEs, and to GLS-type estimators. Applyied to the problem of forecasting U.S. inflation and industrial production using the 'diffusion indexes' framework of Stock and Watson [2002a], we find relative improvement in accuarcy. However, the gains are not substantial and depend on the target series.

The rest of the paper is organized as follows. Section 2 reviews some results of the dynamic factor models and the method of principal components. Section 3 introduces the C-PC estimator and Section 4 establishes asymptotic convergence result and its relative efficiency to PC estimator. The small sample properties of the estimators are compared in Section 5 by means of Monte Carlo simulations. Finally, Section 5 concludes. Proofs are deffered to the Appendix.

Notation

The following notation is used throughout the paper: $E(.|Z_t)$ and $E_t(.)$ denote conditional expectation given variables in Z_t and given information at time t respectively, A' denotes the transpose of A, when $A = [a_{i,j}]$ is $q \times p$ matrix, $A' = [a_{j,i}]$ is of dimensions $p \times q$, $A \otimes B$ denotes the Kronecker product of matrices A and B, for $A = [a_{ij}]$ and $B = [b_{ij}]$, $A \otimes B = [a_{ij}B]$, A^{-1} denotes the inverse of a matrix A, ι_m is a m-vector of ones, I_m is an $m \times m$ identity matrix, $\text{diag}(A) = (a_{1,1}, a_{2,2}, ..., a_{n,n})$ when $A = [a_{i,j}]$ and, by "vector" we mean column vector, for any positive number a, $[a]$ is the largest integer smaller than or equal to a.

2 Preliminaries

Let X_{it} be the observed data for the i^{th} cross-section unit at time t ($i = 1, \cdots, N, t = 1, \cdots, T$). Consider the static factor model representation of the data:

$$X_{it} = \lambda_i' F_t + e_{it}, \tag{2.1}$$

where $F_t = \{F_{kt}\}_{1\leq k\leq r}$, is an $r \times 1$ vector of common factors, $\lambda_i = \{\lambda_{ik}\}_{1\leq k\leq r}$ is the corresponding vector of factor loading for cross-section unit i, and e_{it} is an idiosyncratic component. The only observable quantities are the X_{it}, both the common factors F_t and the loadings λ_i are not observed and are estimated. In fact, the number of factors r is also in principle unknown. For the purpose of this analysis, r is assumed to be known. There are several methods for determining the number of factors r. Some methods are based on the fit of the factors based forecasts, Stock and Watson [1998] . While other methods use information criteria, Bai and Ng [2002].

Let $\underline{X}_1, \cdots, \underline{X}_T$ be observations from the $N-$variate response variable, the factor structure in vector form:

$$\underline{X}_t = \Lambda F_t + \underline{e}_t, t = 1, \cdots, T, \tag{1.1}$$

where F_t is the $r \times 1$ vector of common factors, Λ is an $N \times r$ matrix of factor loadings, $\Lambda = \{\lambda_1', \cdots, \lambda_N'\}$; $\underline{e}_t$ is the $N \times 1$ vector of idiosyncratic component of the model. Let Ψ_N be the covariance matrix for the $N - variate$ response variable, $\Psi_N = \mathbf{E}(\underline{X}_t \underline{X}_t')$. Then the factor structure implies a variance decomposition in the form

$$\Psi_N = \Lambda_N \Omega_F \Lambda_N' + \Omega_N, \tag{1.2}$$

where the subscript N is explicit to show that the factor structure depends on the number of cross-sections. The existence and uniqueness of the approximate factor structure requires that the largest r eigenvalues of Ψ_N are unbounded with respect to N, the remaining eigenvalues are constant (Chamberlain and Rothschild [1983],Brown [1989] and Connor and Korajczyk [1993]). The approximate factor structure of Chamberlain and Rothschild [1983] generalizes the strict factor model which assumes diagonal error covariance Ω_N to allow for a more general covariance structure of the error term allowing for both time and cross-sectional dependence amongst the errors. The correlation between the idiosyncratic components is assumed to be weak both serially and cross-sectionally to allow for identification and estimation of the factor structure. The dimension of the panel in Chamberlain and Rothschild [1983] approximate factor model can be large in both N and T. In fact the high-dimensional property is needed to derive the desirable statistical properties of the estimate of both the factors and the loadings in an approximate factor model. The model assumes weak cross-sectional correlation, which is literally defined at the limit: as the number of variables grows larger, the correlation between these variables becomes smaller. At the limit, when N goes to infinity, the correlation dies out which ensures consistent estimation of the number of factors and the space spanned by the common factors (Stock and Watson [2002a] and Bai and Ng [2002]), and inferential theory (Bai [2003], Bai and Ng [2003]). The consistency result is achieved even if the estimation method doesn't exploit features of the data, such as heterogeneity in the signal to noise ratio, and non-spherical error component.

In matrix notation, the model is written as

$$\mathbf{X} = \mathbf{F^0 \Lambda^{0\prime}} + \mathbf{e}, \tag{2.2}$$

where $\mathbf{X} = [\underline{X}_1, \cdots, \underline{X}_T]'$ is the $T \times N$ matrix of observations, $\mathbf{e} = [\underline{e}_1, \cdots, \underline{e}_T]'$ is a $T \times N$ matrix of idiosyncratic errors, $\mathbf{F}^0 = [F_1^0, \cdots, F_T^0]'$ is the $T \times r$ matrix of common factors and $\mathbf{\Lambda}^0 = [\lambda_1^0, \cdots, \lambda_N^0]'$ is $N \times r$ matrix of factor loadings.

The underlying assumptions of the approximate factor structure are standard in the literature and are being reproduced here for completeness (see)Bai and Ng [2002] and Bai [2003]).

Assumption A1 (**Factors**). *$E\|F_t^0\|^4 < \infty$ and $T^{-1}\sum_{t=1}^{T} F_t^0 F_t^{0\prime} \to \Sigma_F$ as $T \to \infty$ for some positive definite matrix Σ_F.*

Assumption A2 (**Factor Loadings**). *$E\|\lambda_i^0\|^4 < \overline{\lambda} < \infty$, and $\|N^{-1}\sum_{i=1}^{T} \lambda_i^0 \lambda_{it}^{0\prime} - \Sigma_\Lambda\| \to 0$ as $N \to \infty$ for some $r \times r$ positive definite matrix Σ_Λ.*

Assumption A3 (**Error term**). *There exists a positive constant $M < \infty$, such that for all N and T,*

1. *$E(e_{it}) = 0, E|e_{it}|^8 \leq M$;*
2. *$E(\underline{e}_s'\underline{e}_t/N) = \gamma_N(s,t)$, $|\gamma_N(s,s)| \leq M$ for all s and $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}|\gamma_N(s,t)| \leq M$;*
3. *$E(e_{it}, e_{jt}) = \tau_{ij,t}$ with $|\tau_{ij,t}| \leq |\tau_{ij}|$ for some τ_{ij} and for all t; in addition,*

$$N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}|\tau_{ij}| \leq M;$$

4. *$E(e_{it}, e_{js}) = \tau_{ij,ts}$ and $(NT)^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}|\tau_{ij,ts}| \leq M$;*
5. *for every (t,s), $E\left|N^{-1/2}\sum_{i=1}^{N}[e_{is}e_{it} - E(e_{is}e_{it})]\right|^4 \leq M$*

Assumption A4. *Weak Dependence between Factors and Idiosyncratic Errors:*

$$E\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t^0 e_{it}\right\|\right) \leq M$$

This paper uses the usual normalization trick in the PC literature to enable identification of the factor structure of either $\frac{1}{T}\sum_{t=1}^{T}F_tF_t' = I_r$ or $\frac{1}{N}\sum_{i=1}^{N}\lambda_i\lambda_i' = I_r$. The model being analyzed in this paper is a static factor model in the sense that $\underline{X}_t$ has a contemporaneous relationship with the factors. Under the regularity conditions in Assumptions **A1-A4** (Bai and Ng [2002], Stock and Watson [2002b]), the factors and factor loadings can be consistently estimated as N and T are both large using the method of *asymptotic principal components*, Connor and Korajczyk [1989]. Technically, the principal component (PC) estimator minimizes the total sum of squares

$$V(\mathbf{\Lambda}, \mathbf{F}) \;=\; \mathrm{tr}\left[(\mathbf{X} - \mathbf{F}\mathbf{\Lambda}')'(\mathbf{X} - \mathbf{F}\mathbf{\Lambda}')\right]. \tag{2.3}$$

The PC estimator minimizes (2.3) subject to the normalization $\mathbf{F}'\mathbf{F}/T = I_r$. The estimator has a simple interpretation in terms of the singular value decomposition of

the sample covariance of the data. Consider the spectral decomposition of the sample covariance matrix of $\mathbf{X}$, $\Psi_N = \frac{1}{T}\mathbf{X}'\mathbf{X}$:

$$\Psi_N \Gamma = \Gamma \Delta,$$

where $\Delta = \text{diag}(d_1, \cdots, d_N)$ is a diagonal matrix with d_l corresponding to the l^{th} highest eigenvalue of Ψ_N, and $\Gamma = (\varphi_1, \cdots, \varphi_N)$ is the matrix whose columns corresponds to the normalized eigenvectors of Ψ_N. The 'normalized' PC estimator of $\mathbf{F}$ are $\widehat{F}_{k,t} = \frac{1}{\sqrt{d_k}}\varphi_k' \underline{X}_t$, for $k = 1, \cdots, r$; De Mol et al. [2008]. The PC estimator for $\mathbf{\Lambda}$ can be computed as OLS projection of $\mathbf{X}$ on the estimated $\hat{\mathbf{F}}$, $\mathbf{\Lambda} = \frac{1}{T}\mathbf{X}'\hat{\mathbf{F}} = \Gamma_{1:r}$. Let us assume that the processes are stationary, abstract from serial correlation and focus only on the role of the cross-sectional dependence assumption. Assumption (**A3.3**) is sufficient and necessary for the asymptotic properties of the PC estimators derived in the literature. Bai and Ng [2002] derive consistency results for the estimated number of factors, and the space spanned by the estimated factors under approximate factor model. Referring to mathematical appendix of Bai and Ng [2002], the result that PC estimated factors $\hat{F}_t$ span (up to an orthogonal rotation H^k) the space of the true factors F_t, ie, $C_{NT}^2 \left(\frac{1}{T} \sum_{t=1}^{T} \| \hat{F}_t^k - H^{k'} F_t \|^2 \right) = O_p(1)$ at a rate $C_{NT} = min\{N, T\}$, requires that $N^{-2} \| e_t' \mathbf{\Lambda} \|^2 = O_p(N^{-1})$. The latter follows from $E\left(T^{-1} \sum_{t=1}^{T} \| N^{-1/2} e_t' \mathbf{\Lambda} \|^2 \right) \leq \bar{\lambda} M$, where $\| \lambda_i \| \leq \bar{\lambda} < \infty$, which is a direct implication of Assumption (**A3.3**) as shown in Bai and Ng [2002]' Lemma 1. Further more, this average convergence rate is sufficient for consistency of the estimated number of factors using Bai and Ng [2002] criteria.

The assumption doesn't explicitly play a role in the PC estimation. In fact, the minimization of the system (1.1) implicitly assumes that the error covariance matrix is diagonal and homoscedastic. A number of studies have shown the importance of deviation from the assumption of spherical e_{it} on the small sample properties of the PC estimators. Boivin and Ng [2006] provide an empirical assessment of the extent of which likely features of the data affect the properties of the PC factors estimates $\hat{\mathbf{F}}$. Their study finds that forecast based on weighted had smaller errors that forecasts based on OLS-PC estimation. Their result points to a need to develop more efficient estimators that fully exploit information in the data. Let us assume for the moment that the errors are independent across time and that the time and cross-sectional dynamics are separable, $E(\underline{e}_t \underline{e}_t') = \Omega$. If Ω is known, a generalized least squares type principal component (GLS-PC) estimator can be constructed by minimizing,

$$V_\Omega(\mathbf{\Lambda}, \mathbf{F}) \;=\; \text{tr}\left[\Omega^{-1} (\mathbf{X} - \mathbf{F}\mathbf{\Lambda}')' (\mathbf{X} - \mathbf{F}\mathbf{\Lambda}') \right]. \tag{2.4}$$

This GLS-PC estimator is studied by Choi [2012] for the case $N < T$ with heteroscedastic errors, where $\Omega = \text{diag}[E(e_{1t}, \cdots, E(e_{Nt}))]$, and with block diagonal cross section dependence with n blocks, where $\Omega = \Omega_1 \bigoplus \Omega_2 \cdots \bigoplus \Omega_n$. Breitung and Tenhofen [2011] considers similar type estimation for dynamic factor models with heteroscedasticity and serial correlation. Estimation requires an estimate for the covariance matrix Ω. Feasible estimators include, the sample covariance matrix $\hat{\Omega}_N = T^{-1}(\mathbf{X} - \hat{\mathbf{F}}\hat{\mathbf{\Lambda}}')'(\mathbf{X} - \hat{\mathbf{F}}\hat{\mathbf{\Lambda}}')$. however, for high-dimensional systems with $N > T$, $\hat{\Omega}_N$ is singular. A candidate estimate for Ω is the sample covariance matrix. However, when $N > T$, $\widehat{\Omega}$ is singular and

minimizing $V_{\hat{\Omega}_N}(\mathbf{\Lambda}, \mathbf{F})$ is unfeasible. To overcome inverting a singular matrix, Boivin and Ng [2006] propose a weighting scheme that accounts for heteroscedasticity and cross correlation. The weighting scheme is then applied to the PC estimator minimize

$$V(\lambda_i, F_t, w_{it}) = \sum_{t=1}^{T} w_{it} \sum_{t=1}^{T} (X_{it} - \lambda_i' F_t)^2, \tag{2.5}$$

where choices of the weights include (i) w_{it} is the inverse of the diagonal element of $\widehat{\Omega}_T$ estimated using data up to time T and, (ii) w_{it} is the inverse of $N^{-1} \sum_{i=1}^{N} |\widehat{\Omega}_T(i,j)|$.

In principle, if an estimator of $\hat{\Omega}^{-1}$ is available, a GLS type PC estimation can be carried out. The random matrix literature has a rich body of work on estimating large dimensional covariance matrices. Some of the results in this literature have been used in the factor model literature. The approach is to estimate a sparse covariance matrix using threshholding or penalized maximum likelihood.

Bai and Liao [2016] apply the estimator principal orthogonal component threshholding estimator of Fan et al. [2013] to derive a two-step estimator, and use a penalized likelihood (Lam and Fan [2009b]) for their proposed joint estimation procedure. In their study, two estimators are proposed. The first is a two-step estimator that minimizes the negative log-likelihood function,

$$-\mathcal{L}_1(\mathbf{\Lambda}, \Omega) = \frac{1}{N} \log|\det(\mathbf{\Lambda}\mathbf{\Lambda}' + \Omega_N)| + \frac{1}{N} \text{tr}\left(S_{\mathbf{X}}(\mathbf{\Lambda}\mathbf{\Lambda}' + \Omega_N)^{-1}\right), \tag{2.6}$$

where $S_{\mathbf{X}}$ is the sample covariance matrix of the data. An estimator of Ω_N is obtained in a first step estimation using threshholding. The second joint estimator they propose is an l_1-penalized maximum likelihood estimator that minimizes,

$$L_2\mathbf{\Lambda}, \Omega) = -\mathcal{L}_1(\mathbf{\Lambda}, \Omega) + \frac{1}{N} \sum_{i \neq j} \mu_T w_{ij} |\Omega_{ij}| \tag{2.7}$$

The second estimator penalizes the off-diagonal elements of the covariance matrix.

3 The Cn-PC Estimator

Let us consider Assumption (**A3.3**) of bounded cross-sectional in an approximate factor model. The eigenvalues of the error covariance matrix $\Omega = E(\underline{e}_t \underline{e}_t')$ in Chamberlain and Rothschild [1983]'s factor model must be bounded. Under the assumption of (covariance) stationarity, $E(e_{it}e_{jt}) = \tau_{ij}$, all the eigenvalues of Ω are bounded by $\max_i \sum_{i=1}^{N} |\tau_{ij}|$. Thus Assumption (**A3.3**) is implied by the assumption of $\sum_{i=1}^{N} |\tau_{ij}| \leq M$ for all i and all N, Bai and Ng [2002]. The ordinary PC (OLS-PC, Breitung and Tenhofen [2011]) minimizes the total sum of squares

$$\underset{\lambda_i, F_t}{\text{minimize}} \; (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}^2. \tag{3.1}$$

The inequality in Assumption (**A3.3**) can be written as:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\text{sign}(\tau_{ij})\tau_{ij} \leq M, \tag{3.2}$$

where $\tau_{ij} = E(e_{it}e_{jt})$ and $e_{it} = X_{it} - \lambda_i' F_t$.

In this article, the estimated factors solve an optimization problem that combines the PC objective function and the assumption of bounded cross-sectional correlation. The factors and the loadings are estimated by solving the PC optimization in (3.1) under the restriction implied by (3.2). The proposed estimation solves:

$$\underset{\lambda_i, F_t}{\text{minimize}}\ (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2 \tag{3.3}$$

$$\text{s.t}\ \sum_{i=1}^{N}\sum_{j=1}^{N}\text{sign}(\tau_{ij})\tau_{ij} \leq M \tag{3.4}$$

The sign function is defined as $\text{sign}(a) = -1, 0, 1$ for $a < 0, = 0, > 0$. For a scalar a, $\text{sign}(a) = a/|a|$, that is, $\text{sign}(a) = a/\sqrt{a^2}$.

Let $\mathcal{S}$ be $N \times N$ matrix with elements $[\mathcal{S}_{ij}]$ defined as,

$$\mathcal{S}_{i,i} = 0 \tag{3.5}$$

$$\mathcal{S}_{i,j} = \text{sign}\left(E(e_{it}e_{jt})\right)\ \text{for}\ i \neq j. \tag{3.6}$$

The elements $\mathcal{S}_{ij}$ are functions of the model parameters, $\mathcal{S}_{ij} \equiv \mathcal{S}(\lambda_i, \lambda_j, F_t)$. An estimate of $\mathcal{S}_{ij}$ can be defined using the estimated parameters of the model,

$$\hat{\mathcal{S}}_{ij} = \hat{\mathcal{S}}\left(\hat{\lambda}_i, \hat{\lambda}_j, \hat{F}_t\right) \tag{3.7}$$

$$= \text{sign}\left(\frac{1}{T}\sum_{t=1}^{T}(X_{it} - \hat{\lambda}_i'\hat{F}_t)(X_{jt} - \hat{\lambda}_j'\hat{F}_t)\right) \tag{3.8}$$

Under regularity conditions, for $i \neq j$ plim $\hat{\mathcal{S}}(\hat{\lambda}_i, \hat{\lambda}_j, \hat{F}_t) = \text{sign}(\tau_{ij})$ as $T \longrightarrow \infty$. See Appendix for details.

In the following, let us assume that the population $\mathcal{S}_{ij}$ are known. This is of course unfeasible since in practice the variance covariance matrix is unknown. We argue that in many applications, institutional knowledge and theory may provide information about the direction of co-variation between variables without the knowledge of its strength. In this sense, inference about the sign of the correlation is more precise than inference about its value.

Let $\mathcal{L}_1(F, \Lambda) = \frac{1}{T}\sum_{t=1}^{T}\underline{e}_t'\underline{e}_t$ and $\mathcal{L}_2(F, \Lambda) = \frac{1}{NT}\sum_{t=1}^{T}\underline{e}_t'\mathcal{S}\underline{e}_t - M$. The optimization in (3.3)-(3.4) can be written as:

$$\underset{\Lambda, F}{\text{minimize}}\ \{\mathcal{L}_1(\Lambda, F, r)|\mathcal{L}_2(F, \Lambda, r) \leq 0\}, \tag{3.9}$$

under the normalization of either $T^{-1}\sum_{t=1}^{T} F_t F_t' = I_r$, or $N^{-1}\sum_{i=1}^{N}\lambda_i\lambda_i' = I_r$. This optimization problem can be solved using the theorem of Kuhn-Tucker.

Note that both $\mathcal{L}_1$ and $\mathcal{L}_2$ are of magnitude of order N. Let us assume that the number of factors r is known and concentrate on the estimation of F and Λ for a given r. Treating the system in (3.9) as a convex programming problem, the Lagrangian is

$$\mathcal{L}(\Lambda, F, \mu) = \frac{1}{N}\mathcal{L}_1(\Lambda, F) + \mu_{NT}\mathcal{L}_2(F, \Lambda). \tag{3.10}$$

The matrix $\mathcal{S}$ has diagonal elements equal to zero and off diagonal elements that are either 1 or -1. The lagrangian is similar to that of a shrinkage regression where the cross correlations are shrunk towards zero. The parameter μ_{NT} represents the cost/penalty for deviation of the solution from (3.2) and thus plays the role of a shrinkage factor.

Proposition 1. *The constrained principal component estimator (CPCE) for F^0, denoted $\hat{F}$, which solves (3.10) is $\sqrt{T}$ times the matrix consisting of the eigenvectors corresponding to the r largest eigenvalues of the matrix $\mathbf{X}\mathcal{A}_N\mathbf{X}'$, where $\mathcal{A}_N = I_N + \mu_{NT}\mathcal{S}$, where μ_{NT} is the Lagrange multiplier parameter. The Cn-PC estimator for Λ^0, denoted $\hat{\Lambda}$ is given by $\hat{\Lambda} = \frac{1}{T}\mathbf{X}\hat{F}$.*

See Appendix A.

Under the additional assumptions

Assumption A5 (**Error term**). *There exists a positive constant $M < \infty$, such that for all N and T,*

1. *let $E(\underline{e}_s'\mathcal{S}\underline{e}_t/N) = \varrho_N(s,t)$, then $\sum_{s=1}^{T}|\varrho_N(s,t)| \le M$ for all t.*
2. *for every t, s, and N, assume that $E\left|N^{-1/2}\left[\underline{e}_s'\mathcal{S}\underline{e}_t - E\left(\underline{e}_s'\mathcal{S}\underline{e}_t\right)\right]\right|^4 \le M$;*
3. *for any t and N, there exists a positive constant $M < \infty$ such that $E\left\|\frac{1}{\sqrt{N}}\Lambda^{0\prime}\mathcal{S}\underline{e}_t\right\|^2 \le M$.*

Theorem 1. *For any fixed (known) $r \ge 1$, there exists a suitable $(r \times r)$ full rank rotation matrix $\mathcal{H}$ such that under Assumption A1-A5*

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t - \mathcal{H}'F_t^0\right\|^2 = O_p(\delta_{NT}^{-2}) + O_p(\mu_{NT}^{-2}\delta_{NT}^{-2}),$$

where $\mathcal{H} = \left(\frac{\Lambda'\mathcal{A}_N\Lambda}{N}\right)\left(\frac{F'\hat{F}}{T}\right)V_{NT}^{-1}$. Or equivalently,

$$\omega_{NT}^2\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t - \mathcal{H}'F_t^0\right\|^2\right) = O_p(1),$$

where $\delta_{NT} = \min\{N, T\}$ and $\omega_{NT} = \min\{\delta_{NT}, \delta_{NT}\mu_{NT}\}$.

See Proof in Appendix B.
As in the standard principal components estimation, the true factors F_t^0 are identified only up to a scale. What is considered is the space spanned by the true factors identified by a rotation $\mathcal{H}F_t^0$ of F_t^0. In Theorem 1, the time average of squared deviations between the Cn-PC estimator and those that lie in the true factor space goes to zero as $N, T \to \infty$. The rate of convergence depends on the panel structure but also on the regularization factor μ_{NT}.

Note that the Cn-PC estimator are implicit functions of μ_{NT}. When $\mu_{NT} = O(1)$ (equivalent to $h = 0$ in Proposition below) the Cn-PC estimator of $\hat{F}$ is the principal component estimator of the factor space consisting of the eigenvectors corresponding to the r largest eigenvalues of $\mathbf{XX}'/T$ (see for example, Stock and Watson [2002a], Bai and Ng [2002], Bai [2003]). In this case, Theorem 1 implies the same rate of convergence as in Bai and Ng [2002] which is equal to δ_{NT} and is determined by the smaller of N or T.

Theorem 1 establishes conditions under which the convergence of the Cn-PC estimator is faster/slower than that of the ordinary PCEs.

Proposition 2. *Let $\mu_{NT} = \delta_{NT}^{-h}$, then the rate of convergence in Theorem 1 is:*

(i) $\omega_{NT}^2 = \delta_{NT}^{2(1-h)}$ *for* $h > 0$,

(ii) $\omega_{NT}^2 = \delta_{NT}^2$ *for* $h \leq 0$.

In the case of $h > 0$, $\omega_{NT}^2 < \delta_{NT}^2$ and thus the Cn-PC estimator converge (in the sense of Theorem 1) to factors that lie in the true factors space at a rate slower than Bai and Ng [2002] ordinary CPEs. The two methods are implying a different rotation matrix $\mathcal{H}$ which means the convergence is towards different rotation of the space spanned by the true factors. Thus the estimated factor spaces are not directly comparable.

Lemma 1. *Assume in addition that, then $\sum_{s=1}^{T}\sum_{s=1}^{T}\gamma_N(s,t)^2 \leq M$ for some $M < \infty$ uniformly in t,*

$$\omega_{NT}^2 \left\| \hat{F}_t - \mathcal{H}' F_t^0 \right\|^2 = O_p(1).$$

The proof is similar to that of Theorem 1.

3.1 Choosing M

The 'shrinkage' parameter μ_{NT} can be estimated from the objective function $\mathcal{L}(\hat{\mu})$. The system is a function of M, the amount of cross-sectional correlation allowed in the approximate factor structure. M is a tuning parameter that controls the amount of shrinkage that is applied to the estimated τ_{ij}. Clearly, μ_{NT} increases as M decreases. The relationship between μ_{NT} and M is a correspondence and not a function. Although positive values of μ_{NT} correspond to a single value of M, the value $\mu_{NT} = 0$ relates to all M in $\left[\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}|E(\hat{e}(0)_{it}\hat{e}(0)_{jt})|, \infty\right)$, $\hat{e}(0)_{it}$ are the residuals from the unconstrained PCA. If the factor structure is strict, then there is no need for shrinkage.

Let $M_0 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|T^{-1}\sum_{t=1}^{T}\hat{e}(0)_{it}\hat{e}(0)_{jt}\right|$, then values of $M < M_0$ will increase shrinkage and induce more sparsity of the error covariance matrix. The complimentary slackness conditions are used to deduce an estimate $\hat{\mu}_{NT}$ of μ_{NT}. If the constraints are not binding and $\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}|E(e_{it}e_{jt})| \leq M$, then the constrained maxima are the PC solution $\left(\hat{F}, \hat{\Lambda}, 0\right)$. On the other hand, if $\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}|E(e_{it}e_{jt})| > M$, then by the complimentary slackness we must have $\hat{\mu} > 0$ and $\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}|E(e_{it}e_{jt})| = M$

Chamberlain and Rothschild (1983) showed that asset prices have an approximate factor structure if the largest eigenvalue of $\Omega = E(e_t e_t')$ is bounded. The largest eigenvalue of Ω is bounded by $\max_i \sum_{i=1}^{N}\|\tau_{ij}\|$, where $\tau_{ij} = E(e_{it}e_{jt})$, (Boivin and Ng [2006]).Under the assumptions of an approximate factor model, there should exist a M such that $\sum_{j=1}^{N}\|\tau_{ij}\| \leq M < \infty$ for all i and N. This assumption is vital in the development of the approximate factor structure theory. However, there is no indication as to how much cross-correlation is permitted in practice. Boivin and Ng [2006] use $\hat{\tau}^* = \max_i \hat{\tau}_i^*/N$, where $\hat{\tau}_i^* = \sum_{j=1}^{N}|T^{-1}\sum_{t=1}^{T}\hat{e}_{it}\hat{e}_{jt}|$ as indicator for M/N, which should be small and decreasing with N. That is, the bounding quantity M is of order $O_p(N)$.

There a correspondence between the tuning parameter μ_{NT} that controls the amount of regularization and the threshold M. If M is greater or equal than the $L_{1,1}-$norm of the PC regression sample covariance matrix, $M_0 = \sum_{j=1}^{N}\sum_{i=1}^{N}|\hat{\tau}_{ij}|$, $\hat{\tau}_{ij} = \sum_{t=1}^{T}\hat{e}_{it}\hat{e}_{jt}/T$, then the PCA estimator is, of course unchanged by the proposed regularization. For smaller values of M, the constrained problem shrinks the estimated cross-sectional correlations towards the origin in the $L_{1,1}$ sense. One-way to calibrate and estimate M is cross-validation. Using a normalized parameter $m = M/M_0$ to index the constrained estimates of F and Λ over a grid of values of s between 0 and 1 inclusive. The value $\hat{m}$ yielding the lowest estimated value for some risk function is selected. The risk can be measured in terms of fit of the factors estimates $\hat{F}$ and/or in terms of prediction error for factor based $h-$steps ahead forecasts. In our analysis, we present the path of solutions indexed by a fraction m of shrinkage factor of M_0.

Penalized principal components regression

A closely related optimization problem to constrained PC regression problem in (3.10) is the constrained regression

$$\underset{\lambda_i, F_t}{\text{minimize}}\ (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}^2 + \kappa_{nt}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}|E(e_{it}e_{jt})| \tag{3.11}$$

Problems (3.10) and (3.11) are equivalent (Osborne et al. [2000]). For a given $\kappa_{nt}, 0 \leq \kappa_{nt} < \infty$, there exists a $M \geq 0$ such that the two problems share the same solution, and vice versa. Generalizing the result in

In (3.11), the parameter κ_{nt} is easily interpreted as shrinkage/regularization parameter applied to large cross-section correlation parameters. The Lagrange multiplier μ_{NT} is the price of deviation from the bounded cross correlation constraint imposed by the

approximate factor structure. The two parameters are exchangeable for all practical purposes.

4 Limiting distributions of constrained principal component estimators

In this section, We study the asymptotic distributions of the proposed constrained PC estimators. In particular, these estimators are compared to the properties of the ordinary PC estimators of Bai [2003] and the generalized PCEs of Choi [2012].

Assumption A6. *Moments and Central Limit Theorem*

1. *for any t, N and T, there exists an $M < \infty$ such that*

$$E \left\| \frac{1}{\sqrt{NT}} \sum_{s=1}^{T} F_s^0 \left[\underline{e}_s' \mathcal{S} \underline{e}_t - E\left(\underline{e}_s' \mathcal{S} \underline{e}_t \right) \right] \right\|^2 \leq M;$$

2. *for any N and T, there exists an $M < \infty$ such that*

$$E \left\| \frac{1}{\sqrt{NT}} \sum_{s=1}^{T} \Lambda^{0\prime} \mathcal{S} \underline{e}_s F_s^{0\prime} \right\|^2 \leq M;$$

3. *for each t, as $N \to \infty$,*

$$\frac{1}{\sqrt{N}} \Lambda^{0\prime} \underline{e}_t \xrightarrow{d} N(0, \Psi_t)$$

where $\Psi_t = \lim_{N\to\infty} \frac{1}{N} \Lambda^{0\prime} E(\underline{e}_t \underline{e}_t') \Lambda^0$;

4. *for each i, as $T \to \infty$,*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} F_t^0 e_{it} \xrightarrow{d} N(0, \Phi_i),$$

where $\Phi_i = plim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} E(F_t^0 F_t^{0\prime} e_{it} e_{is})$.

Assumption A7 (Factor Loadings*). $\| N^{-1} \Lambda^{0\prime} \mathcal{A}_N \Lambda^0 - \Sigma_{\Lambda *} \| \to 0$ *as* $N \to \infty$ *for some* $r \times r$ *positive definite matrix* $\Sigma_{\Lambda *}$.

Assumption A8. *The eigenvalues of the* $r \times r$ *matrix* $(\Sigma_{\Lambda *} \cdot \Sigma_F)$ *are distinct.*

Assumption A9. *The tuning parameter* μ_{NT} *satisfy:*

1. $\frac{1}{\delta_{NT}} = o(\mu_{NT})$, $\frac{\sqrt{N}}{\sqrt{T}\delta_{NT}} = o(\mu_{NT})$ *and* $\mu_{NT} = o(1)$;
2. $\mu_{NT} \sum_{i \neq j} |\tau_{ij}| \to 0$.

Theorem 1. *Suppose that Assumptions A1-A7 hold.*

1. *If* $\frac{\sqrt{N}}{T\mu_{NT}} \to 0$,

$$\sqrt{N}\left(\hat{F}_t - \mathcal{H}' F_t^0\right) \xrightarrow{d} N(0, V^{-1/2}\mathcal{Q}\Psi_t\mathcal{Q}'V^{-1/2}). \tag{4.1}$$

Proof in Appendix B.2

Efficiency of Cn-PC estimator

The main motivation of this paper is to improve on the existing estimators in terms of efficiency. The ordinary PCEs have asymptotic distribution (Theorem 1 of Bai and Ng [2003]):

$$\sqrt{N}\left(\hat{F}_t - H' F_t^0\right) \xrightarrow{d} N(0, V_{opc}^{-1/2} Q_{opc}\Psi_t Q'_{opc} V_{opc}^{-1/2}), \tag{4.2}$$

where $Q_{opc} = \Sigma_\Lambda^{-1/2}\Upsilon_{opc}V_{opc}^{1/2}$, Υ_{opc} is eigenvector of $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}$, and $V_{opc} = Q_{opc}\Sigma_\Lambda Q'_{opc}$. Let X_{it} be the observed data for the i^{th} cross-section unit at time t ($i = 1, \cdots, N, t = 1, \cdots, T$). Consider the static factor model representation of the data:

$$X_{it} = \lambda_i' F_t + e_{it}, \tag{4.3}$$

where $F_t = \{F_{kt}\}_{1\le k\le r}$, is an $r \times 1$ vector of common factors, $\lambda_i = \{\lambda_{ik}\}_{1\le k\le r}$ is the corresponding vector of factor loading for cross-section unit i, and e_{it} is an idiosyncratic component. The only observable quantities are the X_{it}, both the common factors F_t and the loadings λ_i are not observed and are estimated. In fact, the number of factors r is also in principle unknown. For the purpose of this analysis, r is assumed to be known. There are several methods for determining the number of factors r. Stock and Watson [1998] develop a consistent estimator for r based on the fit of factor based forecasts. Bai and Ng [2002] use information criteria to penalize the sum of squared residuals in model (4.4) to construct consistent estimator for r.Let X_{it} be the observed data for the i^{th} cross-section unit at time t ($i = 1, \cdots, N, t = 1, \cdots, T$). Consider the static factor model representation of the data:

$$X_{it} = \lambda_i' F_t + e_{it}, \tag{4.4}$$

where $F_t = \{F_{kt}\}_{1\le k\le r}$, is an $r \times 1$ vector of common factors, $\lambda_i = \{\lambda_{ik}\}_{1\le k\le r}$ is the corresponding vector of factor loading for cross-section unit i, and e_{it} is an idiosyncratic component. The only observable quantities are the X_{it}, both the common factors F_t and the loadings λ_i are not observed and are estimated. In fact, the number of factors r is also in principle unknown. For the purpose of this analysis, r is assumed to be known. There are several methods for determining the number of factors r. See for example, Stock and Watson [1998] and Bai and Ng [2002]. It is very hard to compare the asymptotic variance covariance matrices in (4.1) and (4.2) because the Cn-PC

estimator and PCEs are estimating different objects. These estimators are estimating different rotations of the true factors because H in the PCEs and $\mathcal{H}$ in the Cn-PC estimator are generally different.

Consider the case of a factor structure with one common factor. This is an interesting case where H and $\mathcal{H}$ are identical and equal to the scalar $\Sigma_F^{-1/2}$. In this case, PCEs and Cn-PC estimator are estimating the same object $F_t/\sqrt{\Sigma_F}$. Since in this case (of $r = 1$), $\Upsilon = \Upsilon_{opc} = \equiv 1$, $\mathcal{Q} = Q_{opc} = \Sigma_F^{-1/2}$, then the Cn-PC estimator of F_t^0

$$\hat{F}_t \backsimeq \frac{F_t^0}{\sqrt{\Sigma_F}} + \frac{1}{\sqrt{N}} N\left(0, \frac{1}{\Sigma_F}\Sigma_{\Lambda*}^{-1}\Psi_t\Sigma_{\Lambda*}^{-1}\right) \tag{4.5}$$

and the PCEs have

$$\hat{F}_{t,opc} \backsimeq \frac{F_t^0}{\sqrt{\Sigma_F}} + \frac{1}{\sqrt{N}} N\left(0, \frac{1}{\Sigma_F}\Sigma_{\Lambda}^{-1}\Psi_t\Sigma_{\Lambda}^{-1}\right), \tag{4.6}$$

where

$$\Sigma_{\Lambda*} = \Sigma_\Lambda + \mu_{NT} \text{ plim } \frac{\Lambda'\mathcal{S}\Lambda}{N} \geq \Sigma_\Lambda, \tag{4.7}$$

because $\mathcal{S}$ is positive definite and $\mu_{NT} > 0$. In this case, the Cn-PC estimator are more efficient than the PCEs with ratio of (asymptotic) variances equal to:

$$\frac{V(\hat{F}_{t,opc})}{V(\hat{F}_t)} = \left(1 + \mu_{NT} \text{ plim } \frac{\Lambda'\mathcal{S}\Lambda}{N}\right)^2.$$

Note that the variance of the Cn-PC estimator decreases to zero as $\mu_{NT} \to \infty$:

$$\lim_{\mu_{NT}\to\infty} V(\hat{F}_t) = \lim_{\mu_{NT}\to\infty} \frac{\Psi_t}{\Sigma_F\left(\Sigma_\Lambda + \mu_{NT}\Sigma_{\mathcal{S}\Lambda}\right)^2}$$

5 Monte Carlo Simulations

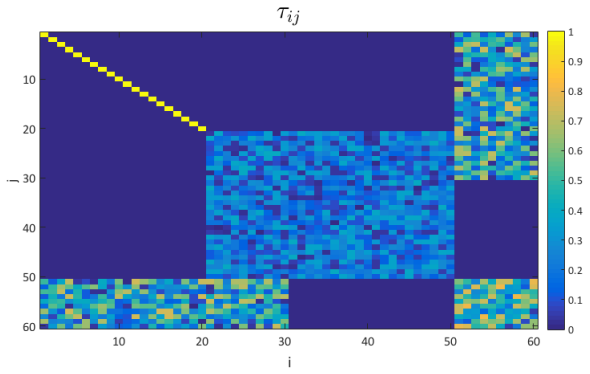
5.1 Simulations designs

This section presents the Monte Carlo experiments designed to study the small sample properties of the proposed Cn-PC estimator and their performance relative to the ordinary PCEs in the presence of cross-correlated errors. The experimental design for the Monte Carlo simulation adopts the same covariance structure as in Boivin and Ng [2006]. Let the total number of cross-sections N be divided into three groups of sizes N_1, N_2 and N_3 such as, $N = N_1 + N_2 + N_3$. Let the errors u_{it} be the building blocks for the errors dynamics with $u_{it} \sim N(0, 1), i = 1, \cdots, N$, and construct the errors e_{it} where

$$N_1 : e_{it} = \sigma_1 u_{it},$$

$$N_2 : e_{it} = \sigma_2 u_{it},$$

Figure 3: Sparsity of the errors covariance matrix, $N = 50$



$$N_3 : e_{it} = \sigma_3 \tilde{e}_{it}, \tilde{e}_{it} = u_{it} + \sum_{j=1}^{C} \rho_{ij} u_{jt}$$

In this experiment, the errors in the first N_1 series are mutually uncorrelated, the errors in the next N_2 are also mutually uncorrelated but their variance differ from the first series, $\sigma_2^2 > \sigma_1^2$. Cross correlation is introduced in the last N_3 series. The latter are correlated with a proportion C from the N_1 group. The cross correlation matrix Ω_{13} therefore has $C \cdot N_3$ non-zero elements. The correlation coefficients ρ_{ij} denote cross-correlation of series $i \in \{1, N_1\}$ and $j \in \{N_1 + N_2 + 1, N\}$ and is drawn from a uniform distribution $U[0.05, 0.7]$. The error variance in the third group is $\sigma_3^2 = \sigma_1^2$.The error covariance matrix takes the form:

$$\begin{aligned}
\Omega_{ii} &= \sigma_1^2, & 1 \leq i \leq N_1 \\
\Omega_{ii} &= \sigma_2^2, & N_1 + 1 \leq i \leq N_1 + N_2 \\
\Omega_{ii} &= \sigma_3^2, & N_1 + N_2 + 1 \leq i \leq N \\
\Omega_{ij} &= 0, & 1 \leq i, j \leq N_1 + N_2 \\
\Omega_{ij} &= \sigma_1 \sigma_3 \rho_{ij}, & i \leq C, N_1 + N_2 + 1 \leq j \leq N.
\end{aligned}$$

Figure 3 displays an example of the pattern of dependence structure in the simulation design. The variances are equal to one, and thus the graph represents a sparcity plot for both the covariance and the correlation matrix. In this example, there is clustering of correlations between group 1 and group 3 as well as within group 1 and group 3 series.

The common factors and the loadings are fixed throughout the simulation, which corresponds to analysis conditional on F^0 and Λ^0. The number of factors r is known and fixed. The study considers two cases, r equal to one and two. The panel dimension takes combinations of $T = 50, 100$, and $N = 50, 100, 150$. Data are generated through $X_{it} = \sum_{m=1}^{r} \lambda_{im} F_{mt} + e_{it}$. Our Monte Carlo results are based on $L = 2,000$ repetitions. For each repetition $l = 1, \cdots, L$, the Monte Carlo experiment is carried out as follows.

(i) Compute the ordinary principal components estimators of $\hat{F}_{OPCE}^{(l)}$, $\hat{\Lambda}_{OPCE}^{(l)'}$ and the estimated errors $\hat{e}_{OPCE}^{l} = \mathbf{X}^{(l)} - \hat{\Lambda}_{OPCE}^{(l)'} \hat{F}_{OPCE}^{(l)}$. Using the sample covariance matrix $\hat{\Omega}_{OPCE} = \hat{e}'\hat{e}/NT$, construct an estimate for sign matrix, $\hat{\mathcal{S}}^{(l)}$.

(ii) Given a value of $M = m \cdot M_0$, where $m \in [0, 1]$, compute $\left(\hat{F}^{(l)}, \hat{\mu}^l_{NT}\right)$:

(a) Begin with a starting value $\mu_{NT} = \mu_0$, here we take $\mu_0 = 0.5\sqrt{tr(\hat{e}'\hat{e})/tr(\hat{e}'\mathcal{A}_N\hat{e})}$, and $\mathcal{A}_\mu = I_N - \mu\mathcal{S}$, find the optimal solution to the dual objective function $\mathcal{L}(\mu)$:

$$\hat{\mu}_{NT} = \arg\max_\mu (NT)^{-1}\left[\text{tr } \mathbf{X}\mathcal{A}_\mu\mathbf{X}' - \text{tr } \hat{F}'_\mu\mathbf{X}\mathcal{A}_\mu\mathbf{X}'\hat{F}_\mu\right] - M, \qquad (5.1)$$

where $\hat{F}_\mu$ is $\sqrt{T}$ times eigenvectors corresponding to the largest r eigenvalues of $\Psi_{N,\mu} = \frac{1}{T}\mathbf{X}'\mathcal{A}_\mu\mathbf{X}$. This is iterated to convergence and to optimal values $\hat{F}^{(l)}, \hat{\mu}^{(l)}_{NT}$.

(b) Compute the Cn-PC estimator for the loadings as a linear projection of $\mathbf{X}$ on $\hat{F}^{(l)}$: $\hat{\Lambda}^{(l)} = \frac{1}{T}\mathbf{X}'\hat{F}^{(l)}$.

(iii) Compute the following measures of performance.

- *Percentage explained variation.* Boivin and Ng [2006] use the percentage of variation in the true factors captured by the estimated structure,

$$S^{(l)}_{\hat{F},F^0} = \frac{\text{tr}\left(F^{0\prime}\hat{F}^{(l)}\left(\hat{F}^{(l)\prime}\hat{F}^{(l)}\right)^{-1}\hat{F}^{(l)\prime}F^0\right)}{\text{tr}(F^{0\prime}F^0)}. \qquad (5.2)$$

- *Small sample bias.* The estimated factors and the true factors are not directly comparable. The estimated factors span a transformation of the true factors. In comparing the small sample bias of the Cn-PC estimator and the benchmark PCEs, one has to account for the differences in the rotation matrices H and $\mathcal{H}$. We compute the small sample bias of the (rotated) factors $\tilde{F}_t \equiv \mathcal{H}^{-1}\hat{F}_t$:

$$\text{bias}^{(l)} \;=\; \frac{1}{L}\sum_{l=1}^{L}\tilde{F}^{(l)}_{tk} - F^0_{tk}, \qquad (5.3)$$

for $k = 1$ and $t = 1, [T/2], T$.

- *Empirical mean squared errors (MSEs).* For each $\hat{F}^{(l)}_t$, we compute

$$MSEs^{(l)} = r^{-1}\left\|\hat{F}^{(l)}_t - F^{0(l)}_t\right\|^2. \qquad (5.4)$$

5.2 The 'Diffusion Index' framework

Consider the forecasting model whereby we are interested in the one h-ahead forecast of a series y_t. The series to be forecasted in both Monte Carlos are generated by

$$y_{t+h} = \beta_0 + \sum_{j=1}^{r}\beta_j F^0_{jt} + \epsilon_{t+h} \equiv y_{F^0,t+h|t} + \epsilon_{t+h},$$

where $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, and σ_ϵ^2 is chosen such that the R^2 of the forecasting equation is κ_y. The infeasible diffusion index forecast is $\hat{y}_{F^0,t+h|t}$, which only requires estimation of β. The feasible diffusion index forecast is denoted $\hat{y}_{\hat{F},t+h|t}$, which requires estimation of both the factors and β. A forecast using the observed N series is not feasible if N is large. However, one can use the factor structure of X_{it} in equation (4.4) and use $F_t^0 \equiv \{F_{jt}^0\}_{j=1}^r$ to account for the important drivers of common variation in $\mathbf{X}$:

$$\hat{y}_{F^0,t+h|\mathcal{I}_t} = \hat{\beta}_0 + F_t^{0'}\hat{\beta}. \tag{5.5}$$

This forecast is unfeasible since the true factors F_t^0 are unobserved. Given estimates $\hat{F}_{t,N} \equiv \{\hat{F}_{jt,N}\}_{j=1}^{\hat{r}}$, using the data from the N series and conditional on information at time $\mathcal{I}t$, a feasible factor augmented forecast, also known as a 'diffusion index' forecast (Stock and Watson [2002a]), is

$$\hat{y}_{\hat{F}_{t,N},t+1|\mathcal{I}_t} = \hat{\beta}_0 + \hat{F}_{t,N}'\hat{\beta}. \tag{5.6}$$

The feasible 'diffusion index' forecast requires the estimation of both F_t and β and thus depends on the properties of the 'generated' regressors $\hat{F}_{t,N}$. We compute the empirical mean-squared-forecast errors (MSFE) and, Boivin and Ng [2006]

$$MSFE_{\hat{y}^{\hat{F}},\hat{y}^{F^0}} = \frac{1}{J}\sum_{t=T}^{T+J-1}\left(\hat{y}_{F^0,t+1|t} - \hat{y}_{\hat{F},t+1|t}\right)^2 \tag{5.7}$$

$$S_{\hat{\beta},\beta} = \frac{1}{J}\sum_{t=T}^{T+J-1}\left(y_{\hat{F},t+1|t} - \hat{y}_{\hat{F},t+1|t}\right)^2 \tag{5.8}$$

The statistic(5.7) measures the loss in forecast accuracy due to F_t being unobserved and estimated. If the estimated factors are consistent and span the same space as the true factors, the difference in forecasting performance of the two predictors $\hat{F}_{t,N}$ and F_t^0 should be negligible and $S_{\hat{y}^{\hat{F}},\hat{y}^{F^0}}$ close to one. The larger is $S_{\hat{y}^{\hat{F}},\hat{y}^{F^0}}$, the closer are the 'diffusion index' forecasts to those generated by the (infeasible) forecasts based on observed factors. The statistic in (5.8) assesses the accuracy of the 'diffusion index' forecasts relative to the conditional mean forecasts which requires only estimation of F_t. Smaller values of $S_{\hat{\beta},\beta}$ are desirable.

A pseudo-out-of-sample forecasting experiment with 10 years rolling window corresponds to $T = 120$ and is carried out for 10 years into the future, $J = 120$. The panel size in this experiment are $T = 120$ and $N = 131$ to reflect the panel dimensions commonly used in macroeconomic forecasting.

5.3 Simulation results

Fixed M

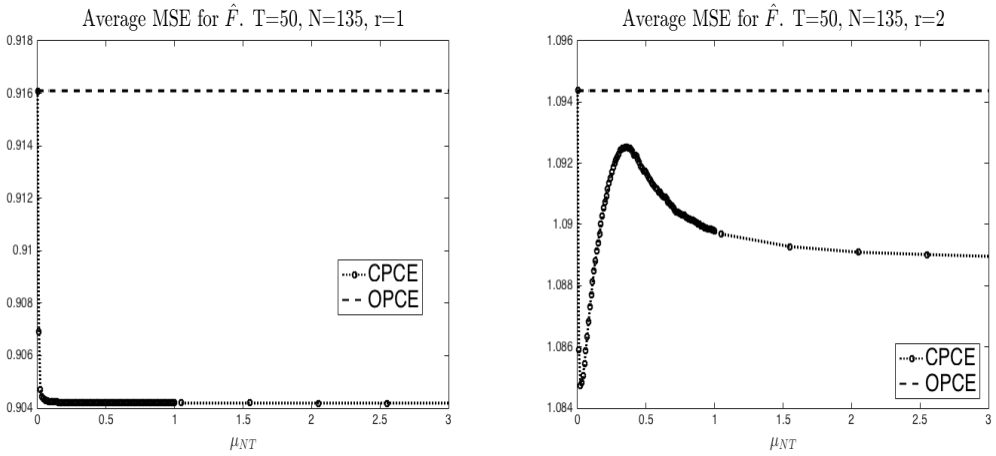
Table 1 reports the small sample bias and sample standard deviation of the estimated factors $\tilde{F}_{tj}$ for $j = 1$. For the sake of brevity, the results are computed for t equals to $[T/2]$ and T. The number of factors in this design experiment is $r = 1$. Note

Table 1: Small sample bias and standard errors for the estimated factors $\tilde{F}_{t,1}$, for $t = [T/2], T$ and $r = 1$

		Cn-PC				PCE			
		$\tilde{F}_{[T/2],1}$		$\tilde{F}_{T,1}$		$\tilde{F}_{[T/2],1}$		$\tilde{F}_{T,1}$	
T	N	bias	std	bias	std	bias	std	bias	std
50	50	-0.018	0.190	-0.092	0.163	-0.024	0.037	-0.139	0.097
	100	0.001	0.179	-0.207	0.092	0.033	0.031	-0.025	0.008
	150	-0.155	0.136	0.293	0.137	-0.211	0.103	0.353	0.166
100	50	-0.004	0.118	0.008	0.092	-0.046	0.025	0.121	0.015
	100	-0.128	0.102	-0.109	0.106	-0.120	0.079	-0.114	0.054
	150	-0.168	0.105	-0.049	0.115	-0.126	0.063	-0.013	0.053
150	50	-0.007	0.089	0.026	0.078	-0.032	0.053	0.070	0.081
	100	0.018	0.097	-0.113	0.086	0.054	0.022	-0.180	0.032
	150	0.093	0.062	0.031	0.065	-0.000	0.020	-0.065	0.021

The results are for the sampling distribution of $\tilde{F}_t = \mathcal{J}^{-1}\hat{F}_t$, $\mathcal{J} = \mathcal{H}$ for Cn-PC and $\mathcal{J} = H$ for PCE. The shrinkage factor M is chosen by a $10-$fold cross-validation.

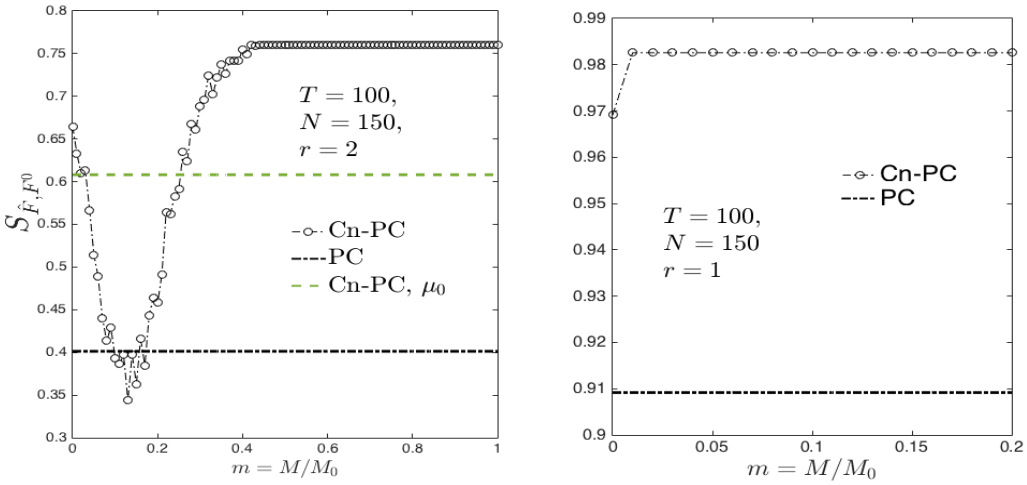
Figure 4: Accuracy of the Cn-PC estimators of common factors: Empirical MSEs



Note: The MSEs are computed for the rotated estimated factor matrix $\tilde{F}_t = \mathcal{J}\hat{F}_t$. The shrinkage factor M is equal to M_0 for which the constrained problem has the same solution as its dual penalized PC regression.

that in all of Monte Carlo results, the number of factors r is assumed to be known. The threshold $M = \overline{M}$ for which (3.11) and (3.9) have the same solution. Results show that overall, the proposed constrained estimators (Cn-PC estimator) have smaller bias compared to the ordinary principal components estimators (PCEs). The sample standard deviation of the Cn-PC estimator is larger than those of the PCEs for panel dimensions considered in the experiment. Figure 4 displays the sample MSEs (5.4) for the rotated factors matrix $\widetilde{F}_t$ estimated using the Cn-PC over a grid of values for the regularization parameter μ_{NT} and for a given M. The results shown are equivalent to the penalized PC estimator that solves (3.11). Results for the PCEs are reported in dashed line. The left panel is for the case with one true factor, and the right panel is for the case of two factors in the population model. As expected, the proposed technique with $\mu_{NT} = 0$ gives the same factors' accuracy in terms of MSEs as the standard principal components method. As the penalization increases, the MSEs for

Figure 5: Accuracy of Cn-PC estimators of common factors $\hat{F}$: $S_{\hat{F},F^0}$



model with $r = 1$, decrease sharply. For $r = 2$ DGP, the MSEs of $\tilde{F}_t$ also reaches a stable value after some dynamics for small μ_{NT}. The relationship between MSE and μ_{NT} is not monotonic.

5.3.1 M indexed path

Figure 5 displays the path of the statistic $S_{\hat{F},F^0}$ indexed by $m = M/M_0$. Note that in this experiment, the Cn-PC estimator $\hat{F}_t$ and $\hat{\mu}_{NT}$ are jointly estimated. The Cn-PC estimator results are shown in circle-dot dashed line, and the PCEs in bold dashed line. The left panel plots the results in case of a covariance matrix Ω^0 of design 1 and the right panel for design 2 covariance matrix. The right panel shows that the PCEs is doing a pretty good job with statistic values in the 90% range. However, there is also a clear advantage of the Cn-PC estimator with values that range from 0.97 to 0.98. For the more elaborate covariance structure in the left panel, the estimated factors span less perfectly the true factors. The PCEs are low in the 0.40 range. The Cn-PC estimator improves the ability of the factors estimates to span the factor space with values reaching 0.75. The plot suggests that the relation between M and $S_{\hat{F},F}$ is not monotonic. There are some values of M for which the Cn-PC estimator do slightly worse than the standard techniques.

These plots display how the estimated statistics are affected by the model threshold selection method. These can be graphical tools for selection of the parameter M in similar way as the 'ridge-trace' plot is used in the context of ridge regression, (Hoerl et al. [1975]). Such plots provide a visual assessment of the effect on coefficient of the choice of the ridge regularization parameter, thus allowing the analyst to make more informed decision. The selected M^* corresponds to the threshold value at which the value of the statistic of interest stabilizes.

In Table 2, the estimator GLS-PC refers to Choi [2012] estimator which uses PCE sample covariance estimator to compute a feasible generalized PC efficient estimator.

Figure 6: Empirical mean squared errors (MSEs) of Cn-PC estimators of common factors $\hat{F}$ and common components $\hat{C}$

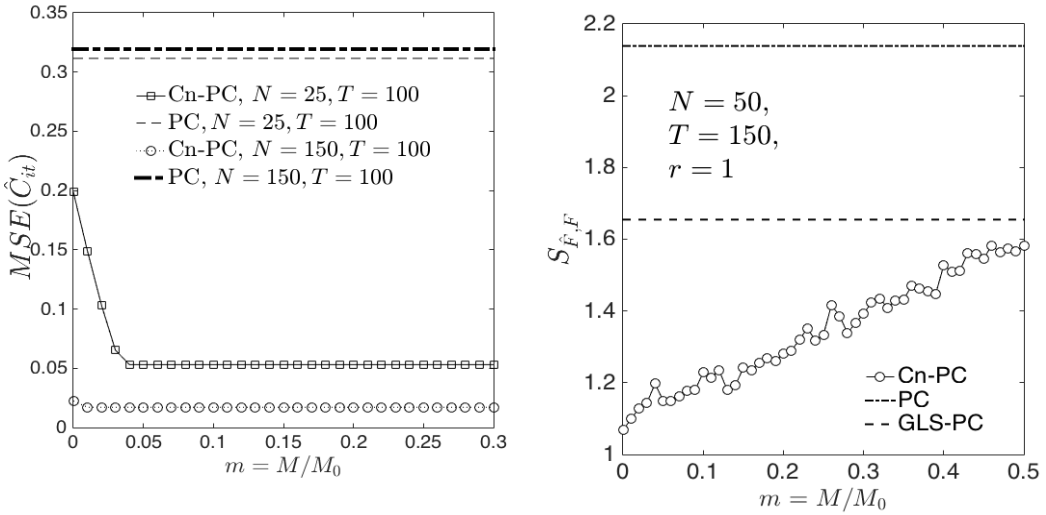


Table 2: Efficiency of estimated common factors: $S_{\hat{F},F^0}$ and $MSE_{\hat{F}}$

		$S_{\hat{F},F^0}$			$MSE_{\hat{F}}$		
T	N	PC	Cn-PC	PC-GLS	PC	Cn-PC	PC-GLS
100	25	0.13	0.319	0.434	2.17	2.16	2.13
	50	0.12	0.382	0.157	1.84	1.76	1.77
150	50	0.10	0.337	0.185	1.78	1.95	1.97
	100	0.10	0.580	0.078	1.83	1.89	1.94
55	50	0.24	0.505	0.072	1.94	1.95	1.75
50	25	0.26	0.341	0.252	1.96	1.36	2.02

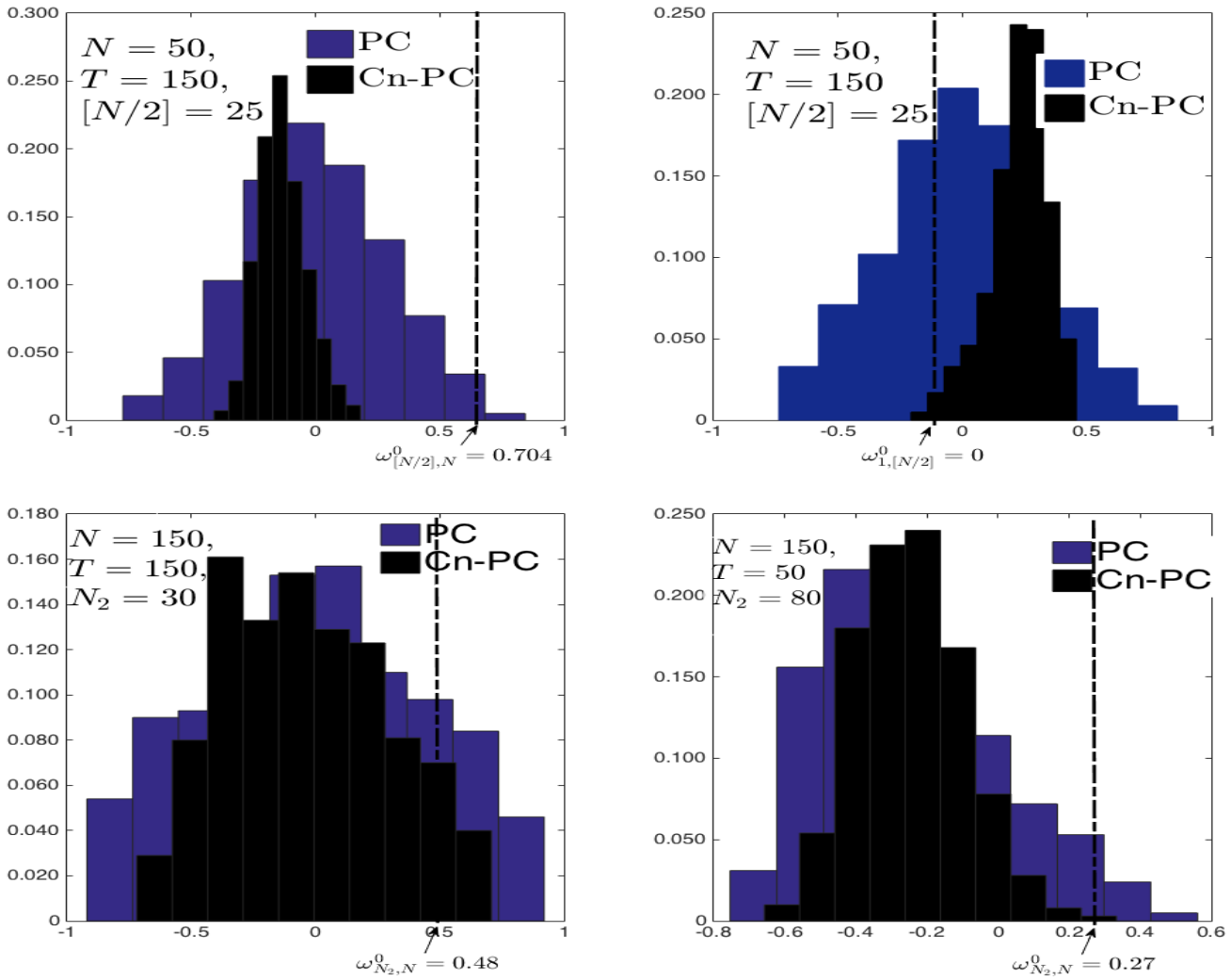
The OLS-PC are very inaccurate in terms of $S_{\hat{F},F^0}$. The GLS-PC performs better in case of T large and N small. However, as N becomes larger, PC-GLS becomes less accuarate. When N is large, GLS-PC performs poorly with S_{F,F^0} considerably lower than the ones for the PC and Cn-PC estimators. The low accuracy of PC-GLS can be explained by the poor accuarcy and unstable estimator of the covariance matrix when N is large and close to T.

Sample correlations

Similar to the use 'ridge-trace' plot which shows graphically the effect of the shrinkage parameter on the coefficients in the linear regression model, one can look at the effect of the threshold M on the elements $|\hat{\tau}_{ij}|$ and the sample cross-section correlations. We use this strategy to select the threshold M for the results in this section.

Figure 7 shows histograms of the sampling distribution of $\hat{\omega}_{ij}$ for a selection of values for i and j. We select cases where $\omega_{ij}^0 = 0$ and $\omega_{ij}^0 \neq 0$ in the population model. The dotted vertical line marks the true population value. The Cn-PC estimators are shown

Figure 7: Distribution of $\hat{\omega}_{i,j}$



in the black colored histogram.

In the top two panels, the results show that for the PCEs estimates of the sample correlations, the distributions is almost symmetric around zero and fat tailed. The Cn-PC estimator estimates are much smaller and concentrated around a small average value. This observation is regardless of the true population value. This is due to the fact that the Cn-PC estimator is shrinking the average absolute value of these correlations. The shrinkage is not applied to each correlation coefficient.

The Cn-PC estimator' correlations are shrunk relative to the PCEs. This reduction in the size of the correlations is less significant for the case of $N = T = 150$, although the spread is still smaller.

Figure 8 shows the sampling distribution of maximum average cross correlation $\hat{\tau}^*$. The results show that overall, the estimated $\hat{\tau}^*$ based on Cn-PC estimator are lower than those based on PCEs. In the first panel with $N = 50$ and $T = 100$, $\hat{\tau}^*$ for Cn-PC

Figure 8: Sampling distribution of $\hat{\tau}^*$

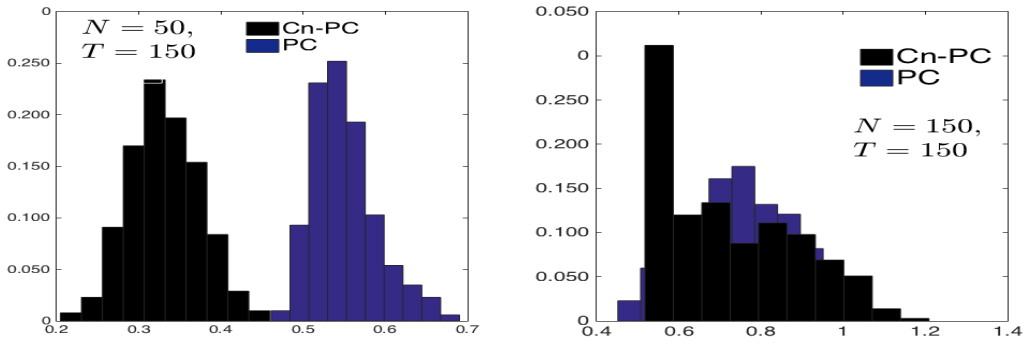
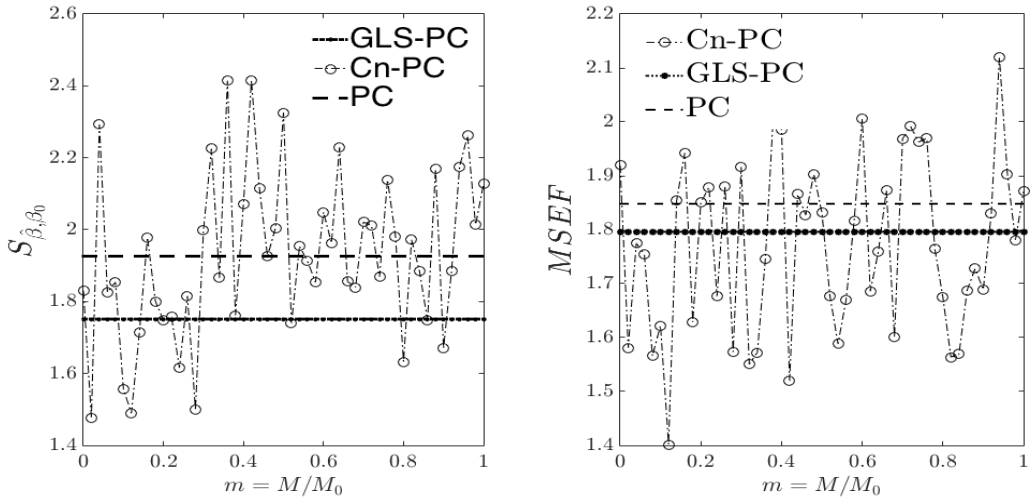


Figure 9: Accuracy of the 'Diffusion Index' forecasts



estimator support ranges from 0.02 to 0.46, while for PCEs, the range starts from 0.47 to 0.66.

These results depend on the panel dimension. For $T = N = 150$, the results are less promising, although the distribution of $\hat{\tau}^*$ is skewed to the left favoring lower values.

Simulated forecasts

Figure 9 displays the statistics $S_{\hat{y},y}$ and $S_{\hat{\beta},\beta_0}$ indexed by $m = M/M_0$. The dotted-dashed circle line plots the results for Cn-PC estimator, while the benchmark PCEs are shown in the straight dashed line. The plot also shows results for the weighted-PC estimator (Boivin and Ng [2006]) which uses as weights, w_{iT} equal to the inverse of $N^{-1}\sum_{j=1}^{N}|\hat{\Omega}_{ij}|$ for each error e_{it} in the PC objective function. The results correspond to a panle with $T = 120$ and $N = 130$ to reflect the panel dimensions that are encountered in macroeconomic forecasting and arbotrage pricing applications. The plots correspond to averages over 1000 replications.

As expected, the weighted-PC estimator outperforms the PCEs with smaller values of $S_{\hat{y},y}$ and $S_{\hat{\beta},\beta_0}$. The shrinkage factor $M = m \cdot M_0$ matters for the performance of the

Table 3: Pseudo-out-of-sample mean squared forecasts errors for US inflation and industrial production

		IPS10				PUNEW			
		$r=10$		$r=5$		$r=10$		$r=5$	
h=12		PC	Cn-PC	PC	Cn-PC	PC	Cn-PC	PC	Cn-PC
1970-2002	$MSFE$	0.51	0.51	0.52	0.50	0.64	0.62	0.57	0.57
	Var	0.85	0.85	0.66	0.66	0.53	0.53	0.60	0.60
1970-1985	$MSFE$	0.32	0.31	0.31	0.31	0.43	0.40	0.38	0.38
	Var	0.95	0.94	0.75	0.75	0.45	0.45	0.56	0.56
1985-2002	$MSFE$	1.09	1.08	1.13	1.11	1.65	1.63	1.46	1.40
	Var	0.53	0.50	0.39	0.43	0.87	0.85	0.77	0.75

		IPS10				PUNEW			
		$h=1$		$h=4$		$h=1$		$h=4$	
r=7		PC	Cn-PC	PC	Cn-PC	PC	Cn-PC	PC	Cn-PC
1970-2002	$MSFE$	0.72	0.70	0.57	0.57	0.78	0.75	0.67	0.67
	Var	0.42	0.38	0.56	0.56	0.27	0.27	0.37	0.37
1970-1985	$MSFE$	0.66	0.61	0.49	0.49	0.75	0.71	0.56	0.55
	Var	0.46	0.43	0.56	0.56	0.26	0.25	0.42	0.41
1985-2002	$MSFE$	0.86	0.86	0.86	0.86	0.82	0.82	0.97	0.97
	Var	0.28	0.28	0.54	0.54	0.28	0.28	0.25	0.25

Cn-PC. Unlike the resuls we have documented earlier with respect to the accuracy of the factors, there is no pattern to the relationship between M and the diffusion index forecasts. But the results, show that, for small values of m, the Cn-PC can outperform the weighted-PC by sizable margins.

6 Empirical Example

This section applies the Cn-PC estimator to a forecasting experiment for the U.S. Index of Industrial Production (IPS10) and Consummer Price Index (PUNEW) using the dataset provided by Stock and Watson [2002a]. The data include real variables such as sectoral industrial production, employment and hours worked; nominal variables such as consumer and price indexes, wages, money aggregates; in addition to stock prices and exchange rates. The data series are transformed to achieve stationarity: monthly growth rates for real variables(industrial production, sales $\cdots$) and first differences for variables already expressed in rates (unemployment rate, capacity utilization, $\cdots$). The dataset comprises of monthly observations from 1959:01 to 2003:12 and 131 time series. The sample is divided into an in-sample portion of size $T = 120$ (1959:01 to 1969:12) and an out-of-sample evaluation portion with first date December 1970 and last date December 2003. Therefore, there are a total of $M = 397$ out-of-sample evaluation points split into pre- and post-1985 periods with cat-off date December 1984. The models and parameters are re-estimated and the 12-step-ahead forecasts are computed for each month $t = T+12, \cdots T+12+M-1$ using a rolling window scheme that uses the most recent 10 years of monthly data, that is data indexed $t-12-T+1, \cdots, t-12$. In this empirical example, the Cn-PC estimator is computed using a threshhold parameter M that is chosen using a ten-fold cross-validation.

Table 3 reports the mean squared forecasts error (MSFE) relative to the random walk and the variance (var) of the forecasts relative to the variance of the series to be forecast. We consider three sample periods and consider different values for the forecast horizon

h. The number of factors r is selected using Bai and Ng [2002] information crieterion IC_{p_1}, which returns and estimate of $\hat{r} = 7$. We also show results for arbitrary values of $r = 5, 10$.

It is observed that the gains in forecasts accuracy depends on the sample period and on the target series. Generally, the gains are not significant and range from 0% to 6% decrease in the pseudo-out-of-sample mean-squared forecast errors.

Consumer price Index forecasts appear to benefit the most from incorporating dependence features using the Cn-PC estimators of the predictors $\hat{F}_t$. These benefits are more appreciable during the period of post moderation of 1985-2002. This is supported by the findings in the literature. During this period, predictability of the price and output series is problematic partly because of instabilities in the data and of the inflation targeting policy of the Federal Reserve Bank.

7 Conclusion

This paper proposes a novel PC-based method for incorporating features of cross corelation in the data in large factor models. The method allows for approximate factor structure in the sense of Chamberlain and Rothschild [1983] and embeds the assumption of bounded cross-sectional dependence as external information in the PC method. This constrained estimation is easily implemented within the classical principal components analysis. The method does not require estimation of large covariance matrices and works through a shrinkage mechanism applied to the sample cross covariances. The Monte Carlo results show that the Cn-PC estimator is generally more efficient than the PC and GLS-PC for large systems. Applied to real data, the results suggest that improvements in the accuracy of the estimated factors does not always leed to improvements into the forecasts accuracy but the results depend on the target series and on the forecast horizon.

Future research into inference using the Cn-PC estimator is warranted. Applications of the Cn-PC may include tests for the number of common factors and tests of hypothesis on the factor space.

Appendix A1: Cn-PC Estimators

The critical points of the function (3.10) are found by solving the first order conditions on the feasible set:

$$(I): \frac{\partial \mathcal{L}(\Lambda, F)}{\partial \Lambda}|_{\hat{\Lambda},\hat{F}} = 0 \tag{7.1}$$

$$(II): \frac{\partial \mathcal{L}(\Lambda, F)}{\partial F}|_{\hat{\Lambda},\hat{F}} = 0 \tag{7.2}$$

$$M \geq (NT)^{-1} \sum_{t=1}^{N} \underline{\hat{e}}_t' \mathcal{S} \underline{\hat{e}}_t, \quad \hat{\mu}_{NT} \geq 0, \quad \hat{\mu}_{NT} \left(M - (NT)^{-1} \sum_{t=1}^{N} \underline{\hat{e}}_t' \mathcal{S} \underline{\hat{e}}_t \right) = 0 \tag{7.3}$$

The conditions in (7.3) are known as the complementary slackness. The first two sets of conditions in (7.4) and (7.6), lead to the following:

$$(I) \quad : \quad \sum_{t=1}^{T} \left(I_N + \hat{\mu}_{NT}\mathcal{S}\right) \underline{\hat{e}}_t \hat{F}_t' = 0 \tag{7.4}$$

$$\hat{\Lambda} = \left(\sum_{t=1}^{T} \underline{X}_t F_t'\right)\left(\sum_{t=1}^{T} F_t F_t'\right)^{-1} \tag{7.5}$$

$$(II) \quad : \quad \sum_{t=1}^{T} \hat{\Lambda}' \left(I_N + \hat{\mu}_{NT}\mathcal{S}\right) \underline{\hat{e}}_t = 0 \tag{7.6}$$

$$\hat{F}_t = \left(\hat{\Lambda}' \left(I_N + \hat{\mu}_{NT}\mathcal{S}\right) \hat{\Lambda}\right)^{-1} \hat{\Lambda}' \left(I_N + \hat{\mu}_{NT}\mathcal{S}\right) \underline{X}_t \tag{7.7}$$

Substituting (7.5) into the Lagrangian and imposing the identification restriction $F'F/T = I_r$, this concentrates out Λ to obtain a reduced Lagrangian that is a function of F and μ:

$$\begin{aligned}
\mathcal{L}(\hat{F}, \hat{\mu}_{NT}, r) &= (NT)^{-1} \sum_{t=1}^{T} \underline{\hat{e}}_t' \underline{\hat{e}}_t - \mu \left[M/N - (N^2T)^{-1} \sum_{t=1}^{N} \underline{\hat{e}}_t' \mathcal{S} \underline{\hat{e}}_t \right] \\
&= (NT)^{-1} \text{trace } \left[\hat{e}\left(I_N + \hat{\mu}_{NT}\mathcal{S}\right)\hat{e}\right] - M \\
&= \frac{\text{trace } X\left(I_N + \hat{\mu}_{NT}\mathcal{S}\right)X'}{NT} - \frac{\text{trace } \hat{F}'\left(\mathbf{X}\left(I_N + \hat{\mu}_{NT}\mathcal{S}\right)\mathbf{X}'\right)\hat{F}}{NT} - \hat{\mu}_{NT} M
\end{aligned}$$

For a given $\hat{\mu}_{NT}$, the optimization problem is identical to maximizing trace $F'\left(\frac{\mathbf{X}(I_N + \hat{\mu}_{NT}\mathcal{S})\mathbf{X}'}{T}\right)F$ with respect to F. The estimated factor matrix, denoted by $\hat{F}_{\hat{\mu}_{NT}}$ to the latter problem is the matrix with columns consisting of the principal components of, $\mathbf{X}\left(I_N + \hat{\mu}_{NT}\mathcal{S}\right)\mathbf{X}'$. Technically, consider the spectral decomposition of the matrix of,

$$\Psi'_{N,\hat{\mu}} = \mathbf{X}\left(I_N + \hat{\mu}_{NT}\mathcal{S}\right)\mathbf{X}',$$

$$\Psi_{N,\hat{\mu}} \Gamma_{\hat{\mu}} = \Gamma_{\hat{\mu}} \Delta_{\hat{\mu}},$$

where $\Delta_{\hat{\mu}} = \text{diag}(d_{1,\hat{\mu}}, \cdots, d_{N,\hat{\mu}})$ is a diagonal matrix with $d_{a,\hat{\mu}}$ corresponding to the a^{th} highest eigenvalue of $\Psi_{N,\hat{\mu}}$, and $\Gamma_{\hat{\mu}} = (\varphi_{1,\hat{\mu}}, \cdots, \varphi_{N,\hat{\mu}})$ is the matrix whose columns corresponds to the normalized eigenvectors of $\Psi_{N,\hat{\mu}}$. The 'normalized' constrained PC estimators (Cn-PC estimator) of $\mathbf{F}(\hat{\mu})$ are $\widehat{F}_{k,t} = \frac{1}{\sqrt{d_{k,\hat{\mu}}}} \varphi'_{k,\hat{\mu}} \underline{X}_t$, for $k = 1, \cdots, r$.

To summarize,

$$\hat{F}_{\hat{\mu}_{NT}} \quad : \quad \sqrt{T} \times \text{first } r \text{ principal components of } \mathbf{X}\left(I_N + \hat{\mu}_{NT}\mathcal{S}\right))\mathbf{X}', \tag{7.8}$$

$$\hat{\Lambda}_{\hat{\mu}_{NT}} \quad : \quad \hat{\Lambda}_{\hat{\mu}_{NT}} = \mathbf{X}'\hat{F}_{\hat{\mu}_{NT}}/T, \tag{7.9}$$

$$\mu \quad : \quad M = (NT)^{-1} \sum_{t=1}^{N} \underline{\hat{e}}_t' \mathcal{S} \underline{\hat{e}}_t \tag{7.10}$$

I solve for $(\hat{F}_{\hat{\mu}}, \hat{\mu})$ which minimizes the reduced Lagrangian $\mathcal{L}(F, \mu)$ in (7.8) subject to the constraint $F'F/T = I_r$. The problem can be solve as in the standard primal-dual procedure, whereby the Lagrangian is further concentrated to a reduced function of μ, after replacing F by $\hat{F}(\mu)$. The dual problem solves the maximum of the concentrated objective function, $\mathcal{L}(\mu)$, which is equal to:

$$(NT)^{-1}\left[\text{tr } X\left(I_N + \hat{\mu}_{NT}\mathcal{S}\right)X' - \text{tr } \hat{F}'_{\hat{\mu}_{NT}}\left(\mathbf{X}\left(I_N + \hat{\mu}_{NT}\mathcal{S}\right)\mathbf{X}'\right)\hat{F}_{\hat{\mu}_{NT}}\right] - \hat{\mu}_{NT}M \quad (7.11)$$

APPENDIX B: Proofs of Main results

B.1. Proof of Theorem 1

The main results in this paper can be proven using some of the Lemma's of Bai and Ng [2002] and Bai [2003]. I will be therefore omit many of the details that are not worth reporting. In all of the proofs, I assume that the true number of factors (in the population) r is known.

Proof of Theorem 1 Let V_{NT} be an $r \times r$ matrix consisting of the largest eigenvalues of the matrix $\frac{1}{NT}\mathbf{X}\left(I_N + \mu_{NT}\mathcal{S}\right)\mathbf{X}'$ in descending order. Denote $\mathcal{A}_N \equiv I_N + \mu_{NT}\mathcal{S}$. The Cn-PC estimator estimates for the common factors $\hat{F}$ are defined as the eigenvectors corresponding to the largest eigenvalues of the matrix $\mathbf{X}\mathcal{A}_N\mathbf{X}'$ and thus satisfy the relation $\hat{F} = \frac{1}{NT}\mathbf{X}\mathcal{A}_N\mathbf{X}'\hat{F}V_{NT}^{-1}$ by the definition of eigenvalues and eigenvectors. Let the rotation matrix $\mathcal{H} = \left(\frac{\Lambda'\mathcal{A}_N\Lambda}{N}\right)\left(\frac{F'\hat{F}}{T}\right)V_{NT}^{-1}$. Then the following relation originates from Choi [2012] (if $\mathcal{A}_N$ is replaced with Ω^{-1})who generalized the original expressions in Bai [2003] and Bai and Ng [2002] (corresponding to $\mu_{NT} = 0$),

$$\begin{aligned}
\hat{F} - F\mathcal{H} &= \frac{1}{NT}\left(\mathbf{X}\mathcal{A}_N\mathbf{X}'\right)\hat{F}V_{NT}^{-1} - \frac{1}{NT}F\left(\Lambda'\mathcal{A}_N\Lambda\right)F'\hat{F}V_{NT}^{-1} \\
&= \frac{1}{NT}\left(\mathbf{X}\mathcal{A}_N\mathbf{X}' - F\left(\Lambda'\mathcal{A}_N\Lambda\right)F'\right)\hat{F}V_{NT}^{-1} \\
&= \frac{1}{NT}\left(\mathbf{e}\mathcal{A}_N\mathbf{e}' + \mathbf{e}\mathcal{A}_N\Lambda F' + F\Lambda'\mathcal{A}_N\mathbf{e}\right)\hat{F}V_{NT}^{-1}.
\end{aligned}$$

In vector form, the relation becomes,

$$\hat{F}_t - \mathcal{H}'F_t = \frac{1}{NT}V_{NT}^{-1}\hat{F}'\left(\mathbf{e}\mathcal{A}_N\underline{e}_t + F^0\Lambda'\mathcal{A}_N\underline{e}_t + \mathbf{e}\mathcal{A}_N\Lambda F_t^0\right) \quad (7.12a)$$

$$= V_{NT}^{-1}\left(\frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\underline{e}'_s\mathcal{A}_N\underline{e}_t + \frac{1}{NT}\sum_{s=1}^{T}\hat{F}_sF_s^{0'}\Lambda'\mathcal{A}_N\underline{e}_t + \frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\underline{e}'_s\mathcal{A}_N\Lambda F_t^0\right) \quad (7.12b)$$

$$= V_{NT}^{-1}\left(a_{NT,t} + b_{NT,t} + c_{NT,t}\right) \quad (7.12c)$$

Let $\mathcal{A}_{Nj}$ be the j^th column of the matrix $\mathcal{A}_N$ with elements $\mathcal{A}_{N,ij}$, then we can write:

$$\begin{aligned} a_{NT,t} = \frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\underline{e}_s'\mathcal{A}_N\underline{e}_t &= \frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\sum_{j=1}^{N}\sum_{i=1}^{N}e_{is}\mathcal{A}_{N,ij}e_{jt} \\ &= \frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\left[\sum_{l=1}^{N}\left(e_{ls}e_{lt}+\mu_{NT}\sum_{k\neq l}^{N}\mathcal{S}_{il}e_{is}e_{lt}\right)\right] \end{aligned}$$

Note that the latter comes from the fact that the elements of $\mathcal{A}_N = I_N + \mu_{NT}\mathcal{S}$ are equal to $\mathcal{A}_{N,ii} = 1$ for $i = 1, \cdots, N$ and $\mathcal{A}_{N,ij} = \mu_{NT}\mathcal{S}_{ij}$ for $1 \leq i \neq j \leq N$.

$$\begin{aligned} a_{NT,t} &= \frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\sum_{l=1}^{N}e_{ls}e_{lt}+\mu_{NT}\frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\sum_{k\neq l}^{N}\mathcal{S}_{il}e_{is}e_{lt} \\ &= \frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\sum_{l=1}^{N}e_{ls}e_{lt}+\mu_{NT}\frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\left(\sum_{l=1}^{N}\sum_{i=1}^{N}\mathcal{S}_{il}e_{is}e_{lt}-\sum_{l=1}^{N}e_{ls}e_{lt}\right) \\ &= \frac{1}{NT}\sum_{s=1}^{T}\sum_{l=1}^{N}\hat{F}_se_{ls}e_{lt}(1-\mu_{NT})+\mu_{NT}\frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\sum_{l=1}^{N}\sum_{i=1}^{N}\mathcal{S}_{il}e_{is}e_{lt} \\ &= \left[\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\gamma_N(s,t)+\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\varsigma_{st}\right](1-\mu_{NT})+\mu_{NT}\left[\frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\sum_{l=1}^{N}\sum_{i=1}^{N}\mathcal{S}_{il}e_{is}e_{lt}\right], \end{aligned}$$

where $\gamma_N(s,t) = E\left(N^{-1}\sum_{i=1}^{N}e_{it}e_{is}\right)$ and $\varsigma_{st} = \frac{\underline{e}_s'\underline{e}_t}{N} - \gamma_N(s,t)$ are defined as in Bai and Ng [2002]. Similarly, we can write:

$$b_{NT,t} = \left[\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\eta_{st}\right](1-\mu_{NT})+\mu_{NT}\left[\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\left(\frac{\Lambda^{0'}(I_N+\mathcal{S})\underline{e}_t}{N}\right),\right]$$

where $\eta_{st} = F_s^{0'}\Lambda^{0'}\underline{e}_t/N$. The last term $c_{NT,t}$ is equal to $b_{NT,t}$ since, $\underline{e}_s'\Lambda F_t^0/N = \eta_{st}$. Using the Cauchy-Schwarz inequality, that states $(\sum_{s=1}^{m}z_s)^2 \leq m\sum_{s=1}^{m}z_s^2$, we have

$$\|\hat{F}_t - \mathcal{H}'F_t^0\|^2 \leq 3\left(\|a_{NT,t}\|^2 + \|b_{NT,t}\|^2 + \|c_{NT,t}\|^2\right),$$

and

$$\frac{1}{T}\sum_{t=1}^{T}\|\hat{F}_t - \mathcal{H}'F_t^0\|^2 \leq \frac{3}{T}\sum_{t=1}^{T}\left(\|a_{NT,t}\|^2 + \|b_{NT,t}\|^2 + \|c_{NT,t}\|^2\right).$$

Now

$$\begin{aligned} \|a_{NT,t}\|^2 &\leq 3(1-\mu_{NT})^2T^{-2}\|\sum_{s=1}^{T}\hat{F}_s\gamma_N(s,t)\|^2 + 3(1-\mu_{NT})^2T^{-2}\|\sum_{s=1}^{T}\hat{F}_s\varsigma_{st}\|^2 \\ &+ 3\mu_{NT}^2T^{-2}\|\sum_{s=1}^{T}\hat{F}_s\sum_{l=1}^{N}\sum_{i=1}^{N}\mathcal{S}_{il}e_{is}e_{lt}/N\|^2. \end{aligned}$$

Bai and Ng [2002] in the proof of their Theorem 1 in page 213, show that

$$\begin{aligned} T^{-1}\sum_{t=1}^{T}\|T^{-1}\sum_{s=1}^{T}\hat{F}_s\gamma_N(s,t)\|^2 &= O_p\left(T^{-1}\right), \\ T^{-1}\|\sum_{s=1}^{T}\hat{F}_s\varsigma_{st}/T\|^2 &= O_p\left(N^{-1}\right). \end{aligned}$$

For the last term in $\sum_{t=1}^{T} a_{NT,t}/T$:

$$\begin{aligned} T^{-1}\sum_{t=1}^{T}\|\sum_{s=1}^{T}\hat{F}_s\sum_{l=1}^{N}\sum_{i=1}^{N}\mathcal{S}_{il}e_{is}e_{lt}/NT\|^2 &= \\ \sum_{s=1}^{T}\hat{F}_s\sum_{l=1}^{N}\sum_{i=1}^{N}\mathcal{S}_{il}e_{is}e_{lt}/NT &= \frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\xi_N(s,t)+\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\varrho_N(s,t) \end{aligned}$$

where

$$\xi_N(s,t) = N^{-1}\underline{e}_s'\mathcal{S}\underline{e}_t - \varrho_N(s,t)$$

Now,

$$\begin{aligned} \|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\varrho_N(s,t)\| &\leq \left(\frac{1}{T}\sum_{s=1}^{T}\|\hat{F}_s\|^2\right)^{1/2}\left(\frac{1}{T}\sum_{s=1}^{T}\left|\varrho_N(s,t)\right|^2\right)^{1/2} \\ &= O_p(1)\cdot O\left(\frac{1}{\sqrt{T}}\right) \end{aligned}$$

because of the normalization $\hat{F}'\hat{F}/T = I_r$ and Assumption A5(1). Thus, $\frac{1}{T}\sum_{t=1}^{T}\|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\varrho_N(s,t)\|^2 = O_p\left(\frac{1}{T}\right)$. Now,

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}\|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\xi_N(s,t)\|^2 &\leq \frac{1}{T}\left(\frac{1}{T}\sum_{s=1}^{T}\|\hat{F}_s\|\right)^{1/2}\left[\frac{1}{T^2}\sum_{s=1}^{T}\sum_{s'=1}^{T}\left(\sum_{t=1}^{T}\xi_N(s,t)\xi_N(s',t)\right)^2\right]^{1/2} \\ &\leq \frac{1}{T}O_p(1)\cdot\frac{T}{N} = O_p\left(\frac{1}{N}\right) \end{aligned}$$

since $E\left(\sum_{t=1}^{T}\xi_N(s,t)\xi_N(s',t)\right)^2 \leq T^2\ \max_{s,t}\ E\left|\xi_N(s,t)\right|^4$, and from Assumption A5(2),

$$E\left|\xi_N(s,t)\right|^4 = \frac{1}{N^2}E\left|N^{-1/2}\left[\underline{e}_s'\mathcal{S}\underline{e}_t - E(\underline{e}_s'\mathcal{S}\underline{e}_t)\right]\right|^4 \leq N^{-2}M.$$

To summarize,

$$\frac{1}{T}\sum_{t=1}^{T}\|a_{NT,t}\|^2 = \left[O_p\left(\frac{1}{T}\right)+O_p\left(\frac{1}{N}\right)\right](2\mu_{NT}^2 - 2\mu_{NT}+1)$$

For $b_{NT,t}$,

$$\frac{1}{T}\sum_{t=1}^{T}\|b_{NT,t}\|^2 \leq \frac{1}{T}\sum_{t=1}^{T}\|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\eta_{st}\|^2(1-\mu_{NT})^2+\mu_{NT}^2\frac{1}{T}\sum_{t=1}^{T}\|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\left(\frac{F_s^{0\prime}\Lambda^{0\prime}\mathcal{S}\underline{e}_t}{N}\right)\|^2.$$

The proof of Theorem 1 in Bai and Ng [2002] show that: $\frac{1}{T}\sum_{t=1}^{T}\|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\eta_{st}\|^2 = O_p\left(N^{-1}\right)$. Consider the second term,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\left(\frac{F_s^{0\prime}\Lambda^{0\prime}\mathcal{S}\underline{e}_t}{N}\right)\right\|^2 &\leq \frac{1}{NT}\sum_{t=1}^{T}\left[\left(\frac{1}{T}\sum_{s=1}^{T}\|\hat{F}_s\|^2\right)\left(\frac{1}{T}\sum_{s=1}^{T}\|F_s^0\|^2\right)\left\|\frac{\Lambda^{0\prime}\mathcal{S}\underline{e}_t}{\sqrt{N}}\right\|^2\right]\\ &= O_p\left(\frac{1}{N}\right),\end{aligned}$$

because of Assumption A5(3), and Assumption A1(1). Thus,

$$\frac{1}{T}\sum_{t=1}^{T}\|b_{NT,t}\|^2 = O_p\left(\frac{1}{N}\right)\left[\mu_{NT}^2+(\mu_{NT}-1)^2\right]$$

Combining all the results, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t-\mathcal{H}'F_t\right\|^2 &= \left[O_p\left(\frac{1}{T}\right)+O_p\left(\frac{1}{N}\right)\right](2\mu_{NT}^2-2\mu_{NT}+1)\\ &= V_{NT}^{-1}\left[O_p\left(\delta_{NT}^{-2}\right)+O_p\left(\mu_{NT}^2\delta_{NT}^{-2}\right)\right].\end{aligned}$$

The last step is to characterize the convergence of the matrix V_{NT},

$$\begin{aligned}\|V_{NT}\| &= \frac{1}{T}\left\|\hat{F}'\left(\mathbf{X}\mathcal{A}_N\mathbf{X}'\right)\hat{F}\right\|\\ &\leq \left\|\hat{F}'\hat{F}/T\right\|\|\mathbf{X}\mathcal{A}_N\mathbf{X}'/N\|\\ &\leq O_p(1)\cdot\mu_N^2\left(\|\mathbf{X}\mathbf{X}'/N\|\,\|\mathbf{X}\mathbf{X}'/N\|\right)\\ &= \mu_N^2O_p(1).\end{aligned}$$

At last,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left\|V_{NT}^{-1}\left(\hat{F}_t-\mathcal{H}'F_t\right)\right\|^2 &\leq \|V_{NT}\|^2\left[\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t-\mathcal{H}'F_t\right\|^2\right]\\ &\leq \mu_{NT}^{-2}\left[O_p\left(\delta_{NT}^{-2}\right)+O_p\left(\mu_{NT}^2\delta_{NT}^{-2}\right)\right].\end{aligned}$$

and thus,

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\left(\hat{F}_t-\mathcal{H}'F_t\right)\right\|^2 \leq \left[O_p\left(\mu_{NT}^{-2}\delta_{NT}^{-2}\right)+O_p\left(\delta_{NT}^{-2}\right)\right].$$

B.2. Proof of Theorem 2

From (7.12),

$$\begin{aligned}\hat{F}_t - \mathcal{H}'F_t &= V_{NT}^{-1}\left(\frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\underline{e}_s'\mathcal{A}_N\underline{e}_t + \frac{1}{NT}\sum_{s=1}^{T}\hat{F}_sF^{0'}\Lambda'\mathcal{A}_N\underline{e}_t + \frac{1}{NT}\sum_{s=1}^{T}\hat{F}_s\underline{e}_s'\mathcal{A}_N\Lambda F_t^0\right)\\ &= V_{NT}^{-1}\left(a_{NT,t} + b_{NT,t} + c_{NT,t}\right)\\ &= V_{NT}^{-1}\left[I + \mu_N II\right],\end{aligned}$$

where

$$I = \frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\gamma_N(s,t) + \frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\varsigma_{st} + \frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\eta_{st} + \frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\xi_{st},$$

where $\xi_{st} = F_t^{0'}\Lambda^0\underline{e}_s/N$, and

$$II = \frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\frac{\underline{e}_t'\mathcal{S}\underline{e}_s}{N} + \frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\left(\frac{F_s^{0'}\Lambda^{0'}\mathcal{S}\underline{e}_t}{N}\right) + \frac{1}{T}\sum_{s=1}^{T}\hat{F}_s\left(\frac{F_t^{0'}\Lambda^{0'}\mathcal{S}\underline{e}_s}{N}\right).$$

Lemma 2. *If $\mu_{NT} = O(1)$. Then*

$$\hat{F}_t - \mathcal{H}'F_t = V_{NT}^{-1}\left[O_p\left(\frac{1}{\sqrt{T}\omega_{NT}}\right) + O_p\left(\frac{1}{\sqrt{N}\omega_{NT}}\right) + O_p\left(\frac{1}{\sqrt{N}}\right) + O_p\left(\frac{1}{\sqrt{N}\omega_{NT}}\right)\right] \tag{7.13}$$

Lemma 2 follows from the earlier result of Theorem 1 and the proof can be carried out in similar way as that of [Bai, 2003, Lemma A.2 pages. 159–160].

The limiting distribution is determined by the dominant term in the expression 7.13 which depends on the panel dimensions and on the tuning parameter.

Lemma 3. *Let $\sqrt{N}/T\mu_{NT} \to 0$. Then under Assumptions A1-A7,*

$$\sqrt{N}\left(\hat{F}_t - \mathcal{H}'F_t\right) = V_{NT}^{-1}\left(\frac{\sum_{s=1}^{T}\hat{F}_sF_s^{0'}}{T}\right)\left[\left(\frac{\Lambda^{0'}\underline{e}_t}{\sqrt{N}}\right) + \mu_{NT}\left(\frac{\Lambda^{0'}\mathcal{S}\underline{e}_t}{\sqrt{N}}\right)\right] + o_p(1) \tag{7.14}$$

We have $\Lambda^{0'}\mathcal{S}\underline{e}_t/\sqrt{N} = O_p(1)$ by Assumption 5(3) and $\mu_{NT} = o_p(1)$ by Assumption 7(1) thus

$$\sqrt{N}\left(\hat{F}_t - \mathcal{H}'F_t\right) = V_{NT}^{-1}\left(\frac{\sum_{s=1}^{T}\hat{F}_sF_s^{0'}}{T}\right)\left(\frac{\Lambda^{0'}\underline{e}_t}{\sqrt{N}}\right) + o_p(1) \tag{7.15}$$

By Assumption A6(3),

$$\left(\frac{\Lambda^{0'}\underline{e}_t}{\sqrt{N}}\right) \xrightarrow{d} N(0, \Psi_t).$$

Lemma 4. *Under Assumptions A1-A5,*

(i)

$$V_{NT} = \frac{1}{T}\hat{F}' \left(\frac{\mathbf{X}\mathcal{A}_N\mathbf{X}'}{TN} \right) \hat{F} \xrightarrow{p} V,$$

(ii)

$$\frac{\hat{F}'F^0}{T} \left(\frac{\Lambda^{0'}\mathcal{A}_N\Lambda^0}{N} \right) \frac{\hat{F}'F^0}{T} \xrightarrow{p} V$$

Lemma 5. *Under Assumptions A1-A4 and A7,*

$$plim_{T,N\to\infty} \frac{\hat{F}'F^0}{T} = \mathcal{Q},$$

where $\mathcal{Q}$ is an invertible $r \times r$ matrix given by $\mathcal{Q} = V^{1/2}\Upsilon\Sigma_{\Lambda*}^{-1/2}$, with V consisting of eigenvalues (in descending order) of $\Sigma_{\Lambda*} \cdot \Sigma_F$ and Υ is the corresponding matrix of eigenvectors.

Proof. The result in lemma 5 can be proven using the same methods as in the proof of Proposition 1 in Bai [2003]. Key elements of the proof. By Lemma 4(ii) and $\mathbf{X} = F^0\Lambda^{0'} + e$, we have (respectively):

$$\left(\frac{\Lambda^{0'}\mathcal{A}_N\Lambda^0}{N} \right)^{1/2} T^{-1}F^{0'} \left(\frac{\mathbf{X}\mathcal{A}_N\mathbf{X}'}{N} \right) \hat{F} = \left(\frac{\Lambda^{0'}\mathcal{A}_N\Lambda^0}{N} \right)^{1/2} \left(\frac{F^{0'}\hat{F}}{T} \right) V_{NT},$$

and

$$(B_{NT} + d_{NT}R_{NT}^{-1})R_{NT} = R_{NT}V_{NT},$$

where $B_{NT} = \left(\frac{\Lambda^{0'}\mathcal{A}_N\Lambda^0}{N} \right)^{1/2} \left(\frac{F^{0'}F^0}{T} \right) \left(\frac{\Lambda^{0'}\mathcal{A}_N\Lambda^0}{N} \right)^{1/2}$, and $R_{NT} = \left(\frac{\Lambda^{0'}\mathcal{A}_N\Lambda^0}{N} \right)^{1/2} \left(\frac{F^{0'}\hat{F}}{T} \right)$. Let $\Upsilon_{NT} = R_{NT}V_{NT}^*$ with V_{NT}^* the matrix consisting of diagonal elements of $R_{NT}'R_{NT}$. Then $(B_{NT} + d_{NT}R_{NT}^{-1})\Upsilon_{NT} = \Upsilon_{NT}V_{NT}$, which implies that Υ_{NT} is an eigenvector of $B_{NT} + d_{NT}R_{NT}^{-1}$, and we have

$$\mathcal{Q} = \text{plim } \frac{F^{0'}\hat{F}}{T} = \text{plim } \left(\frac{\Lambda^{0'}\mathcal{A}_N\Lambda^0}{N} \right)^{-1/2} \Upsilon_{NT}V_{NT}^* = \Sigma_{\Lambda*}^{-1/2}\Upsilon V^{1/2},$$

where Υ is the eigenvectors for the matrix $B = \text{plim } B_{NT} + d_{NT}R_{NT}^{-1} = \Sigma_{\Lambda*}^{1/2}\Sigma_F\Sigma_{\Lambda*}^{1/2}$. **Proof** of Theorem 7.15 follows due to Lemma 2 and Lemma 4 and we have limiting distribution of $\sqrt{N}\left(\hat{F}_t - \mathcal{H}'F_t^0 \right)$ is therefore a $N(0, \Xi_t)$ where

$$\Xi_t = V^{-1}\mathcal{Q}\Psi_t\mathcal{Q}'V^{-1} \tag{7.16}$$

References

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–172.

Bai, J., Liao, Y., 2016. Efficient estimation of approximate factor models via penalized maximum likelihood. Journal of Econometrics.

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

Bai, J., Ng, S., 2003. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. Mimeo, University of Michigan.

Boivin, J., Ng, S., 2006. Are more data always better for factor analysis? Journal of Econometrics 132, 169–194.

Breitung, J., Tenhofen, J., 2011. GLS estimation of dynamic factor models. Journal of the American Statistical Association 106 (495), 1150–1166.

Brown, S. J., 1989. The number of factors in security returns. Journal of Finance 44, 1247–1262.

Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure and mean-variance analysis in large asset markets. Econometrica 51, 1305–1324.

Choi, I., 4 2012. Efficient estimation of factor models. Econometric Theory 28, 274–308.

Connor, G., Korajczyk, R. A., 1989. An intertemporal equilibrium beta pricing model. Review of Financial Studies 2 (3), 255–289.

Connor, G., Korajczyk, R. A., 1993. A test for the number of factors in an approximate factor model. Journal of Finance XLVIII (4), 1263–1291.

De Mol, C., Giannone, D., Reichlin, L., 2008. Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? Journal of Econometrics 146, 318–328.

Doz, C., Giannone, D., Reichlin, L., 2012. A quasi maximum likelihood approach for large approximate dynamic factor models. Review of Economics and Statistics (REStat) 94 (4), 1014–1024.

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75 (4).

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2004. The generalized dynamic factor model: consistency and rates. Journal of Econometrics 119, 231–245.

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

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2009. Opening the black box: Structural factor models with large cross-sections. Econometric Theory 25 (05), 1319–1347.

Hoerl, A. E., Kennard, R. W., Baldwin, K. F., 1975. Ridge regression: Some simulations. Communications in Statistics 4 (2), 105—123.

Kapetanios, G., 2010. A testing procedure for determining the number of factors in approximate factor models with large datasets. Journal of Business and Economics Statistics 28, 397–409.

Lam, C., Fan, J., 2009a. The Annals of Statistics 37 (6B), 4254–4278.

Lam, C., Fan, J., 2009b. Sparsistency and rates of convergence in large covariance matrix estimation. The Annals of Statistics 37 (6B), 4254–4278.

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. Journal Multivariate Analysis 88, 365–411.

Ledoit, O., Wolf, M., 2012. Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics 40 (2), 1024—-1060.

Onatski, A., 2010. Determining the number of factors from empirical distribution of eigenvalues. Review of Economics and Statistics 92, 1004–1016.

Osborne, M. R., Presnell, B., Turlach, B. A., 2000. On the LASSO and its dual. Journal of Computational and Graphical Statistics 9 (2), 319–337.

Stock, J., H., Watson, M., W., 1998. Diffusion indexes. NBER, Working Papers 6702.

Stock, J. H., Watson, M. ., 2006. Forecasting with many predictors. In Handbook of Economic Forecasting 1, 551–554.

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

Stock, J. H., Watson, M. W., 2002b. Macroeconomic forecasting using diffusion indexes. Journal of Business and Economic Statistics 20 (2), 147–162.