ARTICLE IN PRESS

Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Maximum likelihood estimation and inference for high dimensional generalized factor models with application to factor-augmented regressions☆

Fa Wang

School of Economics, Peking University, 5 Yiheyuan Road, Haidian District, Beijing, 100871, China

ARTICLE INFO

Article history:
Received 30 April 2018
Received in revised form 22 October 2020
Accepted 15 November 2020
Available online xxxx

JEL classification:
C13
C35

Keywords:
Factor model
Mixed measurement
Maximum likelihood
High dimension
Factor-augmented regression
Forecasting

ABSTRACT

This paper reestablishes the main results in Bai (2003) and Bai and Ng (2006) for generalized factor models, with slightly stronger conditions on the relative magnitude of N (number of subjects) and T (number of time periods). Convergence rates of the estimated factor space and loading space and asymptotic normality of the estimated factors and loadings are established under mild conditions that allow for linear, Logit, Probit, Tobit, Poisson and some other single-index nonlinear models. The probability density/mass function is allowed to vary across subjects and time, thus mixed models are also allowed for. For factor-augmented regressions, this paper establishes the limit distributions of the parameter estimates, the conditional mean, and the forecast when factors estimated from nonlinear/mixed data are used as proxies for the true factors.

1. Introduction

High dimensional factor models where a large number of time series are simultaneously driven by a small number of latent factors provide a powerful framework to analyze high dimensional data. Accompanied by an ever-increasing data size, the literature for this model recently experienced a wave of development. For example, Bai and Ng (2002) and Bai (2003) respectively show that utilizing the high dimensionality, we are able to consistently determine the number of factors and establish the asymptotic normality of the estimated factors and loadings. High dimensional factor models have also been successfully used in macroeconomic monitoring and forecasting, business cycle analysis, asset pricing, risk measurement, see for example Stock and Watson (2002, 2016), Bernanke et al. (2005), Ross (1976) and Campbell et al. (1997), to name a few.

The literature recently starts to pay attention to nonlinearity in factor models. Fan et al. (2019) extend linear trace regression to generalized trace regression to accommodate categorical dependent variables. Fan et al. (2017) consider unknown nonlinear relationship between the factors and the forecast target. However, when the relationship between the dependent variables and the factors is nonlinear (e.g., categorical data), results remain scant. Direct extension of existing

☆ We would like to thank the editor Jianqing Fan, the associate editor and two referees for their valuable comments and suggestions. We would also like to thank Jushan Bai, Mingli Chen, Xiaohong Chen, Ivan Fernandez-Val, Kunpeng Li, Giovanni Urga, Martin Weidner and seminar participants at Chinese University of Hong Kong, Cass Business School, University of Connecticut and 2017 Asian Meeting of Econometric Society for helpful comments and suggestions.

E-mail address: fa.wang@pku.edu.cn.

https://doi.org/10.1016/j.jeconom.2020.11.002
0304-4076/

theory, e.g., Bai (2003) and Bai and Li (2012, 2016), to categorical data is not feasible because essentially both methods are based on the covariance matrix of the continuously distributed dependent variable. This paper seeks to establish a new estimation and inferential theory for high dimensional generalized factor models. More specifically, consider the following single-index factor model: For $i = 1, \ldots, N$ and $t = 1, \ldots, T$,

$$x_{it} \sim g_{it}(\cdot \,|\pi_{it}^0). \tag{1}$$

x_{it} is the observed data for the ith subject at time t. $g_{it}(\cdot\,|\cdot)$ is some known probability (density or mass) function of x_{it} allowed to vary across i and t. Note that $g_{it}(\cdot\,|\cdot)$ is the conditional probability function. Weak cross-sectional and serial dependence of x_{it} is allowed. $\pi_{it}^0 = f_t^{0\prime}\lambda_i^0$, and f_t^0 and λ_i^0 are r dimensional vector of factors and loadings respectively. Both factors and loadings are unobservable. Both N and T are large. The number of factors r is known. How to determine the number of factors is studied in a separate paper.

For engineering, this model has been successfully used in data compression, visualization, pattern recognition and machine learning. For social sciences, this model also plays important role in psychology and education. For economics and finance, possible applications are partially listed below:

(1) Macroeconomic forecasting, factor-augmented vector autoregression and business cycle analysis: In these areas, common factors are predominantly estimated by principal components using continuous data, see Stock and Watson (2002), Bernanke et al. (2005) and Bai and Ng (2006). Little attention has been paid to the treatment of categorical or mixed measurement data even though many data sets are of this type. For example, let x_{1t} be the GDP, x_{2t} be the consumer confidence index (categorical), x_{3t} be the interest rate announcement of FOMC, etc, at time t. Let f_t^0 denote some macroeconomic factors, then x_{it} is nonlinearly linked to $\pi_{it}^0 = f_t^{0\prime}\lambda_i^0$ through some known link function. While mixed measurement data are quite informative, they cannot be directly handled by principal components estimation. This paper provides a rigorous solution to this issue.

(2) Credit risk analysis: Default correlation modeling has direct implications for CDO (collateralized debt obligations) pricing, bond portfolio management and commercial bank risk management. Intuitively, default correlation originates from common exposures to business cycle, monetary policy, market sentiment and other financial or sector factors. Factor models provide a parsimonious way for analyzing default correlation and underlies many risk models used in practice. In a representative case, $\pi_{it}^0 + e_{it}$ is the value of company i at time t, e_{it} is the idiosyncratic error term, f_t^0 is the common factors and x_{it} is nonlinearly linked to π_{it}^0. x_{it} could be rating category company i belongs to, or the binary variable describing the default event, or the credit spread of its bond, or its stock return, or its stock volatility at time t. For more details on default correlation modeling and estimation, see Schonbucher (2000), McNeil and Wendin (2007), Koopman and Lucas (2008), Koopman et al. (2008, 2011), Creal et al. (2014) and the references therein.

(3) Socio-economic status measurement: In development economics, health economics, welfare economics and economics of education, researchers frequently encounter the problem of measuring the socio-economic status (more specifically the wealth or consumption) of a household or an individual. A good measure, serving as either the explanatory or the dependent variable, is crucial for these studies. Direct accurate measures of household wealth or consumption usually are not available or not reliable. Instead, the survey data contains many reliable yet categorically distributed proxies, such as living conditions and ownership of durables or assets. Treating these proxies as the dependent variables and household wealth as the latent explanatory factor, household wealth could be estimated from the data of these proxies. For example, let x_{it} be the ith proxy of household t and let f_t^0 be the wealth of household t, then x_{it} is nonlinearly linked to $\pi_{it}^0 = f_t^{0\prime}\lambda_i^0$ through some known link function implied by economic theory. Filmer and Pritchett (2001) follow this approach to construct wealth index for estimating the effect of wealth on educational enrollments in India. The Filmer–Pritchett procedure simply extracts the factor from the binary proxies directly by principal component. Rigorously speaking, this procedure lacks of theoretical support and may lead to misleading results.

For all the above and future applications, it is in urgent need to develop a theoretically justified method for estimating the factors and loadings from high dimensional nonlinear/mixed data. It is also necessary to establish the asymptotic properties of the proposed estimator under the high dimensional setup. Such asymptotic properties are needed to characterize the conditions under which the estimation error is negligible when estimated factors are used as regressors and to construct confidence intervals when estimated factors represent economic indices.

This paper considers maximum likelihood for estimating the factors and loadings from nonlinear/mixed data. Both factors and loadings are treated as parameters to be estimated and a penalty function is added to the log-likelihood function to guarantee the uniqueness of the solution of the likelihood maximization problem. This paper establishes the convergence rates of the estimated factor space and loading space, and asymptotic normality of the estimated factors and loadings, given that the probability function satisfies some regularity conditions. These regularity conditions allow for linear, Logit, Probit, Tobit and Poisson models. Thus Bai (2003) is a special case of this paper. The probability function is also allowed to vary across i and t, thus a mixture of these models is allowed for. This paper also establishes the limit distributions of the parameter estimates, the conditional mean as well as the forecast for factor-augmented regression models when the estimated factors are used as proxies for the true factors. This result generalizes Bai and Ng (2006) to allow us using factors extracted from nonlinear/mixed data.

In the statistics literature, classic factor analysis has been successfully extended to categorical data and mixed data, see for example, Bartholomew (1980), Moustaki (1996), Bartholomew and Knott (1999), Moustaki (2000), Moustaki and Knott (2000) and Joreskog and Moustaki (2001), to name a few. All these papers assume N is fixed and much smaller

than T. While factors are typically of primary interest in economic applications, factors can not be consistently estimated under the fixed N large T setup. This limitation and the urgent need to handle high dimensional mixed data recently has motivated researchers to explore possible solution. Ng (2015) reviews alternative methods of constructing factors that can potentially be extended to categorical data and explores their numerical properties.

This paper provides a general theory for factor analysis of high dimensional nonlinear data. Since factors and loadings are treated as parameters to be estimated, the number of parameters tend to infinity as N and T tend to infinity jointly. This paper solves this problem by utilizing the fact that for factor model, the Hessian is asymptotically block diagonal and the tensor of third order derivatives is sparse. More specifically, elements in the diagonal blocks of the Hessian are $O_p(N)$ or $O_p(T)$ while elements in the off-diagonal blocks are $O_p(1)$. This paper shows that under relevant regularity conditions, the presence of these nonzero off-diagonal blocks has no effect on the asymptotic properties of the estimated factors and loadings. Asymptotic block diagonality of the Hessian also provides explanation for Bai (2003)'s results from the perspective of extremum estimation.

This paper's solution is reminiscent of the diagonalization approaches discussed in Cox and Reid (1987) and Lancaster (2000, 2002). The difference is that in this paper the diagonality comes from the factor structure and high dimensionality and holds only when N and T tend to infinity jointly, while in those papers the diagonality comes from artificial reparametrization. More recently, Fernandez-Val and Weidner (2016) and Chen et al. (2014, 2020) utilize asymptotic diagonality of the incidental parameter Hessian to derive the limit distributions of the regression coefficients and the average partial effects in nonlinear panel models. For the estimated factors and loadings, Chen et al. (2014, 2020) establish the average consistency, while this paper also establishes the convergence rates, the limit distributions and the effect of using estimated factors in factor-augmented regression.

The rest of the paper is organized as follows. Section 2 introduces notations and preliminaries. Section 3 discusses the assumptions. Section 4 presents the limit theory. Section 5 presents results for factor-augmented regressions. Section 6 introduces computation algorithms. Section 7 presents simulation results. Section 8 concludes. All proofs are relegated to the appendix.

2. Notations and preliminaries

The log-likelihood[1] function is

$$L(X\,|f,\lambda) = \sum_{i=1}^{N}\sum_{t=1}^{T} l_{it}(f_t'\lambda_i), \tag{2}$$

where $l_{it}(\pi_{it}) = \log g_{it}(x_{it}\,|\pi_{it})$ and $\pi_{it} = f_t'\lambda_i$, X is the $T\times N$ matrix of observed data and x_{it} is the element on the tth row and the ith column, $f = (f_1', \ldots, f_T')'$ a Tr dimensional vector and $\lambda = (\lambda_1', \ldots, \lambda_N')'$ is a Nr dimensional vector. $g_{it}(\cdot\,|\cdot)$ is allowed to vary across i and t, thus data following different models (e.g., discretely and continuously distributed time series) can be merged directly to extract common factors. We consider the following representative examples.

Example 1 (*Linear*). $l_{it}(f_t'\lambda_i) = -\frac{1}{2}(x_{it} - f_t'\lambda_i)^2$.

Example 2 (*Probit*). $l_{it}(f_t'\lambda_i) = x_{it}\log\Phi(f_t'\lambda_i) + (1-x_{it})\log(1-\Phi(f_t'\lambda_i))$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Example 3 (*Logit*). $l_{it}(f_t'\lambda_i) = x_{it}\log\Psi(f_t'\lambda_i) + (1-x_{it})\log(1-\Psi(f_t'\lambda_i))$, where $\Psi(\cdot)$ is the CDF of the logistic distribution.

Example 4 (*Tobit*). Suppose $x_{it} = x_{it}^*$ if $x_{it}^* > 0$ and $x_{it} = 0$ if $x_{it}^* \le 0$, where $x_{it}^* = f_t'\lambda_i + e_{it}$ and e_{it} is $N(0,1)$. The likelihood function is $l_{it}(f_t'\lambda_i) = -\frac{1}{2}(x_{it} - f_t'\lambda_i)^2\mathbf{1}(x_{it} > 0) + \log(1-\Phi(f_t'\lambda_i))\mathbf{1}(x_{it} = 0)$, where $\mathbf{1}(\cdot)$ is the indicator function.

Example 5 (*Poisson*). $l_{it}(f_t'\lambda_i) = -e^{f_t'\lambda_i} + kf_t'\lambda_i - \log k!$, because $P(x_{it} = k) = p(k,\lambda) = e^{-\lambda}\lambda^k/k!$ and $\lambda = e^{f_t'\lambda_i}$.

Let $\phi = (\lambda', f')'$, $F = (f_1, \ldots, f_T)'$, $\Lambda = (\lambda_1, \ldots, \lambda_N)'$. Similarly, for the true values of the factors and the loadings, let $f^0 = (f_1^{0\prime}, \ldots, f_T^{0\prime})'$, $\lambda^0 = (\lambda_1^{0\prime}, \ldots, \lambda_N^{0\prime})'$, $\phi^0 = (\lambda^{0\prime}, f^{0\prime})'$, $F^0 = (f_1^0, \ldots, f_T^0)'$ and $\Lambda^0 = (\lambda_1^0, \ldots, \lambda_N^0)'$. Both factors and loadings are treated as parameters and estimated by maximum likelihood. Note that for any F, Λ and any $r\times r$ invertible matrix G, FG and $\Lambda(G')^{-1}$ has the same likelihood as F and Λ. To uniquely fix F and Λ, we impose the normalization such that

$$\text{(i) } F'F \text{ is diagonal, (ii) } \Lambda'\Lambda \text{ is diagonal, (iii) } \frac{1}{T}F'F = \frac{1}{N}\Lambda'\Lambda,$$

i.e., the estimated factors and loadings are the solution of maximizing $L(X\,|f,\lambda)$ under constraints (i)–(iii).

[1] When x_{it} is cross sectionally or serially dependent, $L(X\,|f,\lambda)$ is the quasi-likelihood function.

As explained in Appendix A, the solution of this constraint maximization problem is the same as the solution of maximizing $Q(f,\lambda) = L(X|f,\lambda) + P(f,\lambda)$, where

$$P(f,\lambda) = -\frac{c}{8}NT\left\|diag(\frac{1}{N}\Lambda'\Lambda - \frac{1}{T}F'F)\right\|_F^2 - \frac{c}{2}NT\left\|ndiag(\frac{1}{N}\Lambda'\Lambda)\right\|_F^2 - \frac{c}{2}NT\left\|ndiag(\frac{1}{T}F'F)\right\|_F^2 \quad (3)$$

is a penalty function, $0 < c < b_L$ and b_L is lower bound of $\left|\partial_{\pi^2} l_{it}(\pi_{it})\right|$ as presented in Assumption 2(ii) below. $\|\cdot\|_F$ denotes Frobenius norm. $diag(\frac{1}{N}\Lambda'\Lambda - \frac{1}{T}F'F)$ is the diagonal matrix consisting of diagonal elements of $\frac{1}{N}\Lambda'\Lambda - \frac{1}{T}F'F$. $ndiag(\frac{1}{N}\Lambda'\Lambda)$ is the upper triangular matrix consisting of nondiagonal element of $\frac{1}{N}\Lambda'\Lambda$, i.e., the (p,q)th element of $ndiag(\frac{1}{N}\Lambda'\Lambda)$ equals zero if $p \geq q$ and equals the (p,q)th element of $\frac{1}{N}\Lambda'\Lambda$ if $p < q$. $ndiag(\frac{1}{T}F'F)$ is defined in the same way. We shall consider the estimated factors and loadings as the solution of maximizing $Q(f,\lambda)$ in asymptotic analysis. For numerical computation, the algorithms in Section 6 still solve the constraint maximization problem.

The normalization (i)–(iii) is slightly different from the classical normalization $\frac{1}{T}F'F = I_r$ and $\Lambda'\Lambda$ being diagonal. We choose this normalization because choosing this normalization is equivalent to adding penalty (3) to $L(X|f,\lambda)$ and with penalty (3), the Hessian matrix of $Q(f,\lambda)$ has some convenient special structure for analyzing its asymptotic behavior.[2] If we choose another normalization, all results of this paper still hold, except for a different rotation matrix.[3]

Throughout the paper, let $(N,T) \to \infty$ denote N and T going to infinity jointly, $\delta_{NT} = \min\{N^{\frac{1}{2}}, T^{\frac{1}{2}}\}$, $D_{NT} = \begin{bmatrix} N \times I_{Nr} & 0 \\ 0 & T \times I_{Tr} \end{bmatrix}$, $D_{TN} = \begin{bmatrix} T \times I_{Nr} & 0 \\ 0 & N \times I_{Tr} \end{bmatrix}$. $\xrightarrow{d}$ denotes convergence in distribution. "w.p.a.1" denotes "with probability approaching 1". For vectors, $\|\cdot\|$ denotes the Euclidean norm. For matrix A, $\rho_{\min}(A)$ denotes its smallest eigenvalue, and $\|A\|$, $\|A\|_F$, $\|A\|_1$, $\|A\|_\infty$ and $\|A\|_{\max}$ denotes its spectral norm, Frobenius norm, 1-norm, infinity norm and max norm respectively. When A has Nr rows, divide A into N blocks with each block containing r rows and let $[A]_{iq}$ denote the qth row in the ith block and $[A]_i = ([A]'_{i1}, \ldots, [A]'_{ir})'$ denote the ith block.

3. Assumptions

Assumption 1. (i) $T^{-1}F^{0\prime}F^0 \xrightarrow{p} \Sigma_F$ for some positive definite Σ_F. There exists $M > 0$ such that $\left\|f_t^0\right\| \leq M$ for all t.

(ii) $N^{-1}\Lambda^{0\prime}\Lambda^0 \xrightarrow{p} \Sigma_\Lambda$ for some positive definite Σ_Λ. There exists $M > 0$ such that $\left\|\lambda_i^0\right\| \leq M$ for all i.

Assumption 1(i) corresponds to Assumption A in Bai (2003). Factors are allowed to be dynamic with arbitrary dynamics. Assumption 1(ii) is exactly the same as Assumption B in Bai (2003), and ensures each factor has a nontrivial contribution. Note that here $\left\|f_t^0\right\|$ and $\left\|\lambda_i^0\right\|$ are assumed to be uniformly bounded. This assumption is the same as Bai and Li (2016), but stronger than Bai (2003), which only assumes uniform boundedness of $\mathbb{E}\left\|f_t^0\right\|^4$ and $\mathbb{E}\left\|\lambda_i^0\right\|^4$. In general, compactness of parameter space is quite common for nonlinear models, e.g., Newey and McFadden (1994), Jennrich (1969) and Wu (1981). Under the current setup, this assumption is necessary because the convergence rate (and hence limit distribution) of $\hat{f}_t$ is not uniform over the parameter space of f_t^0 if $\left|\partial_{\pi^2} l_{it}(f_t^{0\prime}\lambda_i^0)\right| \to 0$ as $\left\|f_t^0\right\| \to \infty$. In other words, in such cases the convergence rates of $\hat{f}_t$ will not be the same[4] for all t.

Let $\partial_\pi l_{it}(\pi_{it})$, $\partial_{\pi^2} l_{it}(\pi_{it})$ and $\partial_{\pi^3} l_{it}(\pi_{it})$ be the first, second and third order derivative of $l_{it}(\cdot)$ evaluated at π_{it}, respectively. When these derivatives are evaluated at π_{it}^0, we suppress the argument and simply write $\partial_\pi l_{it}$, $\partial_{\pi^2} l_{it}$ and $\partial_{\pi^3} l_{it}$.

Assumption 2. (i) $l_{it}(\cdot)$ is three times differentiable.

(ii) There exists $b_U > b_L > 0$ such that $b_L \leq -\partial_{\pi^2} l_{it}(\pi_{it}) \leq b_U$ within a compact space of π_{it}.

(iii) $\left|\partial_{\pi^3} l_{it}(\pi_{it})\right| \leq b_U$ within a compact space of π_{it}.

Assumption 2(i) imposes smoothness condition on the log-likelihood function. Assumption 2(ii) and (iii) assumes that the log-likelihood function is concave, the second order derivatives are bounded below and above, and the third order derivatives are bounded above. The boundedness of the second and third order derivatives is needed to control the remainder term in the expansion of the first order condition.[5] The boundedness from below of the second order

[2] See equations (A-4)–(A-5) in Appendix B for the Hessian of $L(X|f,\lambda)$ and equation (A-11) for the Hessian of $P(f,\lambda)$, and see Lemmas 2–3 for how the special structure of the Hessian is utilized to analyze its asymptotic properties.

[3] To show this, we first prove the results for this normalization, and then prove the results still hold after changing the rotation.

[4] For example, consider the case f_t^0 is one dimensional and $\left|\partial_{\pi^2} l_{it}(f_t^{0\prime}\lambda_i^0)\right|$ converges to zero monotonically as $f_t^0 \to \infty$. Let $t^* = \arg\max f_t^0$ and $t^{**} = \arg\min f_t^0$. Then convergence rate of $\hat{f}_{t^*}$ could be slower than $\hat{f}_{t^{**}}$ as $(N,T) \to \infty$.

[5] Newey and McFadden (1994) only require two times continuously differentiable because it expands the first order condition only to the second order and utilizes Lemma 2.4 to establish the convergence of the Hessian. In this paper we expand the first order condition to the third order and utilize the uniform boundedness of the third order derivatives to explicitly calculate the magnitude of the third order term. Lemma 2.4 in Newey and McFadden (1994) is no longer applicable here because the dimension of the parameter space and the dimension of the Hessian also tend to infinity.

derivatives together with boundedness of π_{it} are used to show consistency of the estimated factors and loadings. We verify in Appendix E that Logit, Probit, Poisson and Tobit all satisfy Assumption 2. For other models, readers can check accordingly.

Assumption 3. There exists $M > 0$ such that for all N and T:
(i) $\mathbb{E}(|\partial_\pi l_{it}|^\xi) \le M$ for some $\xi > 14$ and all i and t.
(ii) $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}(\gamma_N(s,t))^2 \le M$, where $\gamma_N(s,t) = N^{-1}\sum_{i=1}^{N}\mathbb{E}(\partial_\pi l_{is}\partial_\pi l_{it})$.
(iii) For every (t,s), $\mathbb{E}(N^{-\frac{1}{2}}\sum_{i=1}^{N}[\partial_\pi l_{is}\partial_\pi l_{it} - \mathbb{E}(\partial_\pi l_{is}\partial_\pi l_{it})])^2 \le M$.

Assumption 4. There exists $M > 0$ such that for some $\zeta > 2$ and for all N and T,
$\mathbb{E}(N^{-1}\sum_{i=1}^{N}\left\|T^{-\frac{1}{2}}\sum_{t=1}^{T}\partial_\pi l_{it}f_t^0\right\|^\zeta) \le M$,
$\mathbb{E}(T^{-1}\sum_{t=1}^{T}\left\|N^{-\frac{1}{2}}\sum_{i=1}^{N}\partial_\pi l_{it}\lambda_i^0\right\|^\zeta) \le M$.

Assumption 5. (i) $\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{t=1}^{T}(T^{-1}\sum_{t=1}^{T}\partial_{\pi^2}l_{it}f_t^0f_t^{0\prime})^{-1}f_t^0\partial_\pi l_{it}\partial_\pi l_{is}\right\|^2 \le M$ for any s and
$\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{t=1}^{T}(N^{-1}\sum_{i=1}^{N}\partial_{\pi^2}l_{it}\lambda_i^0\lambda_i^{0\prime})^{-1}\lambda_i^0\partial_\pi l_{it}\partial_\pi l_{jt}\right\|^2 \le M$ for any j.
(ii) $\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{t=1}^{T}(T^{-1}\sum_{t=1}^{T}\partial_{\pi^2}l_{it}f_t^0f_t^{0\prime})^{-1}\partial_\pi l_{it}f_t^0\lambda_i^{0\prime}\right\|^2 \le M$ and
$\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{t=1}^{T}(N^{-1}\sum_{i=1}^{N}\partial_{\pi^2}l_{it}\lambda_i^0\lambda_i^{0\prime})^{-1}\partial_\pi l_{it}\lambda_i^0f_t^{0\prime}\right\|^2 \le M$.
$\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{t=1}^{T}(T^{-1}\sum_{t=1}^{T}\partial_{\pi^2}l_{it}f_t^0f_t^{0\prime})^{-1}\partial_\pi l_{it}f_t^0\lambda_i^{0\prime}\partial_{\pi^2}l_{is}\right\|^2 \le M$ for any s and
$\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{t=1}^{T}(N^{-1}\sum_{i=1}^{N}\partial_{\pi^2}l_{it}\lambda_i^0\lambda_i^{0\prime})^{-1}\partial_\pi l_{it}\lambda_i^0f_t^{0\prime}\partial_{\pi^2}l_{jt}\right\|^2 \le M$ for any j.
(iii) for any i, $-T^{-1}\sum_{t=1}^{T}\partial_{\pi^2}l_{it}f_t^0f_t^{0\prime} \xrightarrow{p} \Sigma_{iF}$ and $T^{-\frac{1}{2}}\sum_{t=1}^{T}\partial_\pi l_{it}f_t^0 \xrightarrow{d} \mathcal{N}(0,\Omega_{iF})$ for some positive definite Σ_{iF} and Ω_{iF}.
(iv) for any t, $-N^{-1}\sum_{i=1}^{N}\partial_{\pi^2}l_{it}\lambda_i^0\lambda_i^{0\prime} \to \Sigma_{t\Lambda}$ and $N^{-\frac{1}{2}}\sum_{i=1}^{N}\partial_\pi l_{it}\lambda_i^0 \xrightarrow{d} \mathcal{N}(0,\Omega_{t\Lambda})$ for some positive definite $\Sigma_{t\Lambda}$ and $\Omega_{t\Lambda}$.

Assumptions 3–5 are generalization of Assumptions C, D and F in Bai (2003) in the current nonlinear setup. When the model is linear, $\partial_\pi l_{it}$ is the error term "e_{it}" and $\partial_{\pi^2}l_{it}$ is a constant, and Assumptions 3–5 reduce to Assumptions C, D and F in Bai (2003) respectively (with slight modification on the value of ξ and ζ and the statement of Assumption F1). As Bai (2003), distribution of x_{it} is allowed to be heterogeneous over i and t, and limited cross-sectional and serial dependence of x_{it} is also allowed. If x_{it} is independent over i and t conditional on the factors and loadings, Assumption 3(ii) and (iii), Assumptions 4 and 5 can be easily verified. If there is no conditional independence, these assumptions still can be verified provided certain weak dependence conditions are imposed on. We follow Bai (2003)'s treatment in presenting Assumptions 3–5.

Assumption 6. The eigenvalues of the $r \times r$ matrix $(\Sigma_F \cdot \Sigma_\Lambda)$ are different.

Assumption 7. $\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}(N+T)^{\frac{1}{\zeta}}}{\delta_{NT}} \to 0$ as $(N,T) \to \infty$.

Assumption 6 is a crucial identification condition and is the same as Assumption G in Bai (2003). It guarantees that there exist unique F and Λ such that $F\Lambda' = F^0\Lambda^{0\prime}$, $F'F$ and $\Lambda'\Lambda$ are diagonal and $F'F/T = \Lambda'\Lambda/N$. Assumption 7 is quite weak if ξ and ζ are large. Except for some well-designed mathematical counterexamples, Assumptions 3(i) and 5 indeed hold with very large ξ and ζ.[6]

4. Limit theory for estimated factors and loadings

For any F^0 and Λ^0, there exists unique G such that the normalized true parameters $F^G \equiv F^0G$ and $\Lambda^G \equiv \Lambda^0(G^{-1})'$ satisfy constraints (i)–(iii) in Section 2. More specifically, let $\rho_1^2 > \cdots > \rho_r^2$ be the eigenvalues of $N^{-1}T^{-1}(\Lambda^{0\prime}\Lambda^0)^{\frac{1}{2}}F^{0\prime}F^0(\Lambda^{0\prime}\Lambda^0)^{\frac{1}{2}}$ and Υ be the matrix of corresponding eigenvectors, and let $\mathcal{V} = diag(\rho_1^2,\ldots,\rho_r^2)$. Assumption 1 implies that $\mathcal{V}$ converges in probability to the diagonal matrix of eigenvalues of $\Sigma_\Lambda^{\frac{1}{2}}\Sigma_F\Sigma_\Lambda^{\frac{1}{2}}$ and Υ converges in probability to the matrix of eigenvectors of $\Sigma_\Lambda^{\frac{1}{2}}\Sigma_F\Sigma_\Lambda^{\frac{1}{2}}$. Let $G = (\frac{\Lambda^{0\prime}\Lambda^0}{N})^{\frac{1}{2}}\Upsilon\mathcal{V}^{-\frac{1}{4}}$, G converges in probability to a constant matrix and Assumption 6

[6] For linear model with heavy-tailed error term, Assumptions 3(i) and 4 could be restrictive when ξ and ζ are large. One promising direction is to relax Assumption 7.

guarantees G is unique for N and T large enough.[7] It can be easily verified that $F^G\Lambda^{G\prime} = F^0\Lambda^{0\prime}$ and

$$\frac{1}{T}F^{G\prime}F^G = \frac{1}{N}\Lambda^{G\prime}\Lambda^G = \mathcal{V}^{\frac{1}{2}}, \tag{4}$$

i.e., (F^G, Λ^G) satisfies the constraints (i)–(iii) in Section 2 and columns of F^G and Λ^G are ordered by their Euclidean norm, from the largest to the smallest. Similar to the notation in Section 2, let $F^G = (f_1^G, \ldots, f_T^G)'$, $\Lambda^G = (\lambda_1^G, \ldots, \lambda_N^G)'$, $f^G = (f_1^{G\prime}, \ldots, f_T^{G\prime})'$, $\lambda^G = (\lambda_1^{G\prime}, \ldots, \lambda_N^{G\prime})'$ and $\phi^G = (\lambda^{G\prime}, f^{G\prime})'$. By definition of F^G, it is easy to see that $f_t^G = G'f_t^0$ and $\lambda_i^G = G^{-1}\lambda_i^0$.

Let $B(\mathcal{D})$ denote the neighborhood $\|f\|_\infty \le \mathcal{D}$ and $\|\lambda\|_\infty \le \mathcal{D}$ for some $\mathcal{D}$ large enough such that the normalized true parameters ϕ^G lie in the interior[8] of $B(\mathcal{D})$. $B(\mathcal{D})$ can be considered as generalization of compact parameter space from fixed dimensional case to infinity dimensional case. Let $\hat{f} = (\hat{f}_1', \ldots, \hat{f}_T')'$ and $\hat{\lambda} = (\hat{\lambda}_1', \ldots, \hat{\lambda}_N')'$ be the solution of maximizing $Q(f, \lambda)$ within $B(\mathcal{D})$, and let $\hat{\pi}_{it} = \hat{f}_t'\hat{\lambda}_i$, $\hat{\phi} = (\hat{\lambda}', \hat{f}')'$, $\hat{F} = (\hat{f}_1, \ldots, \hat{f}_T)'$ and $\hat{\Lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_N)'$. Columns of $\hat{F}$ and $\hat{\Lambda}$ are sorted by their Euclidean norm in decreasing order.

In addition, let $S(\phi) = \partial_\phi Q(\phi)$, $S_\lambda(\phi) = \partial_\lambda Q(\phi)$ and $S_f(\phi) = \partial_f Q(\phi)$ denote the score, it follows that $S(\phi) = (S_\lambda'(\phi), S_f'(\phi))'$. Let $H(\phi) = \partial_{\phi\phi'}Q(\phi)$ denote the Hessian matrix. Decomposition of $H(\phi)$ and the expression of each component is presented in Appendix B. We suppress the argument when $S(\phi)$ and $H(\phi)$ are evaluated at ϕ^G, i.e., $S = S(\phi^G)$ and $H = H(\phi^G)$.

4.1. Consistency

There are two difficulties in establishing consistency. First, the number of parameters tends to infinity jointly with N and T. Thus the classical procedure for extremum estimators, e.g., Newey and McFadden (1994), is no longer applicable. Second, the parameters are present in both dimensions and the likelihood function is nonconcave with respect to the parameters. Thus it is not feasible to extend the proof strategy of large dimensional nonlinear panels to the current setup, because they either require there is only individual effects or time effects (see for example, Hahn and Newey (2004) and Hahn and Kuersteiner (2011)), or require global concavity of the likelihood function (Fernandez-Val and Weidner, 2016). Inspired by Lemma 1 of Chen et al. (2014), this paper solves the difficulties by utilizing the boundedness from below of $-\partial_{\pi^2} l_{it}(\pi_{it})$ over the compact parameter space.

Proposition 1 (*Average Consistency*). *Under Assumptions 1–3 and 6, as* $(N,T) \to \infty$, $\frac{1}{T}\left\|\hat{f} - f^G\right\|^2 = O_p(\frac{1}{\delta_{NT}})$ *and* $\frac{1}{N}\left\|\hat{\lambda} - \lambda^G\right\|^2 = O_p(\frac{1}{\delta_{NT}})$.

To move forward to convergence rates and limit distributions of the estimated parameters, we need to utilize the first order conditions $S(\hat{\phi}) = 0$. The following proposition proves that $S(\hat{\phi}) = 0$ w.p.a.1.

Proposition 2. *Under Assumptions 1–4, 6 and 7,* $S(\hat{\phi}) = 0$ *w.p.a.1.*

Proposition 2 is nontrivial because the dimension of $\hat{\phi}$ increases with N and T. When the parameter space is fixed dimensional, consistency of the estimated parameters and the assumption that the true parameters is an interior point of the parameter space together implies that the estimated parameters is also an interior point of the parameter space, and consequently the first order conditions are satisfied. However, when the dimension of the parameter space tends to infinity jointly with N and T, average consistency of $\hat{\phi}$ as proved in Proposition 1 is not enough to guarantee that $\hat{\phi}$ is an interior point of the parameter space. We need uniform consistency of $\hat{\phi}$.

Proposition 3 (*Uniform Consistency*). *Under Assumptions 1–4 and 6,*

(i) $\left\|\hat{\lambda} - \lambda^G\right\|_\infty = O_p(\frac{N^{\frac{2}{\xi}}T^{\frac{2}{\xi}}(N+T)^{\frac{1}{\zeta}}}{T^{\frac{1}{2}}})$, *(ii)* $\left\|\hat{f} - f^G\right\|_\infty = O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}(N+T)^{\frac{1}{\zeta}}}{N^{\frac{1}{2}}})$.

Note that normally ξ and ζ could be large, and in such case $\left\|\hat{\lambda} - \lambda^G\right\|_\infty$ and $\left\|\hat{f} - f^G\right\|_\infty$ is approximately $O_p(T^{-\frac{1}{2}})$ and $O_p(N^{-\frac{1}{2}})$, respectively. Thus these rates are more accurate than Bai (2003)'s Proposition 2 when ξ and ζ are large.

All subsequent results do not rely on Assumption 7 directly. They rely on Assumption 7 purely because they rely on Proposition 2. Bai and Ng (2002) and Bai (2003) do not need any condition on the relative magnitude of N and T because in the linear setup the principal component estimator is just the global maximum, i.e., Bai and Ng (2002) and Bai (2003) do not have the difficulty[9] we encounter here.

---

[7] Assumption 6 guarantees Υ is unique, thus G is also unique. If there are repeated eigenvalues among $\rho_1^2, \ldots, \rho_r^2$, Υ is not unique because the orthonormal basis for the eigenspace corresponding to repeated eigenvalue is identifiable up to an orthogonal transformation.

[8] Note that $\|f^G\|_\infty$ and $\|\lambda^G\|_\infty$ are bounded w.p.a.1, because f_t^0 and λ_i^0 are uniformly bounded and $\|G\|$ is bounded w.p.a.1. Thus $\|\phi^G\|_\infty < \mathcal{D}$ w.p.a.1 when $\mathcal{D}$ is large enough.

[9] If we can find a better strategy to handle this difficulty, then we may get rid of Assumption 7.

4.2. Convergence rates

Now we utilize the first order conditions. Using the integral form of the mean value theorem for vector-valued functions[10] to expand the first order conditions, we have $0 = S(\hat{\phi}) = S + \tilde{H} \times (\hat{\phi} - \phi^G)$, where $\tilde{H} = \int_0^1 H(\phi^G + s(\hat{\phi} - \phi^G))ds \equiv \int_0^1 H(s)ds$. It follows that

$$\hat{\phi} - \phi^G = -\tilde{H}^{-1}S, \text{ or equivalently} \tag{5}$$

$$\begin{pmatrix} N^{-\frac{1}{2}}(\hat{\lambda} - \lambda^G) \\ T^{-\frac{1}{2}}(\hat{f} - f^G) \end{pmatrix} = D_{NT}^{-\frac{1}{2}}(\hat{\phi} - \phi^G) = (NT)^{-\frac{1}{2}}(-D_{TN}^{-\frac{1}{2}}\tilde{H}D_{TN}^{-\frac{1}{2}})^{-1}D_{TN}^{-\frac{1}{2}}S, \tag{6}$$

where D_{NT} and D_{TN} are normalization matrices defined in Section 2. Given Assumption 4, it is easy to see that $\left\|D_{TN}^{-\frac{1}{2}}S\right\| = O_p((N+T)^{\frac{1}{2}})$. Utilizing the structure of $H(\phi)$ and eigenvalue perturbation technique, we show in the Appendix (Lemma 3) that the largest eigenvalue of $(-D_{TN}^{-\frac{1}{2}}H(\phi)D_{TN}^{-\frac{1}{2}})^{-1}$ is $O_p(1)$ uniformly within the neighborhood $B(\mathcal{D}) \cap \left\|D_{NT}^{-\frac{1}{2}}(\phi - \phi^G)\right\| \leq m$ for some $m > 0$. Since $\hat{\phi}$ lies in $B(\mathcal{D}) \cap \left\|D_{NT}^{-\frac{1}{2}}(\phi - \phi^G)\right\| \leq m$ w.p.a.1, this implies that $\left\|(-D_{TN}^{-\frac{1}{2}}\tilde{H}D_{TN}^{-\frac{1}{2}})^{-1}\right\|$ is $O_p(1)$. Thus we have:

Theorem 1 (*Average Rate*). *Under Assumptions 1–4, 6 and 7,* $\frac{1}{T}\left\|\hat{f} - f^G\right\|^2 = O_p(\frac{1}{\delta_{NT}^2})$ *and* $\frac{1}{N}\left\|\hat{\lambda} - \lambda^G\right\|^2 = O_p(\frac{1}{\delta_{NT}^2})$.

Theorem 1 establishes the convergence rate of the estimated factor space and the estimated loading space. In applications where estimated factors are used as proxies for the true factors, e.g., forecasting, portfolio construction, Theorem 1 provides the foundation for characterizing the effect of using estimated factors. In this paper, we shall use Theorem 1 to show the limit distributions of $\hat{\lambda}_i - \lambda_i^G$ and $\hat{f}_t - f_t^G$, and limit distribution of the parameter estimates in factor-augmented regressions.

Remark 1. The rate $O_p(\frac{1}{\delta_{NT}^2})$ is the same as the convergence rate in Theorem 1 of Bai and Ng (2002) for linear factor models. The latter is a sharp rate unless the error term is spherical, see Bai and Li (2012). Since our setup includes linear factor models as special cases, the average rate $O_p(\frac{1}{\delta_{NT}^2})$ in Theorem 1 is also sharp.

Remark 2. The key step for Theorem 1 is to show that $\left\|(-D_{TN}^{-\frac{1}{2}}H(\phi)D_{TN}^{-\frac{1}{2}})^{-1}\right\|$ is $O_p(1)$ uniformly within $B(\mathcal{D}) \cap \left\|D_{NT}^{-\frac{1}{2}}(\phi - \phi^G)\right\| \leq m$. Lemma 5 of Chen et al. (2014) proves similar result for the case of one factor. To generalize from one factor to multiple factors, there are some purely mathematical difficulties. This paper solves the difficulties in step (2) of Lemma 2 and Lemma 3. Step (1) of Lemma 2 is similar to (and inspired by) Lemma 5 of Chen et al. (2014).

4.3. Limit distributions

Now we proceed to the limit distributions of the estimated factors and loadings. First, it is not feasible to extend Bai (2003)'s method of deriving the limit distribution of $\hat{f}_t - f_t^G$ to the current nonlinear setup, because Bai (2003)'s method relies on expression A.1 in Appendix A of Bai (2003), a crucial decomposition identity that does not hold in nonlinear setup. Second, noting that $\hat{\lambda}_i$ can be regarded as the maximum likelihood estimator when $\hat{f}$ is used for f^G and vice versa, another choice is to expand the first order conditions $\sum_{t=1}^T \partial_\pi l_{it}(\hat{f}_t'\hat{\lambda}_i)\hat{f}_t = 0$ at λ_i^G and use Theorem 1 to study the effect of using $\hat{f}$ for f^G and $\hat{\lambda}$ for λ^G. When the model is linear, Bai (2003) uses this method to establish the limit distributions of $\hat{\lambda}_i - \lambda_i^G$. However, as explained in Appendix A, this method is not promising when the model is nonlinear.

To solve this problem, we expand the first order conditions to higher order.

$$0 = S(\hat{\phi}) = S + H \times (\hat{\phi} - \phi^G) + \frac{1}{2}R,$$

where $R = (R_\lambda', R_f')'$. R_λ and R_f is Nr and Tr dimensional with element $R_{\lambda,iq} = (\hat{\phi} - \phi^G)'\partial_{\phi\phi'\lambda_{iq}}Q(\phi_{iq}^*)(\hat{\phi} - \phi^G)$ and $R_{f,tq} = (\hat{\phi} - \phi^G)'\partial_{\phi\phi'f_{tq}}Q(\phi_{tq}^*)(\hat{\phi} - \phi^G)$ respectively. ϕ_{iq}^* and ϕ_{tq}^* are linear combinations of $\hat{\phi}$ and ϕ^G. Thus

$$\hat{\phi} - \phi^G = -H^{-1}S - \frac{1}{2}H^{-1}R, \tag{7}$$

$$\text{and } \hat{\lambda}_i - \lambda_i^G = [\hat{\phi} - \phi^G]_i = -[H^{-1}S]_i - \frac{1}{2}[H^{-1}R]_i. \tag{8}$$

[10] Note that the standard mean value theorem does not hold for vector-valued functions. For more details, also see Feng et al. (2013).

Utilizing the structure of H, we show in Appendix D.5 that

$$[H^{-1}S]_i = (\sum_{t=1}^{T} \partial_{\pi^2} l_{it} f_t^G f_t^{G\prime})^{-1} \sum_{t=1}^{T} \partial_{\pi} l_{it} f_t^G + O_p(N^{-\frac{1}{2}} T^{-\frac{1}{2}}). \tag{9}$$

The intuition behind Eq. (9) is that H is approximately block diagonal. If the Hessian is block diagonal, asymptotic behavior of the estimates for parameters within different blocks will not affect each other. Thus as long as the dimension of each block is fixed, whether the dimension of the whole Hessian tends to infinity does not matter. In current context, H is not block diagonal, but the elements in its diagonal blocks are much larger than the elements in its off-diagonal blocks ($O_p(N^{\frac{1}{2}})$ or $O_p(T^{\frac{1}{2}})$ versus $O_p(1)$). Based on this observation and the structure of H, we show that in the expansion of $[H^{-1}S]_i$, the extra terms resulting from those nonzero off-diagonal blocks together have order $O_p(N^{-\frac{1}{2}} T^{-\frac{1}{2}})$.

Based on the structure of H, Theorems 1 and 4 presented below, we show in Appendix D.5 that

$$\left\| [H^{-1}R]_i \right\| = O_p(\frac{N^{\frac{3}{\xi}} T^{\frac{3}{\xi}}}{\delta_{NT}^2}). \tag{10}$$

Thus if $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2} N^{\frac{3}{\xi}} T^{\frac{3}{\xi}} \to 0$, $\left\| [H^{-1}R]_i \right\|$ would be $o_p(T^{-\frac{1}{2}})$ and hence dominated by the first term on the right hand side of Eq. (9). Intuitively, the reason that the remainder term $[H^{-1}R]_i$ is asymptotically negligible is because the tensor of third order derivatives is sparse. For example, it is easy to see that $\sum_{i=1}^{N} \sum_{t=1}^{T} \partial_{\lambda_k \lambda_j f_s} l_{it}(\cdot) = 0$ if $k \neq j$, and $\sum_{i=1}^{N} \sum_{t=1}^{T} \partial_{\lambda_k f_l f_s} l_{it}(\cdot) = 0$ if $l \neq s$.

Proposition 4 (*Individual Rate*)**.** *Under Assumptions 1–4, 6 and 7,* $\left\| \hat{\lambda}_i - \lambda_i^G \right\| = O_p(\frac{1}{\delta_{NT}})$ *for each* i *and* $\left\| \hat{f}_t - f_t^G \right\| = O_p(\frac{1}{\delta_{NT}})$ *for each* t.

The rate $O_p(\frac{1}{\delta_{NT}})$ is not sharp, but enough for calculating the order of $[H^{-1}R]_i$. From Eqs. (9) and (10), and the symmetry between $\hat{\lambda}_i$ and $\hat{f}_t$, we have the following theorem.

Theorem 2 (*Individual Limit Distribution*)**.** *Under Assumptions 1–7,*

$$T^{\frac{1}{2}}(\hat{\lambda}_i - \lambda_i^G) \xrightarrow{d} \mathcal{N}(0, \bar{G}^{-1} \Sigma_{iF}^{-1} \Omega_{iF} \Sigma_{iF}^{-1} \bar{G}'^{-1}) \text{ if } \frac{T^{\frac{1}{2}}}{\delta_{NT}^2} N^{\frac{3}{\xi}} T^{\frac{3}{\xi}} \to 0,$$

$$N^{\frac{1}{2}}(\hat{f}_t - f_t^G) \xrightarrow{d} \mathcal{N}(0, \bar{G}' \Sigma_{t\Lambda}^{-1} \Omega_{t\Lambda} \Sigma_{t\Lambda}^{-1} \bar{G}) \text{ if } \frac{N^{\frac{1}{2}}}{\delta_{NT}^2} N^{\frac{3}{\xi}} T^{\frac{3}{\xi}} \to 0,$$

where $\bar{G} = plim G$, *and* Σ_{iF}, Ω_{iF}, $\Sigma_{t\Lambda}$ *and* $\Omega_{t\Lambda}$ *are defined in Assumption 5. Asymptotic variance of* $\hat{\lambda}_i$ *and* $\hat{f}_t$ *can be estimated by*

$$var_\lambda = T(\sum_{t=1}^{T} \partial_{\pi^2} l_{it}(\hat{f}_t' \hat{\lambda}_i) \hat{f}_t \hat{f}_t')^{-1} (\sum_{t=1}^{T} (\partial_{\pi} l_{it}(\hat{f}_t' \hat{\lambda}_i))^2 \hat{f}_t \hat{f}_t') (\sum_{t=1}^{T} \partial_{\pi^2} l_{it}(\hat{f}_t' \hat{\lambda}_i) \hat{f}_t \hat{f}_t')^{-1},$$

$$var_f = N(\sum_{i=1}^{N} \partial_{\pi^2} l_{it}(\hat{f}_t' \hat{\lambda}_i) \hat{\lambda}_i \hat{\lambda}_i')^{-1} (\sum_{i=1}^{N} (\partial_{\pi} l_{it}(\hat{f}_t' \hat{\lambda}_i))^2 \hat{\lambda}_i \hat{\lambda}_i') (\sum_{i=1}^{N} \partial_{\pi^2} l_{it}(\hat{f}_t' \hat{\lambda}_i) \hat{\lambda}_i \hat{\lambda}_i')^{-1}.$$

Theorem 2 not only allows discrete dependent variables but also allows the probability function to differ across individuals and time. The huge amount of discrete data in macroeconomic and financial studies thus can be utilized, either by themselves or merged with continuous data, to extract information on common shocks or the state of the economy or other relevant variables. In real applications, we may simply choose normal density for continuous x_{it}. For discrete x_{it}, specific parametric model is needed.

Theorem 2 allows us to construct confidence intervals for the true factor process. This is useful since in various applications factors represent economic indices. Theorem 2 also has implication for factor-augmented forecasting. Since the estimated factors will be used as proxies for true factors, the estimation error $\hat{f}_t - f_t^G$ will be reflected in the forecasting error. We shall study this in Section 5.

Remark 3. To have limit normal distribution, Bai (2003) assumes $T^{\frac{1}{2}}/N \to 0$ for estimated loadings and $N^{\frac{1}{2}}/T \to 0$ for estimated factors. It is not difficult to see that when ξ is large, our condition is approximately the same as Bai (2003)'s condition.

Remark 4. "$N^{\frac{3}{\xi}} T^{\frac{3}{\xi}}$" appears because we choose to calculate $\|R\|_1$ rather than $\|R\|$. If we choose to calculate $\|R\|$, then due to the presence of the term "$L1i$" in Lemma 9 in the Appendix, we need to calculate the exact rate of $\left\| \hat{\lambda} - \lambda^G \right\|_4$, which

seems infeasible (note that unlike the linear case, we do not have accurate analytical expression of $\hat{\lambda}_i - \lambda_i^G$). If the model is linear, then $\partial_{\pi^3} l_{it}(\cdot) = 0$ and "$L1i$" would disappear, then there is no need to calculate $\|R\|_1$ and "$N^{\frac{3}{\xi}} T^{\frac{3}{\xi}}$" in all results of this paper except for Proposition 3 would also disappear.

Remark 5. Let $\mathcal{V}^{-}$=$plim\mathcal{V}$. If the model is linear, $\bar{G}'\Sigma_{iF}\bar{G} = \bar{\mathcal{V}}^{\frac{1}{2}}$ and $\bar{G}^{-1}\Sigma_{t\Lambda}\bar{G}'^{-1} = \bar{\mathcal{V}}^{\frac{1}{2}}$, and the limit variance of $\hat{\lambda}_i - \lambda_i^G$ and $\hat{f}_t - f_t^G$ become $\bar{\mathcal{V}}^{-\frac{1}{2}}\bar{G}'\Omega_{iF}\bar{G}\bar{\mathcal{V}}^{-\frac{1}{2}}$ and $\bar{\mathcal{V}}^{-\frac{1}{2}}\bar{G}^{-1}\Omega_{t\Lambda}\bar{G}'^{-1}\bar{\mathcal{V}}^{-\frac{1}{2}}$ respectively. If $\Sigma_{iF} = \Omega_{iF}$ and $\Sigma_{t\Lambda} = \Omega_{t\Lambda}$, the limit variance of $\hat{\lambda}_i - \lambda_i^G$ and $\hat{f}_t - f_t^G$ becomes $\bar{G}^{-1}\Sigma_{iF}^{-1}\bar{G}'^{-1}$ and $\bar{G}'\Sigma_{t\Lambda}^{-1}\bar{G}$ respectively.

4.4. Relationship of G and Bai (2003)'s Rotation Matrix

Bai (2003)'s rotation matrix is $H_{Bai} \equiv \frac{\Lambda^{0\prime}\Lambda^0}{N}\frac{F^{0\prime}\tilde{F}}{T}\mathcal{V}_{NT}^{-1}$, where $\tilde{F} = \hat{F}\mathcal{V}_{NT}^{-\frac{1}{4}}$, $\mathcal{V}_{NT} = diag(\hat{\rho}_1^2, \ldots, \hat{\rho}_r^2)$ and $\hat{\rho}_1 > \cdots > \hat{\rho}_r$ are the singular values of $N^{-\frac{1}{2}}T^{-\frac{1}{2}}\hat{F}\hat{\Lambda}'$. G depends only on f^0 and λ^0, while H_{Bai} depends not only on f^0 and λ^0 but also on the dependent variable. Moreover, we show in Appendix D.6 that

Proposition 5. *Under Assumptions 1–4, 6 and 7,*

$$\|\mathcal{V}_{NT} - \mathcal{V}\| = O_p\left(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2}\right) \tag{11}$$

$$\left\|G\mathcal{V}_{NT}^{-\frac{1}{4}} - H_{Bai}\right\| = O_p\left(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2}\right). \tag{12}$$

Theorem 1 in Bai and Ng (2002) and Lemma A.1 in Bai (2003) show $\left\|\tilde{F} - F^0H_{Bai}\right\|$ is $O_p(\frac{T^{\frac{1}{2}}}{\delta_{NT}})$, while Theorem 1 shows $\left\|\hat{F} - F^0G\right\|$ is $O_p(\frac{T^{\frac{1}{2}}}{\delta_{NT}})$. Given expressions (11)–(12) and $\tilde{F} = \hat{F}\mathcal{V}_{NT}^{-\frac{1}{4}}$, it is easy to see that $\left\|\tilde{F} - F^0H_{Bai}\right\| \leq \left\|\hat{F} - F^0G\right\|\left\|\mathcal{V}_{NT}^{-\frac{1}{4}}\right\| + T^{\frac{1}{2}}O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2})$. Under Assumption 7, $O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}}) = o_p(1)$, thus the result of Bai and Ng is a corollary (and thus special case) of Theorem 1.

Corollary 1. *Under Assumptions 1–4, 6 and 7,* $\left\|\tilde{F} - F^0H_{Bai}\right\| = O_p(\frac{T^{\frac{1}{2}}}{\delta_{NT}})$.

Theorem 1 and Theorem 2 in Bai (2003) show that $N^{\frac{1}{2}}(\tilde{f}_t - H_{Bai}'f_t^0)$ and $T^{\frac{1}{2}}(\tilde{\lambda}_i - H_{Bai}^{-1}\lambda_i^0)$ have limit normal distribution, while Theorem 2 shows that $N^{\frac{1}{2}}(\hat{f}_t - G'f_t^0)$ and $T^{\frac{1}{2}}(\hat{\lambda}_i - G^{-1}\lambda_i^0)$ has limit normal distribution. Since $\tilde{f}_t - H_{Bai}'f_t^0 = \mathcal{V}_{NT}^{-\frac{1}{4}}(\hat{f}_t - G'f_t^0) + (G\mathcal{V}_{NT}^{-\frac{1}{4}} - H_{Bai})'f_t^0$, expressions (11)–(12) and the condition $\frac{N^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \to 0$ imply that Bai's result is a corollary (and thus special case) of Theorem 2.

Corollary 2. *Under Assumptions 1–7,*

$$T^{\frac{1}{2}}(\tilde{\lambda}_i - H_{Bai}^{-1}\lambda_i^0) \xrightarrow{d} \mathcal{N}(0, \bar{\mathcal{V}}^{\frac{1}{4}}\bar{G}^{-1}\Sigma_{iF}^{-1}\Omega_{iF}\Sigma_{iF}^{-1}\bar{G}'^{-1}\bar{\mathcal{V}}^{\frac{1}{4}}) \text{ if } \frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \to 0,$$

$$N^{\frac{1}{2}}(\tilde{f}_t - H_{Bai}'f_t^0) \xrightarrow{d} \mathcal{N}(0, \bar{\mathcal{V}}^{-\frac{1}{4}}\bar{G}'\Sigma_{t\Lambda}^{-1}\Omega_{t\Lambda}\Sigma_{t\Lambda}^{-1}\bar{G}\bar{\mathcal{V}}^{-\frac{1}{4}}) \text{ if } \frac{N^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \to 0.$$

5. Inference and forecasting for factor-augmented regressions

In this section we shall use the results and techniques developed in Section 4 to study the effect of using estimated factors on factor-augmented regressions. Consider the following factor-augmented regression model:

$$y_{t+h} = \alpha' f_t^0 + \beta' W_t + \epsilon_{t+h}, \tag{13}$$

where f_t^0 is a r dimensional vector of factors, W_t is a q dimensional vector of other variables and h is the lead time between the dependent variable and information available. W_t and y_{t+h} are both observable. f_t^0 is unobservable, but a large number of predictors $x_{it}(i = 1, \ldots, N; t = 1, \ldots, T)$ are observable and can be used to estimate f_t^0. The probability function of x_{it} is $g_{it}(\cdot\,|f_t^{0\prime}\lambda_i^0)$, as introduced in Section 1. $g_{it}(\cdot\,|\cdot)$ satisfies the regularity conditions listed in Assumption 2.

When y_{t+h} is a scalar and $x_{it} = f_t^{0\prime}\lambda_i^0 + e_{it}$, this is the "diffusion index forecasting model" of Stock and Watson (2002). When $h = 1$ and $y_{t+1} = (f_{t+1}^{0\prime}, W_{t+1}')'$, this is the FAVAR of Bernanke et al. (2005). When $h = 0$, y_t is a scalar and x_{it} is discretely distributed, this is the model considered in Filmer and Pritchett (2001). When y_{t+h} is a scalar and x_{it} is discretely

distributed for some i and continuously distributed for the other i, this model can be used to analyze and forecast credit risk.

We shall use $\hat{F}$ as proxy for F^0. The objective is to characterize the effect of using $\hat{F}$ for F^0 on the limit distributions of the parameter estimates, the conditional mean as well as the forecast. Bai and Ng (2006) study this effect when the factors are estimated by principal components and $x_{it} = f_t^{0\prime}\lambda_i^0 + e_{it}$. The results in this section generalize Bai and Ng (2006)'s results to allow x_{it} to have nonlinear relationship with the factors for all or some i.

Assumption 8. Let $z_t = (f_t^{0\prime}, W_t')'$. $\mathbb{E}\|W_t\|^{\xi} \leq M$ and $\mathbb{E}(\epsilon_t^{\xi}) \leq M$ for some $\xi > 14$ and all t. $\mathbb{E}(\epsilon_{t+h}|y_t, z_t, y_{t-1}, z_{t-1}, \dots) = 0$ for all $h > 0$. ϵ_t is independent with x_{is} for all i and s. Furthermore,

(i) $T^{-1}\sum_{t=1}^{T} z_t z_t' \xrightarrow{p} \Sigma_{zz}$,

(ii) $T^{-\frac{1}{2}}\sum_{t=1}^{T} z_t\epsilon_{t+h} \xrightarrow{d} \mathcal{N}(0, \Sigma_{zz\epsilon})$, where $\Sigma_{zz\epsilon} = plim T^{-1}\sum_{t=1}^{T}\epsilon_{t+h}^2 z_t z_t'$.

(iii) $\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{t=1}^{T}(N^{-1}\sum_{i=1}^{N}\partial_{\pi^2} l_{it}\lambda_i^0\lambda_i^{0\prime})^{-1}\partial_{\pi} l_{it}\lambda_i^0 W_t'\right\|^2 \leq M$,

$\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{t=1}^{T}(N^{-1}\sum_{i=1}^{N}\partial_{\pi^2} l_{it}\lambda_i^0\lambda_i^{0\prime})^{-1}\partial_{\pi} l_{it}\lambda_i^0\epsilon_{t+h}\right\|^2 \leq M$.

Assumption 8 corresponds to Assumption E in Bai and Ng (2006). Part (i) and part (ii) are exactly the same as part (1) and (2) of Assumption E in Bai and Ng (2006). Bai and Ng (2006) also assume that W_t and ϵ_t are independent with "e_{is}" for all i and s, where "e_{is}" is the error term. The independence between ϵ_t and x_{is} here corresponds to their independence between ϵ_t and "e_{is}". The second condition of Assumption 8(iii) is not difficult to verify using the independence between ϵ_t and x_{is}. The first condition of Assumption 8(iii) corresponds to the independence between W_t and "e_{is}" in Bai and Ng (2006).

We shall only consider the case where y_t is a scalar. When y_t is a vector, the results are conceptually the same. Let $\hat{z}_t = (\hat{f}_t', W_t')'$ and $\theta = ((G^{-1}\alpha)', \beta')'$. Let $\hat{\theta} = (\hat{\alpha}', \hat{\beta}')'$ be the least squares estimator of regressing y_{t+h} on $\hat{z}_t$, i.e., $\hat{\alpha}$ is an estimate of $G^{-1}\alpha$.

Theorem 3 (*Inference*). *Under Assumptions 1–4, 6–8, assume* $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{4}{\xi}} \to 0$ *as* $(N, T) \to \infty$,

$$T^{\frac{1}{2}}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma_{\theta}),$$

where $\Sigma_{\theta} = \bar{\Xi}^{-1}\Sigma_{zz}^{-1}\Sigma_{zz\epsilon}\Sigma_{zz}^{-1}\bar{\Xi}'^{-1}$ *and* $\bar{\Xi} = diag(\bar{G}, I_q)$. *A consistent estimator of* Σ_{θ} *is* $\hat{\Sigma}_{\theta} = (T^{-1}\sum_{t=1}^{T-h}\hat{z}_t\hat{z}_t')^{-1}(T^{-1}\sum_{t=1}^{T-h}\hat{\epsilon}_{t+h}^2\hat{z}_t\hat{z}_t')(T^{-1}\sum_{t=1}^{T-h}\hat{z}_t\hat{z}_t')^{-1}$.

Theorem 3 implies that using the estimated factors does not affect the limit distribution of $\hat{\theta}$ when the factors are estimated by maximum likelihood and the probability function of x_{it} satisfy Assumption 2. Theorem 3 generalizes Theorem 1 of Bai and Ng (2006) to allow factors to be extracted from discrete or some other nonlinear data. This generalization should be valuable as in many factor-augmented regressions the information about the common factors are contained in discrete or mixed data. Theorem 3 provides theoretical support and guidance for exploiting these information.

For factor-augmented vector autoregression (FAVAR), the result and proof are conceptually the same. We do not repeat here. Thus Theorem 2 of Bai and Ng (2006) is also a special case of this paper.

Remark 6. Theorem 1 of Bai and Ng (2006) requires $T^{\frac{1}{2}}/N \to 0$. When ξ is large, the condition $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{4}{\xi}} \to 0$ is close to $T^{\frac{1}{2}}/N \to 0$.

Now consider forecasting for factor-augmented regression models. By Assumption 8, $\mathbb{E}(\epsilon_{t+h}|y_t, z_t, y_{t-1}, z_{t-1}, \dots) = 0$. Thus the conditional mean $y_{T+h|T}$ equals $\alpha' f_T^0 + \beta' W_T$. Let $\hat{y}_{T+h|T} = \hat{\theta}'\hat{z}_T$ be the forecast of $y_{T+h|T}$.

Theorem 4 (*Forecasting*). *Under Assumptions 1–8 and assume* $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{4}{\xi}} \to 0$ *and* $\frac{N^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \to 0$ *as* $(N, T) \to \infty$,

$$(\hat{y}_{T+h|T} - y_{T+h|T})/B_T \xrightarrow{d} \mathcal{N}(0, 1),$$

where $B_T^2 = T^{-1}z_T'\Sigma_{zz}^{-1}\Sigma_{zz\epsilon}\Sigma_{zz}^{-1}z_T + N^{-1}\alpha'\Sigma_{T\Lambda}^{-1}\Omega_{T\Lambda}\Sigma_{T\Lambda}^{-1}\alpha$. *A consistent estimator of* B_T^2 *is* $\hat{B}_T^2 = T^{-1}\hat{z}_T'\hat{\Sigma}_{\theta}\hat{z}_T + N^{-1}\hat{\alpha}'var_f^{-1}\hat{\alpha}$.

Theorem 4 generalizes Theorem 3 of Bai and Ng (2006) to allow factors to be extracted from discrete or some other nonlinear data. The variance of the estimated conditional mean has two components, one from the estimated parameters $\hat{\theta}$ and the other one from the estimated factors $\hat{f}_T$. Compared to cases where factors are observable, the presence of the latter component is the effect of using estimated factors on the estimated conditional mean.

Since $y_{T+h} = y_{T+h|T} + \epsilon_{T+h}$, the forecasting error is

$$\hat{\epsilon}_{T+h} = \hat{y}_{T+h|T} - y_{T+h|T} - \epsilon_{T+h}.$$

Given Theorem 4 and assume ϵ_t is $i.i.d.\mathcal{N}(0,\sigma_\epsilon^2)$, we have $\hat{\epsilon}_{T+h} \sim \mathcal{N}(0,\sigma_\epsilon^2 + var(\hat{y}_{T+h|T}))$. σ_ϵ^2 can be consistently estimated by $T^{-1}\sum_{t=1}^{T}\hat{\epsilon}_t^2$ and $var(\hat{y}_{T+h|T})$ can be consistently estimated by $\hat{B}_T^2$. Prediction intervals can be constructed correspondingly.

Remark 7. Theorem 3 of Bai and Ng (2006) requires $T^{\frac{1}{2}}/N \to 0$ and $N^{\frac{1}{2}}/T \to 0$. When ξ is large, the conditions $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{4}{\xi}} \to 0$ and $\frac{N^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \to 0$ are close to $T^{\frac{1}{2}}/N \to 0$ and $N^{\frac{1}{2}}/T \to 0$.

6. Algorithms

We shall introduce two algorithms, alternating maximization and minorization maximization, to numerically calculate the maximum likelihood estimator. The latter is computationally simpler, but so far we can only show it applies to Probit, Logit and Tobit. Whether it applies to more general models is unknown.

6.1. Alternating maximization (AM)

Algorithm. Step 1 (Initial values): Randomly generate initial values of the factors, $\hat{f}^{(0)}$.

Step 2 (Iterate): For $k = 0, \ldots$, calculate

$$\hat{\lambda}^{(k)} = \arg\max L(X \left| \hat{f}^{(k)}, \lambda \right.),$$
$$\hat{f}^{(k+1)} = \arg\max L(X \left| f, \hat{\lambda}^{(k)} \right.).$$

Iterate until $L(X \left| \hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)} \right.) - L(X \left| \hat{f}^{(k)}, \hat{\lambda}^{(k)} \right.) \le error$, where $error$ is the level of tolerated numerical error.

Step 3 (Repeat): Repeat step 1 and step 2 many times to get many local maximum. Take the one with the largest likelihood.

Step 4 (Normalize): Suppose $\hat{f}^{(s)}$ and $\hat{\lambda}^{(s)}$ be the estimator from step 3. Let $\hat{F}^{(s)} = (\hat{f}_1^{(s)}, \ldots, \hat{f}_T^{(s)})'$ and $\hat{\Lambda}^{(s)} = (\hat{\lambda}_1^{(s)}, \ldots, \hat{\lambda}_N^{(s)})'$. Let $\hat{V}^{(s)}$ be the diagonal matrix of eigenvalues of $N^{-1}T^{-1}(\hat{\Lambda}^{(s)\prime}\hat{\Lambda}^{(s)})^{\frac{1}{2}}\hat{F}^{(s)\prime}\hat{F}^{(s)}(\hat{\Lambda}^{(s)\prime}\hat{\Lambda}^{(s)})^{\frac{1}{2}}$ and $\hat{\Upsilon}^{(s)}$ be the corresponding matrix of eigenvectors, and let $\hat{G}^{(s)} = (\frac{1}{N}\hat{\Lambda}^{(s)\prime}\hat{\Lambda}^{(s)})^{\frac{1}{2}}\hat{\Upsilon}^{(s)}(\hat{V}^{(s)})^{-\frac{1}{4}}$. Choose $\hat{F} = \hat{F}^{(s)}\hat{G}^{(s)}$ and $\hat{\Lambda} = \hat{\Lambda}^{(s)}((\hat{G}^{(s)})^{-1})'$ as the solution of the likelihood maximization problem.

This algorithm is not totally new. In the machine learning literature, similar algorithm has been proposed in Collins et al. (2001) and Schein et al. (2003). The name "Alternating Maximization" comes from step 2, where we choose $\hat{\lambda}^{(k)}$ to maximize the likelihood for given $\hat{f}^{(k)}$ and then choose $\hat{f}^{(k+1)}$ to maximize the likelihood for given $\hat{\lambda}^{(k)}$. This is based on the fact that $L(X|f,\lambda)$ is globally concave with respect to λ for given f and vice versa. Because the likelihood is maximized alternately, we have $L(X \left| \hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)} \right.) \ge L(X \left| \hat{f}^{(k+1)}, \hat{\lambda}^{(k)} \right.) \ge L(X \left| \hat{f}^{(k)}, \hat{\lambda}^{(k)} \right.)$. Thus convergence of step 2 to a local maximum is guaranteed.

Whether the local maximum is global depends on the initial values $(\hat{f}^{(0)}, \hat{\lambda}^{(0)})$. To search the global maximum, a common practice is to randomly choose initial values many times and take the one with the largest likelihood among all local maximum. We follow this common practice in step 3. Step 4 normalizes the estimator from step 3 so that $\hat{F}'\hat{F}$ equals $\hat{\Lambda}'\hat{\Lambda}$ and both are diagonal.

6.2. Minorization maximization (MM)

Algorithm. Step 1 (Initial values): Randomly generate initial values of the factors and the loadings, $(\hat{f}^{(0)}, \hat{\lambda}^{(0)})$.

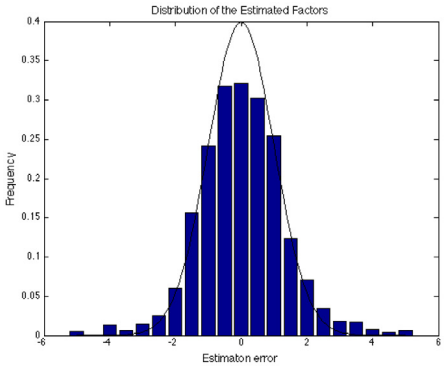
Step 2 (Iterate): For $k = 0, \ldots$, first calculate $\hat{x}_{it}^{(k)} = \hat{f}_t^{(k)\prime}\hat{\lambda}_i^{(k)} + \frac{1}{b_U}\partial_\pi l_{it}(\hat{f}_t^{(k)\prime}\hat{\lambda}_i^{(k)})$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, then $(\hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)}) = \arg\min \sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{x}_{it}^{(k)} - f_t'\lambda_i)^2$. Iterate until $L(X \left| \hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)} \right.) - L(X \left| \hat{f}^{(k)}, \hat{\lambda}^{(k)} \right.) \le error$, where $error$ is the level of tolerated numerical error.

Step 3 (Repeat): Repeat step 1 and step 2 many times to get many local maximum. Take the one with the largest likelihood.

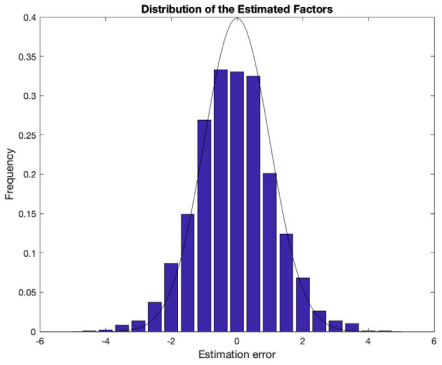
Step 4 (Normalize): Suppose $\hat{f}^{(s)}$ and $\hat{\lambda}^{(s)}$ be the estimator from step 3. Define $\hat{F}^{(s)}$, $\hat{\Lambda}^{(s)}$ and $\hat{G}^{(s)}$ in the same way as step 4 of the AM algorithm. Choose $\hat{F} = \hat{F}^{(s)}\hat{G}^{(s)}$ and $\hat{\Lambda} = \hat{\Lambda}^{(s)}((\hat{G}^{(s)})^{-1})'$ as the solution of the likelihood maximization problem.

Chen (2016) first proposes this algorithm for nonlinear panel models. Unlike the AM algorithm, for MM algorithm we do not need to do alternation. We only need to calculate the eigenvectors, which can be very fast using standard software package.

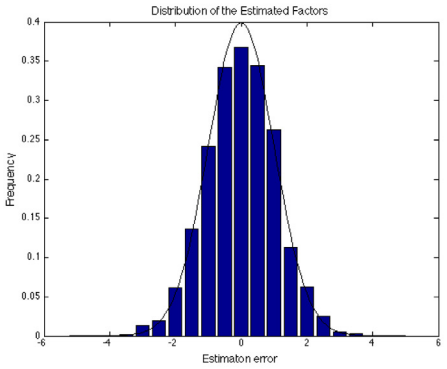
Minorization maximization is a class of algorithm similar to but more general than the expectation maximization (EM). See Appendix A(3) for a brief introduction and the proof for convergence of step 2 to local maximum. For more details on the MM algorithm, see Bohning and Lindsay (1988), de Leeuw (2006), Hunter and Lange (2004) and Lange et al. (2000), to name a few.



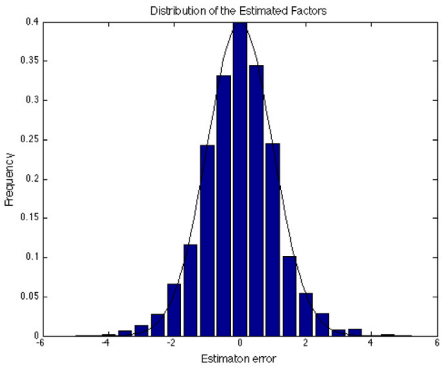
Logit, $N = 50, T = 50$.



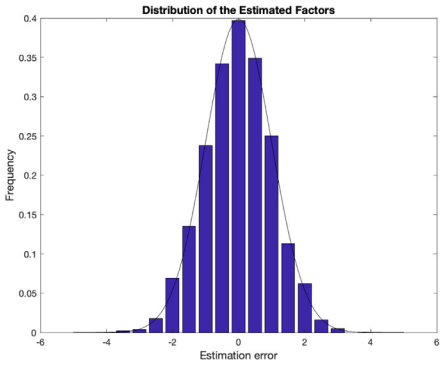
Probit, $N = 50, T = 50$.



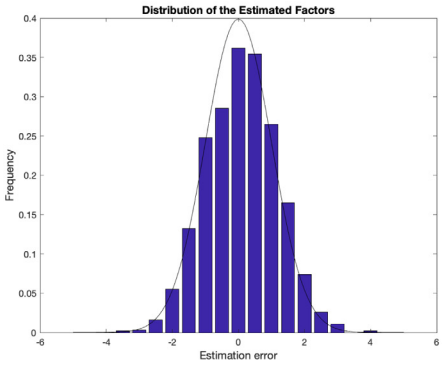
Logit, $N = 100, T = 100$.



Probit, $N = 100, T = 100$.



Logit, $N = 500, T = 200$

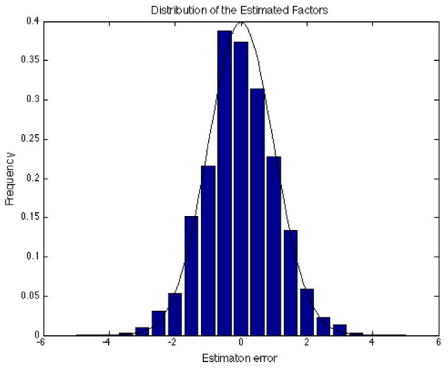


Probit, $N = 500, T = 200$

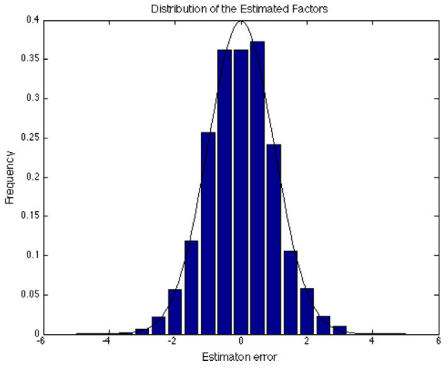
Fig. 1. Distribution of the estimated factors: (logit and probit). Notes: These histograms are for the standardized estimated factors. The curve overlaid on the histograms is the standard normal density function.

7. Simulations

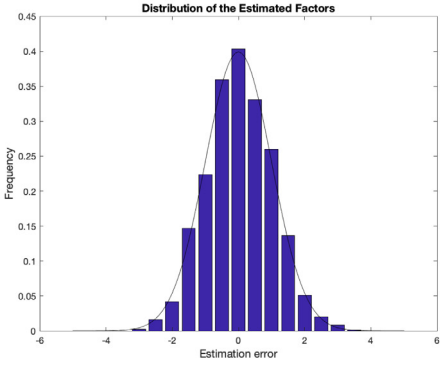
The main purpose of this section is to access the adequacy of the asymptotic distributions in approximating their finite sample counterparts. To allow graphically presenting the distribution of the estimated factors and loadings, we consider



Mixed, $N = 50, T = 50$.



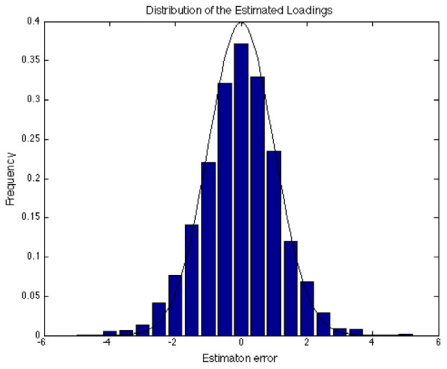
Mixed, $N = 100, T = 100$.



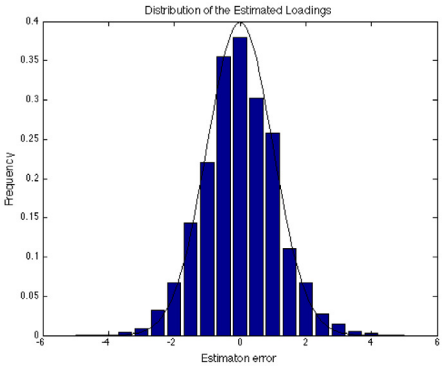
Mixed, $N = 500, T = 200$

Fig. 2. Distribution of the estimated factors (mixed). Notes: These histograms are for the standardized estimated factors. The curve overlaid on the histograms is the standard normal density function.

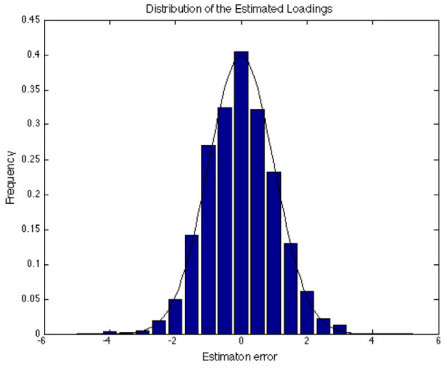
the case with one factor. For $i = 1, \ldots, N$ and $t = 1, \ldots, T$, f_t and λ_i are $i.i.d.\mathcal{N}(0, 1)$ and once generated, they are normalized to f_t^G and λ_i^G such that $\frac{1}{T}\sum_{t=1}^{T}(f_t^G)^2 = \frac{1}{N}\sum_{i=1}^{N}(\lambda_i^G)^2$. f_t^G and λ_i^G are fixed down for each simulation. For the given f_t^G and λ_i^G, we consider three data generating processes (DGPs) for x_{it}. Results for more DGPs, e.g. Poisson, Tobit or others, can be provided if requested.



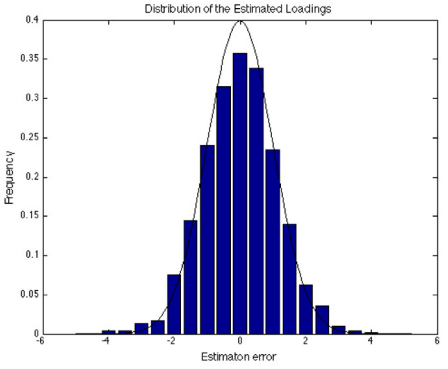
Logit, $N = 50, T = 50$.



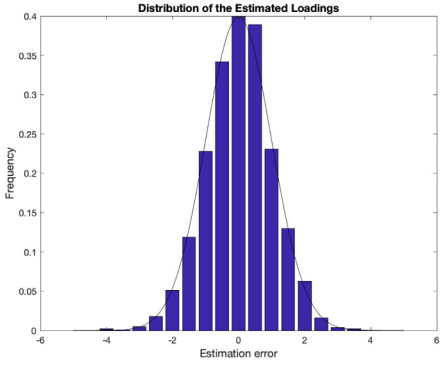
Probit, $N = 50, T = 50$.



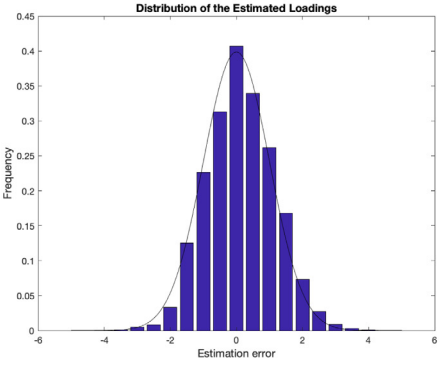
Logit, $N = 100, T = 100$.



Probit, $N = 100, T = 100$.



Logit, $N = 500, T = 200$

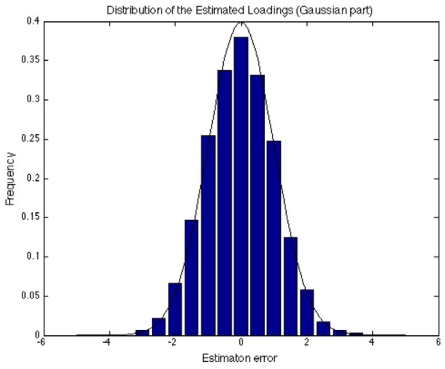


Probit, $N = 500, T = 200$

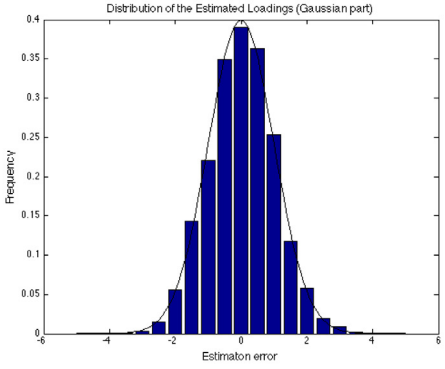
Fig. 3. Distribution of the estimated loadings (logit and probit). Notes: These histograms are for the standardized estimated loadings. The curve overlaid on the histograms is the standard normal density function.

DGP 1 (Logit): For $i = 1, \ldots, N$ and $t = 1, \ldots, T$, x_{it} is a binary random variable and $P(x_{it} = 1) = \Psi(f_t^G \lambda_i^G)$, where $\Psi(z) = 1/(1 + e^{-z})$.

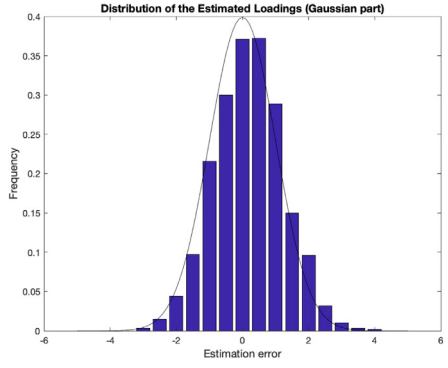
DGP 2 (Probit): For $i = 1, \ldots, N$ and $t = 1, \ldots, T$, x_{it} is a binary random variable and $P(x_{it} = 1) = \Phi(f_t^G \lambda_i^G)$, where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution.



Mixed (Gaussian part), $N = 50, T = 50$.



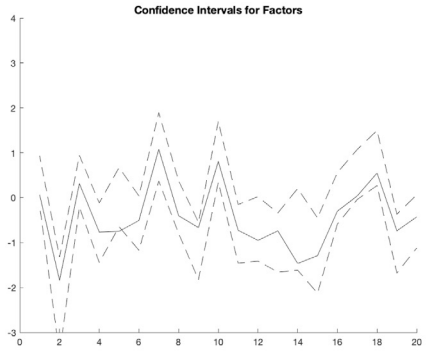
Mixed (Gaussian part), $N = 100, T = 100$.



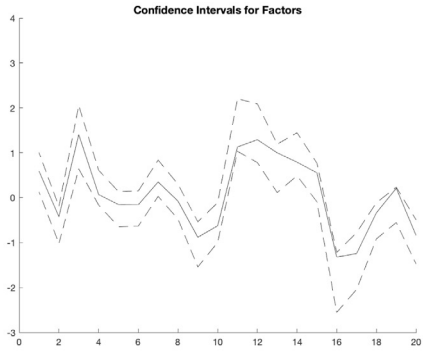
Mixed (Gaussian part), $N = 500, T = 200$

Fig. 4. Distribution of the estimated loadings (mixed: Gaussian part). Notes: These histograms are for the standardized estimated loadings. The curve overlaid on the histograms is the standard normal density function.

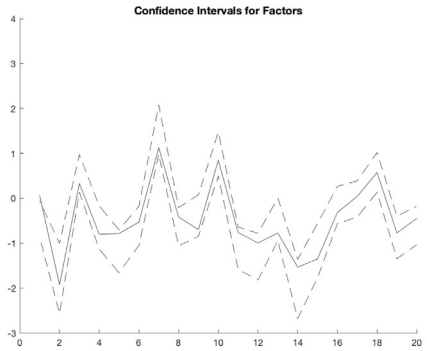
DGP 3 (Mixed): For $i = 1, \ldots, 2N/5$ and $t = 1, \ldots, T$, x_{it} is a binary random variable and $P(x_{it} = 1) = \Psi(f_t^G \lambda_i^G)$; for $i = 2N/5+1, \ldots, 4N/5$ and $t = 1, \ldots, T$, x_{it} is binary random variable and $P(x_{it} = 1) = \Phi(f_t^G \lambda_i^G)$; for $i = 4N/5+1, \ldots, N$ and $t = 1, \ldots, T$, x_{it} is normally distributed with mean $f_t^G \lambda_i^G$ and variance 1.



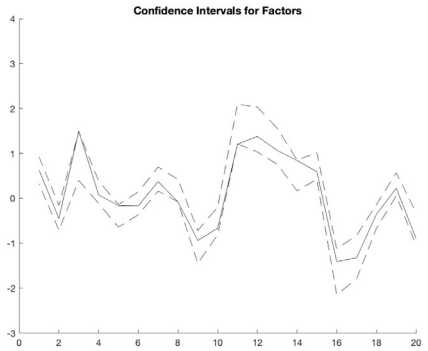
Logit, $N = 50, T = 50$.



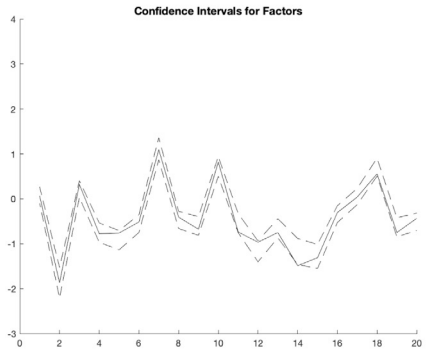
Probit, $N = 50, T = 50$.



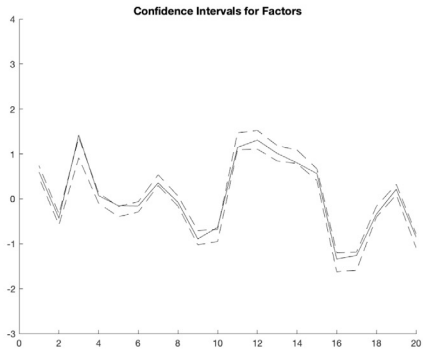
Logit, $N = 100, T = 100$.



Probit, $N = 100, T = 100$.



Logit, $N = 500, T = 200$.

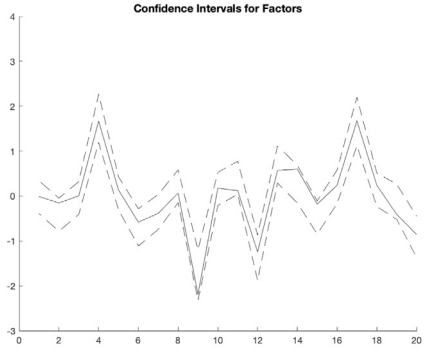


Probit, $N = 500, T = 200$.

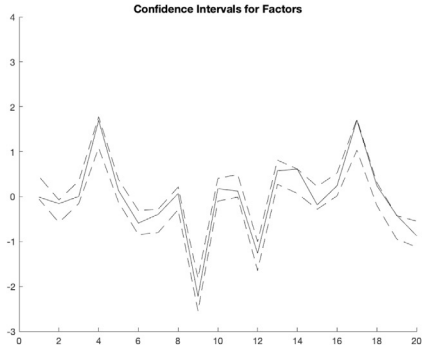
Fig. 5. Confidence intervals for factors (logit and probit). Notes: These are the 95% confidence intervals for the true factor process. The middle curve (solid line) is the true factor process.

Once $\{x_{it}; i = 1, \ldots, N, t = 1, \ldots, T\}$ is generated, we use the MM algorithm[11] to calculate the maximum likelihood estimators, $\{\hat{f}_t, t = 1, \ldots, T\}$ and $\{\hat{\lambda}_i, i = 1, \ldots, N\}$. For step 1, the initial values of the factors and loadings, $(\hat{f}_t^{(0)}, \hat{\lambda}_i^{(0)})$

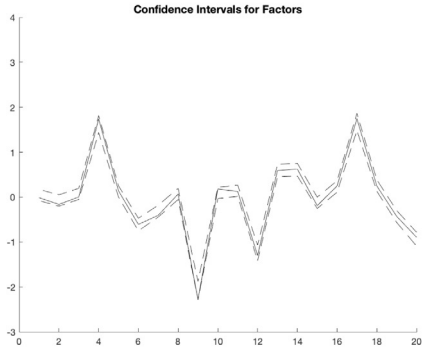
11 We choose the MM algorithm because it is computationally simpler than the AM algorithm.



Mixed, $N = 50, T = 50$.



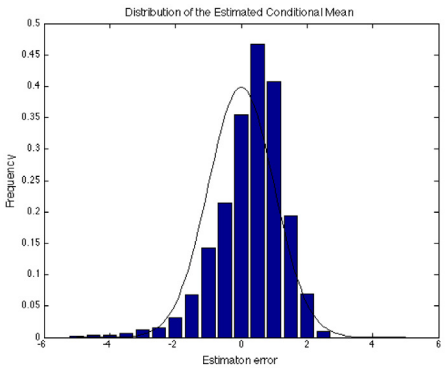
Mixed, $N = 100, T = 100$.



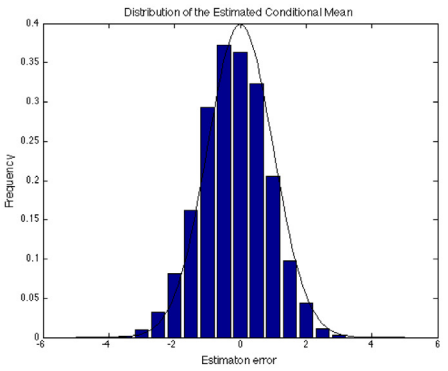
Mixed, $N = 500, T = 200$.

Fig. 6. Confidence intervals for factors (mixed). Notes: These are the 95% confidence intervals for the true factor process. The middle curve (solid line) is the true factor process.

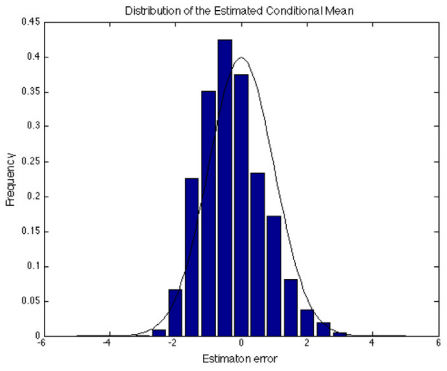
are randomly generated from standard normal distribution for DGP1 and *Uniform*$(-2, 2)$ for DGP2 and DGP3. For step 2, we choose $b_U = \frac{1}{4}$ for DGP1 and $b_U = 1$ for DGP2 and DGP3. This is because $-\partial_{\pi^2} l_{it}(\cdot)$ is bounded by $\frac{1}{4}$ for the Logit case, by 1 for the Probit case and equals 1 for the Gaussian case. For step 3, the maximum number of iteration is 20. In



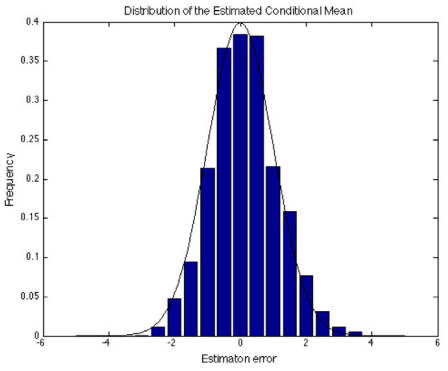
Logit, $N = 50, T = 50$.



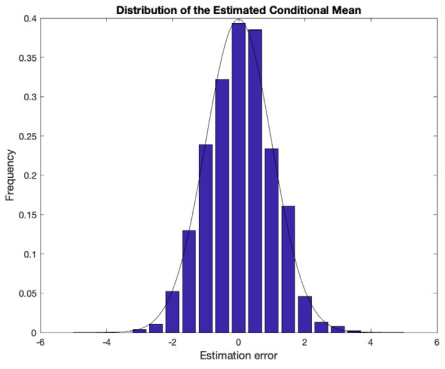
Probit, $N = 50, T = 50$.



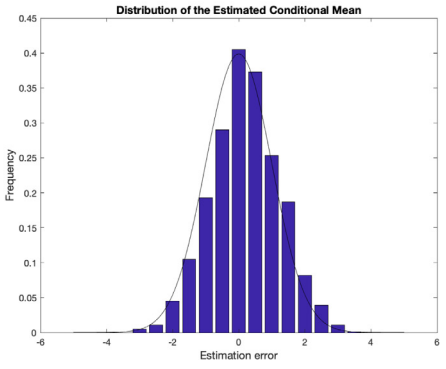
Logit, $N = 100, T = 100$.



Probit, $N = 100, T = 100$.



Logit, $N = 500, T = 200$

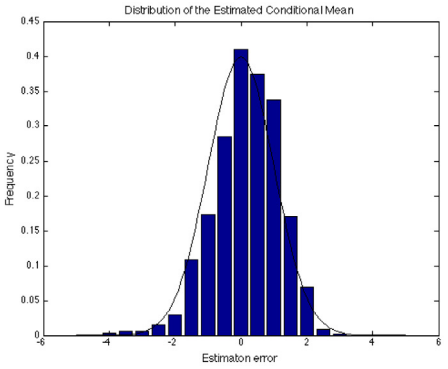


Probit, $N = 500, T = 200$

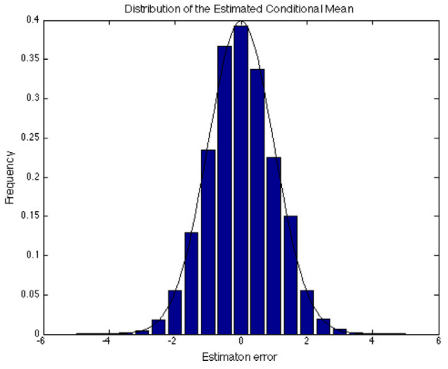
Fig. 7. Distribution of the estimated conditional mean (logit and probit). Notes: These histograms are for the standardized estimated conditional mean. The curve overlaid on the histograms is the standard normal density function.

simulations, we find the convergence speed is very fast at the beginning. The difference between the fourth iteration and the twentieth iteration is not large. The number of simulations is 2000.

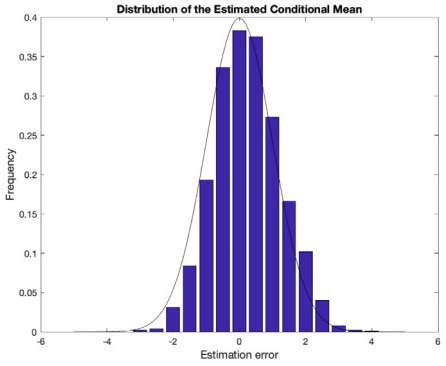
Due to limited space, we only present results for $(N, T) = (50, 50)$, $(100, 100)$ and $(500, 200)$. According to Theorem 2, $N^{\frac{1}{2}}[\bar{G}^{-1}\Sigma_{t\Lambda}\bar{G}'^{-1}]^{\frac{1}{2}}(\hat{f}_t - f_t^G)$ follows standard normal distribution for each t and so does $T^{\frac{1}{2}}[\bar{G}'\Sigma_{iF}\bar{G}]^{\frac{1}{2}}(\hat{\lambda}_i - \lambda_i^G)$ for each i.



Mixed, $N = 50, T = 50$.



Mixed, $N = 100, T = 100$.



Mixed, $N = 500, T = 200$

Fig. 8. Distribution of the estimated conditional mean (mixed). Notes: These histograms are for the standardized estimated conditional mean. The curve overlaid on the histograms is the standard normal density function.

Figs. 1–2 display the histograms of $N^{\frac{1}{2}}[\bar{G}^{-1}\Sigma_{T/2,\Lambda}\bar{G}'^{-1}]^{\frac{1}{2}}(\hat{f}_{T/2} - f^G_{T/2})$ for the three DGPs. Fig. 3 displays the histograms of $T^{\frac{1}{2}}[\bar{G}'\Sigma_{N/2,F}\bar{G}]^{\frac{1}{2}}(\hat{\lambda}_{N/2} - \lambda^G_{N/2})$ for DGP1 and DGP2. Fig. 4 displays the histograms of $T^{\frac{1}{2}}[\bar{G}'\Sigma_{0.9N,F}\bar{G}]^{\frac{1}{2}}(\hat{\lambda}_{0.9N} - \lambda^G_{0.9N})$ for DGP3, where $i = 0.9N$ corresponds to the Gaussian part of DGP3. The histograms are normalized to be a density function and

Table 1
Average correlation coefficients of the estimated factors.

N	T	Logit		Probit		Mixed	
		MLE	PC	MLE	PC	MLE	PC
50	50	0.9352	0.0174	0.9610	0.0444	0.9821	0.9571
100	100	0.9598	0.2307	0.9787	0.0939	0.9886	0.9672
500	200	0.9927	0.0848	0.9949	0.0226	0.9974	0.9773

Notes: These are the correlation coefficients between the true factors and the factors estimated by MLE or PC, averaged over 2000 simulations.

Table 2
Coverage rates of confidence intervals.

N	T	Logit		Probit		Mixed	
		$\hat{y}_{T+1\|T}$	$\hat{y}_{T+1}$	$\hat{y}_{T+1\|T}$	$\hat{y}_{T+1}$	$\hat{y}_{T+1\|T}$	$\hat{y}_{T+1}$
50	50	0.954	0.947	0.946	0.948	0.959	0.950
50	100	0.955	0.951	0.961	0.950	0.943	0.952
100	50	0.931	0.943	0.961	0.951	0.954	0.952
100	100	0.962	0.944	0.941	0.950	0.948	0.951
500	200	0.958	0.952	0.939	0.946	0.941	0.951

Notes: These are the coverage rates of 95% confidence intervals for the conditional mean and the one step ahead forecast.

the standard normal density curve is overlaid on them for comparison. It is easy to see that in all subfigures, the standard normal density curve provides good approximation to the normalized histograms. Note that for different subfigures, the variance of the unnormalized estimation error, i.e., $\hat{f}_t - f_t^G$ and $\hat{\lambda}_i - \lambda_i^G$, varies with N, T and DGP of x_{it}. But once normalized, the estimation errors always approximately follow the standard normal distribution.

To graphically illustrate the estimated factors, we also construct confidence intervals of the true factor process $\{f_t^G, t = 1, \ldots, 20\}$ in Figs. 5–6 for the three DGPs. Confidence intervals for $\{f_t^G, t = 21, \ldots, T\}$ are not presented to avoid too many points in one graph. The solid middle curves in Figs. 5–6 are the true factor processes. It is easy to see that the true factors rarely fall outside of the confidence intervals and the confidence intervals become narrower as N and T increase. For $(N, T) = (500, 200)$, the confidence intervals almost coincide with the true factors. These together lend strong support to the theoretical results.

To compare the factors estimated by MLE with the factors estimated by brutal PCA, we present in Table 1 the correlation coefficient between the estimated factor and the true factor averaged over 2000 simulations. It is easy to see that PCA performs poorly for Logit and Probit, and the performance does not improve as N and T increase. For the mixed DGP, due to the presence of some Gaussian time series, PCA performs much better but still significantly worse than MLE.

Now we consider the factor-augmented regression, $y_{t+1} = \alpha' f_t^0 + \beta' W_t + \epsilon_{t+1}$. We already have f_t^0 and $\hat{f}_t$. W_t is $i.i.d.\mathcal{N}(0, 1)$ and is fixed down once generated. $\{\epsilon_{t+1}, t = 1, \ldots, T\}$ is $i.i.d.\mathcal{N}(0, 1)$ and generated 2000 times. For the regression coefficients, we choose $\alpha = \beta = 1$. According to Theorem 4, $(\hat{y}_{T+1|T} - y_{T+1|T})/B_T$ should follow standard normal distribution. Figs. 7–8 display its histograms for the three DGPs. As Figs. 1–4, the standard normal density curve is overlaid on the normalized histograms. On the whole, standard normal distribution provides reasonable approximation, but here the approximation is not as good as Figs. 1–4. This is because $\hat{y}_{T+1|T} - y_{T+1|T}$ contains two sources of errors, one from $\hat{\theta} - \theta$ and the other from $\hat{f}_T - f_T^G$. The slight skewness of the histograms for the Logit case in Fig. 7 disappears if we increase (N, T) to $(500, 200)$.

Theorem 4 also allows constructing confidence intervals for the conditional mean $y_{T+1|T}$ and the one step ahead forecast. The 95% confidence interval is $(\hat{y}_{T+1|T} - 1.96B_T, \hat{y}_{T+1|T} + 1.96B_T)$ for $y_{T+1|T}$ and $(\hat{y}_{T+1|T} - 1.96\sqrt{B_T^2 + \sigma_\epsilon^2}, \hat{y}_{T+1|T} + 1.96\sqrt{B_T^2 + \sigma_\epsilon^2})$ for the one step ahead forecast. Table 2 reports the coverage rates for the three DGPs. In all cases, the coverage rate is close to the nominal level 95%. To compare the accuracy of forecasts using factors estimated by MLE with factors estimated by brutal PCA, we present in Table 3 the mean squared errors of forecasts using these two methods. It is easy to see that forecasts using MLE factors performs significantly better in all cases. This demonstrates the advantage of using MLE when the DGP is nonlinear or mixed.

8. Conclusions

This paper studies maximum likelihood estimation of factor models with high dimensional nonlinear/mixed data. Convergence rates of the estimated factor space and loading space and asymptotic normality of the estimated factors and loadings are established under mild conditions that allows for linear, Logit, Probit, Tobit, Poisson and some other single-index nonlinear models. This paper also establishes the limit distributions of the parameter estimates, the conditional mean as well as the forecast when these estimated factors are used as proxies in factor-augmented regressions. These

Table 3
Mean squared error of forecasts using MLE factors or PC factors.

N	T	Logit		Probit		Mixed	
		MLE	PC	MLE	PC	MLE	PC
50	50	0.0637	0.3019	0.1327	2.1925	0.0500	0.0833
100	100	0.0969	1.1572	0.0640	1.5993	0.0802	0.1799
500	200	0.0098	0.0637	0.0093	0.0396	0.0079	0.0263

Notes: These are the squared errors of forecasts using factors estimated by MLE or PCA, averaged over 2000 simulations.

results provide a rigorous treatment of high dimensional nonlinear/mixed data in factor analysis and factor-augmented regressions. Given the prevalence of nonlinear/mixed data, empirical applications of the results developed in this paper should be fairly fruitful, especially to the topics discussed in the Introduction. For example, it would be interesting to apply this paper's method to real credit default data. We hope this paper would trigger further developments in the analysis of high dimensional nonlinear data.

Appendix. Mathematical proofs and technical details

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.11.002.

References

Bai, 2003. Inferential theory for factor models of large dimensions. Econometrica 71, 135–171.
Bai, Li, 2012. Statistical analysis of factor models of high dimension. Ann. Statist. 43, 6–465.
Bai, Li, 2016. Maximum likelihood estimation and inference for approximate factor models of high dimension. Rev. Econ. Stat. 98, 298–309.
Bai, Ng, 2002. Determining the number of factors in approximate factor models. Econometrica 70, 191–221.
Bai, Ng, 2006. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. Econometrica 74, 1133–1150.
Bartholomew, 1980. Factor analysis for categorical data. J. R. Stat. Soc. Ser. B Stat. Methodol. 29, 3–321.
Bartholomew, Knott, 1999. Latent Variable Models and Factor Analysis. Edward Arnold.
Bernanke, Boivin, Eliasz, 2005. Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. Q. J. Econ. 120, 387–422.
Bohning, Lindsay, 1988. Monotonicity of quadratic-approximation algorithms. Ann. Inst. Statist. Math. 40, 641–663.
Campbell, Lo, Mackinlay, 1997. The Econometrics of Financial Markets. Princeton University Press, New Jersey.
Chen, 2016. Estimation of Nonlinear Panel Models with Multiple Unobserved Effects. In: Warwick Economics Research Paper Series.
Chen, Fernandez-Val, Weidner, 2014. Nonlinear panel models with interactive effects. arXiv preprint arXiv:1412.5647.
Chen, Fernandez-Val, Weidner, 2020. Nonlinear factor models for network and panel data. J. Econometrics forthcoming.
Collins, Dasgupta, Schapire, 2001. A generalization of principal component analysis to the exponential family. Adv. Neural Inform. Proces. Syst. (13).
Cox, Reid, 1987. Parameter orthogonality and approximate conditional inference. J. R. Stat. Soc. Ser. B Stat. Methodol. 1–39.
Creal, Schwaab, Koopman, Lucas, 2014. Observation-driven mixed-measurement dynamic factor models with an application to credit risk. Rev. Econ. Stat. 96, 898–915.
de Leeuw, 2006. Principal component analysis of binary data by iterated singular value decomposition. Comput. Statist. Data Anal. 50, 21–39.
Fan, Gong, Zhu, 2019. Generalized high-dimensional trace regression via nuclear norm regularization. J. Econometrics 212, 177–202.
Fan, Xue, Yao, 2017. Sufficient forecasting using factor models. J. Econometrics 201, 292–306.
Feng, Wang, Han, Xia, Tu, 2013. The mean value theorem and taylor's expansion in statistics. Amer. Statist. 67, 245–248.
Fernandez-Val, Weidner, 2016. Individual and time effects in nonlinear panel models with large N, T. J. Econometrics 192, 291–312.
Filmer, Pritchett, 2001. Estimating wealth effects without expenditure data—or tears: An application to educational enrollments in states of India. Demography 38, 115–132.
Hahn, Kuersteiner, 2011. Bias reduction for dynamic nonlinear panel models with fixed effects. Econometric Theory 27, 1152–1191.
Hahn, Newey, 2004. Jackknife and analytical bias reduction for nonlinear panel models. Econometrica 72, 1295–1319.
Hunter, Lange, 2004. A tutorial on MM algorithms. Amer. Statist. 58, 30–37.
Jennrich, 1969. Asymptotic properties of non-linear least squares estimators. Ann. Math. Stat. 40, 633–643.
Joreskog, Moustaki, 2001. Factor analysis of ordinal variables: A comparison of three approaches. Multivariate Behav. Res. 36, 347–387.
Koopman, Lucas, 2008. A non-Gaussian panel time series model for estimating and decomposing default risk. J. Bus. Econom. Statist. 26, 510–525.
Koopman, Lucas, Monteiro, 2008. The multi-state latent factor intensity model for credit rating transitions. J. Econometrics 142, 399–424.
Koopman, Lucas, Schwaab, 2011. Modeling frailty-correlated defaults using many macroeconomic covariates. J. Econometrics 162, 312–325.
Lancaster, 2000. The incidental parameter problem since 1948. J. Econometrics 95, 391–413.
Lancaster, 2002. Orthogonal parameters and panel data. Rev. Econom. Stud. 69, 647–666.
Lange, Hunter, Young, 2000. Optimization transfer using surrogate objective functions. J. Comput. Graph. Statist. 9, 1–20.
McNeil, Wendin, 2007. Bayesian inference for generalized linear mixed models of portfolio credit risk. J. Empir. Financ. 14, 131–149.
Moustaki, 1996. A latent trait and a latent class model for mixed observed variables. Br. J. Math. Stat. Psychol. 49, 313–334.
Moustaki, 2000. A latent variable model for ordinal variables. Appl. Psychol. Meas. 24, 211–223.
Moustaki, Knott, 2000. Generalized latent trait models. Psychometrika 65, 391–411.
Newey, McFadden, 1994. Large sample estimation and hypothesis testing. In: Handbook of Econometrics, IV. pp. 2111–2245.
Ng, 2015. Constructing common factors from continuous and categorical data. Econometric Rev. 34, 1141–1171.
Ross, 1976. The arbitrage theory of capital asset pricing. J. Finance 13, 341–360.
Schein, Saul, Ungar, 2003. A generalized linear model for principal component analysis of binary data, In: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics.
Schonbucher, 2000. Factor models for portfolio credit risk. In: Bonn Econ Discussion Papers 16.
Stock, Watson, 2002. Forecasting using principal components from a large number of predictors. J. Amer. Statist. Assoc. 97, 1167–1179.
Stock, Watson, 2016. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In: Handbook of Macroeconomics, Vol. 2. pp. 415–525.
Wu, 1981. Asymptotic theory of nonlinear least squares estimation. Ann. Statist. 50, 1–513.