CENTRE FOR ECONOMETRIC ANALYSIS
CEA@Cass

http://www.cass.city.ac.uk/cea/index.html

Cass Business School
Faculty of Finance
106 Bunhill Row
London EC1Y 8TZ

Maximum Likelihood Estimation and Inference for High Dimensional Nonlinear Factor Models

Fa Wang

CEA@Cass Working Paper Series

WP–CEA–09-2017

Maximum Likelihood Estimation and Inference for High Dimensional Nonlinear Factor Models

Fa Wang*

May 14, 2017

Abstract

This paper considers maximum likelihood for estimating the factors and the loadings from high dimensional categorical/mixed-measurement data. Both factors and loadings are treated as parameters to be estimated. Convergence rate of the estimated factor space, consistency, convergence rate and asymptotic normality of the estimated factors and the estimated loadings are established under mild conditions that allow for linear models, Logit, Probit and some other nonlinear models. The density function is also allowed to vary across subjects, thus mixed-measurement models are explicitly allowed for. 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. All these results are derived under the asymptotic framework where the cross-section dimension and the time dimension tend to infinity jointly at the same rate.

Keywords: Factor model, Discrete data, Maximum likelihood, High dimension, Factor-augmented regression, Forecasting

JEL Classification: C13, C35

*School of Economics, Shanghai University of Finance and Economics, Shanghai, China

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 least squares estimator of the 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), Forni and Reichlin (1998), Bernanke, Boivin and Eliasz (2005), Ross (1976) and Campbell, Lo and Mackinlay (1997), to name a few.

So far the literature only considers linear factor models. However, in some macroeconomic or financial applications and in most microeconomic applications, the relationship between the dependent variable and the factors could be nonlinear. A representative case is when the dependent variable is categorical. Extending the existing theory, e.g., Bai (2003) and Bai and Li (2012), 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 nonlinear factor models. More specifically, this paper considers the following single-index factor model: For $i = 1, ..., N$ and $t = 1, ..., T$,

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

x_{it} is the observed data for the i-th subject at time t. $g_i(\cdot\,|\cdot)$ is some known density function of x_{it} allowed to vary across i. $\pi_{it}^0 = f_t^{0\prime}\lambda_i^0$ and f_t^0 and λ_i^0 is a r dimensional vector of factors and loadings respectively. Both N and T are large. The number of factors r is known. Neither factors nor loadings are observable. When factors or loadings are random, expression (1) is the conditional density. For simplicity, this paper considers nonrandom factors and loadings.

For engineering, this model has been successfully used in data compression, visu-

alization, 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. Little attention has been paid to the treatment of categorical data or mixed-measurement data even though many data sets are of this type. For example, consumer confidence index is categorical, credit rating is categorical and loss given default is bounded between zero and one. While these data sets 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 modelling has direct implications for CDO (collateralized debt obligations) pricing, bond portfolio management and commercial bank risk management. Factor models provide a parsimonious way for analyzing default correlation and underlies many risk models used in practice. In this case, $\pi_{it}^0 + e_{it}$ is the value of company i at time t, e_{it} a idiosyncratic term, $\pi_{it}^0 = f_t^{0\prime}\lambda_i^0$ and f_t^0 is vector of common factors. 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, readers are referred to Schonbucher (2000), Creal, Schwaab, Koopman and Lucas (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. Filmer and

Pritchett (2001) follows 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. This procedure is lack 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 categorical/mixed-measurement 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 the loadings from categorical/mixed-measurement data. Both factors and loadings are treated as parameters to be estimated and a penalty[1] 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 rate of the estimated factor space, consistency, convergence rate and asymptotic normality of the estimated factors and the estimated loadings as N and T tend to infinity at the same rate, given that the log-likelihood function satisfies some regularity conditions. These regularity conditions are mild enough to allow for linear models, Logit, Probit and some other nonlinear models. The density function is allowed to vary across i, thus a mixture of these models is also 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 model when these estimated factors are used as proxies for the true factors.

In the statistics literature, classic factor analysis has been successfully extended to categorical data and mixed-measurement 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

[1] In this paper, the penalty is considered solely for the purpose of fixing down the rotation matrix of the factor model. It has nothing to do with the LASSO literature.

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-measurement data recently has motivated some researchers to explore possible solution. Creal, Schwaab, Koopman and Lucas (2014) chooses a observation driven[2] framework to model the unobservable factors and then uses maximum likelihood to estimate the parameters. 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 and estimated simultaneously by maximum likelihood, this paper faces the incidental parameter problem as discussed in Neyman and Scott (1948), Heckman (1981), Lancaster (2000) and Greene (2004). This paper solves this problem by utilizing the fact that the Hessian of factor model is "asymptotically diagonal[3]" as N and T tend to infinity jointly. This 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 specifically, for factor models, the diagonal blocks of the Hessian are of magnitude $O_p(T)$ or $O_p(N)$ while the off-diagonal blocks are of magnitude $O_p(1)$. This paper shows these nonzero off-diagonal blocks are asymptotically negligible as N and T tend to infinity jointly. Asymptotic diagonality of the Hessian also provides explanation for Bai (2003)'s results from the perspective of extremum estimation, and roughly speaking, for the results in Hahn and Newey (2004) and many other nonlinear panel papers.

The rest of the paper is organized as follows. Section 2 introduces notations and

[2]In observation driven models, the parameters are allowed to vary over time as functions of lagged dependent variables and exogenous variables, i.e., they are perfectly predictable given the past information. For more details, see Creal, Koopman and Lucas (2013).

[3]More accurately, it should be block-diagonal. We use "diagonal" here because as N and T tend to infinity, the dimension of each block always equals to the number of factors while the number of blocks tends to infinity.

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 function is

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

where $l_{it}(\pi_{it}) = \log g_i(x_{it}\,|\pi_{it})$ 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 t-th row and the i-th column, $f = (f_1', ..., f_T')'$ a Tr dimensional vector and $\lambda = (\lambda_1', ..., \lambda_N')'$ is a Nr dimensional vector. The functional form of the density, $g_i(\cdot\,|\cdot)$, is allowed to vary across i. Thus data following different models can be merged directly. For example, discretely distributed time series could be used together with continuously distributed time series to extract common factors. Let $\phi = (\lambda', f')'$, $F = (f_1, ..., f_T)'$, $\Lambda = (\lambda_1, ..., \lambda_N)'$. Similarly, for the true values of the factors and the loadings, let $f^0 = (f_1^{0\prime}, ..., f_T^{0\prime})'$, $\lambda^0 = (\lambda_1^{0\prime}, ..., \lambda_N^{0\prime})'$, $\phi^0 = (\lambda^{0\prime}, f^{0\prime})'$, $F^0 = (f_1^0, ..., f_T^0)'$ and $\Lambda^0 = (\lambda_1^0, ..., \lambda_N^0)'$. Also, let $\partial_\pi l_{it}(\pi_{it})$ and $\partial_{\pi^2} l_{it}(\pi_{it})$ respectively be the first order and the second order derivative of $l_{it}(\cdot)$ evaluated at π_{it}.

Both factors and loadings are treated as parameters to be estimated through maximum likelihood. Note that for any F, Λ and $r \times r$ invertible matrix G, FG and $\Lambda(G')^{-1}$ has the same likelihood as F and Λ. To uniquely fix F and Λ, we add the following penalty function to the log-likelihood[4]:

$$\begin{aligned} P(f,\lambda) \;=\; & -\frac{c}{8}\sum\nolimits_{p=1}^{r}\Big(\sum\nolimits_{i=1}^{N}\lambda_{ip}^2 - \sum\nolimits_{t=1}^{T} f_{tp}^2\Big)^2 \\ & -\frac{c}{2}\sum\nolimits_{p=1}^{r}\sum\nolimits_{q=p+1}^{r}\Big(\sum\nolimits_{i=1}^{N}\lambda_{ip}\lambda_{iq}\Big)^2 \\ & -\frac{c}{2}\sum\nolimits_{p=1}^{r}\sum\nolimits_{q=p+1}^{r}\Big(\sum\nolimits_{t=1}^{T} f_{tp}f_{tq}\Big)^2, \end{aligned} \tag{3}$$

[4]This penalty function is inspired by and generalizes the penalty function in Chen, Fernandez-Val and Weidner (2014).

where $0 < c < b_L$ and b_L is lower bound of $-\partial_{\pi^2} l_{it}(\pi_{it})$ as presented in Assumption 2(ii) below. Thus the criterion function to be maximized is

$$Q(f, \lambda) = L(X \,|f, \lambda) + P(f, \lambda). \tag{4}$$

Let $\hat{f} = (\hat{f}_1', ..., \hat{f}_T')'$ and $\hat{\lambda} = (\hat{\lambda}_1', ..., \hat{\lambda}_N')'$ be the solution and let $\hat{\pi}_{it} = \hat{f}_t' \hat{\lambda}_i$, $\hat{F} = (\hat{f}_1, ..., \hat{f}_T)'$ and $\hat{\Lambda} = (\hat{\lambda}_1, ..., \hat{\lambda}_N)'$. From expression (3), it is not difficult to see that $\hat{F}'\hat{F}$ and $\hat{\Lambda}'\hat{\Lambda}$ must be diagonal and equal to each other. Next, let $S(\phi) = \partial_\phi Q(\phi)$, $S_\lambda(\phi) = \partial_\lambda Q(\phi)$ and $S_f(\phi) = \partial_f Q(\phi)$, it follows that $S(\phi) = (S_\lambda'(\phi), S_f'(\phi))'$. Let $H(\phi) = \partial_{\phi\phi'} Q(\phi)$ be the Hessian matrix. Decomposition of $H(\phi)$ and the expression of each component is presented in Appendix A.

Throughout the paper, $(N, T) \to \infty$ denotes N and T going to infinity jointly. $\xrightarrow{d}$ denotes convergence in distribution. "w.p.a." denotes "with probability approaching". If the argument of a function is suppressed, this means the true values of the parameters are plugged in. For example, $S = S(\phi^0)$ and $H = H(\phi^0)$. For matrix A, let $\rho_{\min}(A)$ denote its smallest eigenvalue and $\|A\|$, $\|A\|_F$, $\|A\|_1$ and $\|A\|_\infty$ denote its spectral norm, Frobenius norm, 1-norm and infinity norm respectively. When A has Nr rows, divide A into N blocks with each block containing r rows and let $[A]_{iq}$ denote the q-th row in the i-th block and $[A]_i = ([A]_{i1}', ..., [A]_{ir}')'$ denote the i-th block.

3 Assumptions

Assumption 1 *$T^{-1}F^{0\prime}F^0 \to \Sigma_F$ as $T \to \infty$ for some positive definite Σ_F and $N^{-1}\Lambda^{0\prime}\Lambda^0 \to \Sigma_\Lambda$ as $N \to \infty$ for some positive definite Σ_Λ.*

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

(ii)[5] *There exists $b_U > b_L > 0$ such that $b_L \leq -\partial_{\pi^2} l_{it}(\pi_{it}) \leq b_U$ and $|\partial_{\pi^3} l_{it}(\pi_{it})| \leq b_U$ a.s..*

(iii) $\mathbb{E}(\partial_\pi l_{it})^{64}$ is uniformly bounded.

[5] This part can be weakened such that it only holds when the independent variable of $l_{it}(\cdot)$ is bounded. This is enough if parameter spaces of factors and loadings are uniformly bounded.

Assumption 3 *Conditioning on the factors and loadings, x_{it} is independent over i and t.*

Assumption 4 *The eigenvalues of $\Sigma_F \Sigma_\Lambda$ are different.*

Assumption 5 *$N/T \to \kappa$ as $(N,T) \to \infty$.*

Assumption 6 *(i) The parameter space of λ_i and f_t is compact. Furthermore, there exists $M > 0$ such that for any i, $\|\lambda_i\| \leq M$ for any possible λ_i and for any t, $\|f_t\| \leq M$ for any possible f_t.*

(ii) $T^{-1}\sum_{t=1}^{T} l_{it}(f_t^{0\prime}\lambda_i) - T^{-1}\sum_{t=1}^{T} \mathbb{E}l_{it}(f_t^{0\prime}\lambda_i)$ is $o_p(1)$ uniformly over the space of λ_i and $N^{-1}\sum_{i=1}^{N} l_{it}(f_t'\lambda_i^0) - N^{-1}\sum_{i=1}^{N} \mathbb{E}l_{it}(f_t'\lambda_i^0)$ is $o_p(1)$ uniformly over the space of f_t.

(iii) $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} \mathbb{E}l_{it}(f_t^{0\prime}\lambda_i)$ is continuous for λ_i and $\lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} \mathbb{E}l_{it}(f_t'\lambda_i^0)$ is continuous for f_t.

(iv) λ_i^0 is the unique maximizer of $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} \mathbb{E}l_{it}(f_t^{0\prime}\lambda_i)$ and f_t^0 is the unique maximizer of $\lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} \mathbb{E}l_{it}(f_t'\lambda_i^0)$.

Assumption 7 *For each i, as $T \to \infty$,*

$$T^{-\frac{1}{2}} \sum_{t=1}^{T} \partial_\pi l_{it} f_t^0 \xrightarrow{d} \mathcal{N}(0, \Sigma_{iF}),$$

where $\Sigma_{iF} = \lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} \mathbb{E}(-\partial_{\pi^2} l_{it}) f_t^0 f_t^{0\prime}$ and for each t, as $N \to \infty$,

$$N^{-\frac{1}{2}} \sum_{i=1}^{N} \partial_\pi l_{it} \lambda_i^0 \xrightarrow{d} \mathcal{N}(0, \Sigma_{t\Lambda}),$$

where $\Sigma_{t\Lambda} = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N} \mathbb{E}(-\partial_{\pi^2} l_{it}) \lambda_i^0 \lambda_i^{0\prime}$.

Assumption 1 ensures that each factor has a nontrivial contribution. For simplicity, this paper only considers nonrandom factors and loadings. If factors and loadings are random, then expression (2) is the conditional likelihood and the results are still valid. Assumption 2(i) imposes smoothness condition on the log-likelihood function. Assumption 2(ii) assumes the log-likelihood function is concave and its third order derivative is uniformly bounded. The existence and boundedness of the third order

derivative is needed to control the remainder term in the expansion of the first order condition[6]. Assumption 2(iii) is used in calculating $\|H^{-1}\|_1$. It can be verified that most frequently-used nonlinear models such as logit, probit, ordered logit and ordered probit satisfy Assumption 2. Assumption 3 assumes independence but not identical distribution of x_{it} over i and t. The independence condition is not uncommon for nonlinear models. With the cost of an increase of the technical complexity, cross-sectional dependence and time dependence could be allowed and the results are conceptually the same, provided the dependence is weak. Assumption 4 is a crucial identification condition. First, it guarantees that there exists unique F and Λ such that $F\Lambda' = F^0\Lambda^{0\prime}$, $\Lambda'\Lambda$ equals $F'F$ and both are diagonal. Second, it guarantees that there exists $m > 0$ such that w.p.a. one, $H(\phi)$ is nonsingular for all $\left\|\phi - \phi^0\right\| \leq mT^{\frac{1}{2}}$, i.e., w.p.a. one $Q(\phi)$ is locally concave for $\left\|\phi - \phi^0\right\| \leq mT^{\frac{1}{2}}$. Assumption 5 is imposed for simplicity. Bai (2003) assumes $T^{\frac{1}{2}}/N \to 0$ and $N^{\frac{1}{2}}/T \to 0$. For all results of this paper, Assumption 5 can be relaxed, but so far how to relax to Bai (2003)'s condition is unknown. Assumption 6 ensures the consistency of the estimated loadings when factors are known and the consistency of the estimated factors when loadings are known[7]. Assumption 7 is simply a result of the central limit theorem and is implied by Assumption 3.

4 Limit Theory

For any F^0 and Λ^0, let $\mathcal{V}$ be the diagonal matrix of eigenvalues of $(\Lambda^{0\prime}\Lambda^0)^{\frac{1}{2}}F^{0\prime}F^0(\Lambda^{0\prime}\Lambda^0)^{\frac{1}{2}}$ and Υ be the corresponding matrix of eigenvectors and let $G = (\Lambda^{0\prime}\Lambda^0)^{\frac{1}{2}}\Upsilon\mathcal{V}^{-\frac{1}{4}}$. It can be verified that $(F^0G)'F^0G = \mathcal{V}^{\frac{1}{2}}$ and $[\Lambda^0(G^{-1})']'[\Lambda^0(G^{-1})'] = \mathcal{V}^{\frac{1}{2}}$. In other words, for any F^0 and Λ^0, there always exists F and Λ such that $F\Lambda' = F^0\Lambda^{0\prime}$, $\Lambda'\Lambda$ equals $F'F$ and both are diagonal. Assumption 4 guarantees such F and Λ is unique for N

[6]Newey and McFadden (1994) only requires 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 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.

[7]See, e.g., Newey and McFadden (1994) for how these conditions can be used to show the consistency.

and T large enough. Without loss of generality, in the following we simply assume

$$F^{0\prime}F^0 = \Lambda^{0\prime}\Lambda^0 \text{ and both are diagonal.} \tag{5}$$

If (5) does not hold, just replace F^0 by F^0G and Λ^0 by $\Lambda^0(G^{-1})'$, then all results below still hold.

Proposition 1 (Average Consistency) *Under Assumptions 1-5, as* $(N,T) \to \infty$, $\left\|\hat{f} - f^0\right\| = O_p(T^{\frac{1}{8}})$, $\left\|\hat{\lambda} - \lambda^0\right\| = O_p(T^{\frac{1}{8}})$ *and w.p.a. one both* $\hat{f}$ *and* $\hat{\lambda}$ *are unique.*

Since w.p.a. one the criterion function is locally concave within a neighborhood of the true parameters[8], average consistency of $\hat{f}$ and $\hat{\lambda}$ implies their uniqueness w.p.a. one. The procedure for establishing consistency here is different from the classical procedure for extremum estimators, e.g., Newey and McFadden (1994) because the number of parameters tends to infinity jointly with N and T, which is also the main difficulty.

Note that although $\hat{f}$ and $\hat{\lambda}$ are estimated simultaneously, $\hat{\lambda}_i$ can be regarded as the maximum likelihood estimator when $\hat{f}$ is used for f^0 because $\hat{\lambda}_i$ maximizes $\sum_{t=1}^{T} l_{it}(\hat{f}_t'\lambda_i)$, and vice versa. Thus to establish the limit distribution of the estimated loadings and the estimated factors, it only remains to study the effect of using $\hat{f}$ for f^0 and $\hat{\lambda}$ for λ^0 respectively. Such effect is well-studied in linear setup, e.g., Bai (2003) and Bai and Ng (2006). But in nonlinear setup, their techniques are no longer useful. Detailed explanation and a new analytical framework to solve this problem will be provided after Proposition 3. We first strengthen Proposition 1 to get the average convergence rate.

Theorem 1 (Average Convergence Rate) *Under Assumptions 1-5 and 6(i), as* $(N,T) \to \infty$, $\left\|\hat{f} - f^0\right\| = O_p(1)$ *and* $\left\|\hat{\lambda} - \lambda^0\right\| = O_p(1)$.

Theorem 1 establishes the convergence rate of the estimated factor space and the estimated loading space. Under the extra condition $N/T \to \kappa$, this result is the same

[8] See Lemma 4 in the Appendix for more details.

as Theorem 1 in Bai and Ng (2002). As in the linear setup, this convergence rate is crucial for analyzing the effect of using estimated factors[9].

Proposition 2 (Individual Consistency) *Under Assumptions 1-6, as* $(N,T) \to \infty$, $\left\|\hat{\lambda}_i - \lambda_i^0\right\| = o_p(1)$ *for each* i *and* $\left\|\hat{f}_t - f_t^0\right\| = o_p(1)$ *for each* t.

Proposition 2 is an intermediate step for establishing Proposition 3 below. Its proof is based on Theorem 1. For example, simply consider the estimated loadings. Given Assumption 6, the estimated loadings are consistent if factors were known. Then Theorem 1 is utilized to show $T^{-1}\sum_{t=1}^{T}[l_{it}(\hat{f}_t'\lambda_i) - l_{it}(f_t^{0\prime}\lambda_i)] = o_p(1)$ uniformly over the space of λ_i.

Proposition 3 (Individual Convergence Rate) *Under Assumptions 1-6, as* $(N,T) \to \infty$, $\left\|\hat{\lambda}_i - \lambda_i^0\right\| = O_p(T^{-\frac{1}{2}})$ *for each* i *and* $\left\|\hat{f}_t - f_t^0\right\| = O_p(N^{-\frac{1}{2}})$ *for each* t.

Proposition 3 is a crucial step for establishing the limit distribution of the estimated factors and loadings. Since $\hat{\lambda}_i$ maximizes the likelihood of the i-th series when $\hat{f}$ is used as the data for f^0, a natural choice for deriving its limit distribution 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^0 and then analyze the effect of using $\hat{f}$ for f^0. This is not feasible. For $q = 1, ..., r$, expand the q-th row of the first order condition at λ_i^0,

$$0 = \sum_{t=1}^{T}\partial_\pi l_{it}(\hat{f}_t'\lambda_i^0)\hat{f}_{tq} + \sum_{t=1}^{T}\partial_{\pi^2} l_{it}(\hat{f}_t'\tilde{\lambda}_{i,iq})\hat{f}_{tq}\hat{f}_t'(\hat{\lambda}_i - \lambda_i^0),$$

where $\tilde{\lambda}_{i,iq} = a_{iq}\lambda_i^0 + (1-a_{iq})\hat{\lambda}_i$ for some $a_{iq} \in (0,1)$. The first term on the right hand side equals

$$\begin{aligned}&\sum_{t=1}^{T}(\partial_\pi l_{it})f_{tq}^0 + \sum_{t=1}^{T}[\partial_\pi l_{it}(\hat{f}_t'\lambda_i^0) - \partial_\pi l_{it}]f_{tq}^0 \\ &+ \sum_{t=1}^{T}(\partial_\pi l_{it})(\hat{f}_{tq} - f_{tq}^0) + \sum_{t=1}^{T}[\partial_\pi l_{it}(\hat{f}_t'\lambda_i^0) - \partial_\pi l_{it}](\hat{f}_{tq} - f_{tq}^0).\end{aligned}$$

By Assumptions 7, the first term is $O_p(T^{\frac{1}{2}})$ and normally distributed in the limit. If the remaining terms are $o_p(T^{\frac{1}{2}})$, then the first term would be dominant and we would

[9]But itself is not enough. In Bai (2003) and Bai and Ng (2006), Lemma B.1 and B.2 in Bai (2003) are also needed.

have normal distribution in the limit. This argument works in linear factor models. Without loss of generality, suppose $l_{it}(\pi_{it}) = -\frac{1}{2}(x_{it} - \pi_{it})^2$, then the second term becomes $-\sum_{t=1}^{T}(\hat{f}_t - f_t^0)'\lambda_i^0 f_{tq}^0$ and the third becomes[10] $\sum_{t=1}^{T} e_{it}(\hat{f}_{tq} - f_{tq}^0)$. Lemma B.2 and Lemma B.1 of Bai (2003) respectively shows both terms are $O_p(T/\min\{N,T\})$, which is $O_p(1)$ under Assumption 5. However, in nonlinear setup we are only able to show they are $O_p(T^{\frac{1}{2}})$ because it is not feasible to reestablish Lemma B.1 and Lemma B.2 of Bai (2003). The proof of these two lemmas are based on a crucial decomposition identity (equation A.1 in Bai (2003)) but that identity no longer exists in nonlinear setup.

To solve this problem, we expand all the first order conditions at ϕ^0.

$$0 = \partial_\phi Q(\hat{\phi}) = \partial_\phi Q + \partial_{\phi\phi'} Q \times (\hat{\phi} - \phi^0) + \frac{1}{2} R,$$

where $R = (R_\lambda', R_f')'$. R_λ and R_f is Nr and Tr dimensional with element $R_{\lambda,iq} = (\hat{\phi} - \phi^0)'\partial_{\phi\phi'\lambda_{iq}} Q(\phi_{iq}^*)(\hat{\phi} - \phi^0)$ and $R_{f,tq} = (\hat{\phi} - \phi^0)'\partial_{\phi\phi' f_{tq}} Q(\phi_{tq}^*)(\hat{\phi} - \phi^0)$ respectively. ϕ_{iq}^* and ϕ_{tq}^* are linear combinations of $\hat{\phi}$ and ϕ^0. Thus

$$\hat{\phi} - \phi^0 = -H^{-1}S - \frac{1}{2}H^{-1}R, \tag{6}$$

and it follows that

$$\hat{\lambda}_i - \lambda_i^0 = [\hat{\phi} - \phi^0]_i = -[H^{-1}S]_i - \frac{1}{2}[H^{-1}R]_i. \tag{7}$$

In the Appendix we show that

$$\left\| [H^{-1}R]_i \right\|_1 = O_p(T^{-\frac{29}{32}}), \tag{8}$$

$$[H^{-1}S]_i = \Big(\sum\nolimits_{t=1}^{T} \partial_{\pi^2} l_{it} f_t^0 f_t^{0\prime}\Big)^{-1} \sum\nolimits_{t=1}^{T} \partial_\pi l_{it} f_t^0 + O_p(T^{-\frac{7}{8}}). \tag{9}$$

Theorem 1 and Proposition 3 are needed to show (8). The limit distribution of $\hat{\lambda}_i$ follows directly from (8) and (9). The limit distribution of $\hat{f}_t$ follows from symmetry.

Theorem 2 (Individual Limit Distribution) *Under Assumptions 1-7, as* (N,T)

[10] Here e_{it} is the error term in Bai (2003).

$\to \infty$, $T^{\frac{1}{2}}(\hat{\lambda}_i - \lambda_i^0) \xrightarrow{d} \mathcal{N}(0, \Sigma_{iF}^{-1})$ *for each* i *and* $N^{\frac{1}{2}}(\hat{f}_t - f_t^0) \xrightarrow{d} \mathcal{N}(0, \Sigma_{t\Lambda}^{-1})$ *for each* t.

The limit distributions of the estimated factors and loadings for linear factor model are presented in Theorem 1 and Theorem 2 of Bai (2003). Theorem 2 includes Bai (2003)'s results as special case. If the density function is normal, Bai (2003)'s results can be obtained directly from Theorem 2.

Theorem 2 is rather useful as it not only allows discrete dependent variables but also allows the density function to differ across individuals. 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.

To understand Theorem 2, the main difficulty is that the dimension of the Hessian tends to infinity jointly as N and T. 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, the Hessian is not block-diagonal, but is asymptotically equivalently block-diagonal. More specifically, due to the factor structure, the Hessian can be decomposed into several parts. The first part is block-diagonal with $\sum_{t=1}^{T} \partial_{\pi^2} l_{it} f_t^0 f_t^{0\prime}$ as the i-th block and $\sum_{i=1}^{N} \partial_{\pi^2} l_{it} \lambda_i^0 \lambda_i^{0\prime}$ as the $(N+t)$-th block. For the remaining parts, each element is of magnitude $O_p(1)$. In the Appendix we show that in the expansion of $[H^{-1}S]_i$, the extra terms resulting from the remaining parts are of magnitude $O_p(T^{-\frac{7}{8}})$, and thus asymptotically negligible. The underlying mechanism of this asymptotic negligibility is: the score is smoothed by the off-diagonal elements of the Hessian when the dimension of the Hessian increases jointly with the sample size.

Remark 1 Σ_{iF} *can be estimated by* $\hat{\Sigma}_{iF} = T^{-1} \sum_{t=1}^{T} (\partial_{\pi} l_{it}(\hat{f}_t' \hat{\lambda}_i))^2 \hat{f}_t \hat{f}_t'$ *and* $\Sigma_{t\Lambda}$ *can be estimated by* $\hat{\Sigma}_{t\Lambda} = N^{-1} \sum_{i=1}^{N} (\partial_{\pi} l_{it}(\hat{f}_t' \hat{\lambda}_i))^2 \hat{\lambda}_i \hat{\lambda}_i'$. *Using Assumption 2(ii), Assumption 3 and Theorem 1, consistency of* $\hat{\Sigma}_{iF}$ *and* $\hat{\Sigma}_{t\Lambda}$ *is not difficult to show.*

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:

$$y_{t+h} = \alpha' f_t^0 + \beta' W_t + \epsilon_{t+h}, \tag{10}$$

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. Let $\delta = (\alpha', \beta')'$ and $z_t = (f_t^{0\prime}, W_t')'$, then $y_{t+h} = \delta' z_t + \epsilon_{t+h}$. W_t and y_{t+h} are both observable. f_t^0 is unobservable, but a large number of predictors $x_{it}(i = 1, ..., N; t = 1, ..., T)$ are observable and can be used to estimate f_t^0. The probability density function of x_{it} is $g_i(\cdot \left| f_t^{0\prime} \lambda_i^0 \right.)$, as introduced in Section 1. $g_i(\cdot \left| \cdot \right.)$ satisfies the regularity conditions listed in Assumptions 2 and 6.

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 other i, this is the model considered in Creal et al. (2014) for the estimation, analysis and forecasting of credit risk.

The objective is to characterize the effect of using the estimated factors $\hat{f}_t$ for f_t^0 on the limit distributions of the parameter estimates, the conditional mean as well as the forecast when the factors are estimated from x_{it} by maximum likelihood. Bai and Ng (2006) studies 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 be discretely distributed for all or some i.

Assumption 8 $\mathbb{E} \left\| z_t \right\|^4 \leq M$ *and* $\mathbb{E}(\epsilon_t^4) \leq M$ *for all* t. $\mathbb{E}(\epsilon_{t+h} \left| y_t, z_t, y_{t-1}, z_{t-1}, ... \right.) = 0$ *for all* $h > 0$. z_t *and* ϵ_t *are independent with* y_{is} *for all* i *and* s. *As* $T \to \infty$,

(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* $\Sigma_{zz\epsilon} = p\lim T^{-1}\sum_{t=1}^{T} \epsilon_{t+h}^2 z_t z_t'$.

Assumption 8 is the same as Assumption E in Bai and Ng (2006), except for an additional condition $\mathbb{E}(\epsilon_t^4) \leq M$. Thus readers are referred to Bai and Ng (2006) for discussion of this assumption. 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 let $\hat{\delta}$ be the least squares estimator of regressing y_{t+h} on $\hat{z}_t$.

Theorem 3 (Inference) *Under Assumptions 1-6(i) and 8, as* $(N,T) \to \infty$, $T^{\frac{1}{2}}(\hat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(0, \Sigma_\delta)$, *where* $\Sigma_\delta = \Sigma_{zz}^{-1}\Sigma_{zz\epsilon}\Sigma_{zz}^{-1}$. *A consistent estimator of* Σ_δ *is*

$$\hat{\Sigma}_\delta = (T^{-1}\sum_{t=1}^{T-h} \hat{z}_t\hat{z}_t')^{-1}(T^{-1}\sum_{t=1}^{T-h} \hat{\epsilon}_{t+h}^2\hat{z}_t\hat{z}_t')(T^{-1}\sum_{t=1}^{T-h} \hat{z}_t\hat{z}_t')^{-1}.$$

Theorem 3 implies that if $N/T \to \kappa > 0$, using the estimated factors does not affect the limit distribution of $\hat{\delta}$ when the factors are estimated by maximum likelihood and the density of x_{it} satisfy Assumptions 2 and 6. Theorem 3 is more general than Theorem 1 of Bai and Ng (2006) since the latter is established under the setup $x_{it} = f_t^{0\prime}\lambda_i^0 + e_{it}$. This generalization should be rather valuable as in many factor-augmented regressions the information about the common factors are contained in discrete or mixed-measurement data. Theorem 3 provides a rigorous way of exploiting these information.

Remark 2 *To eliminate the effect of using estimated factors, Theorem 3 requires* $N/T \to \kappa$ *while Bai and Ng (2006) only requires* $T^{\frac{1}{2}}/N \to 0$. *It is worth pointing out that* $N/T \to \kappa$ *is not a necessary condition for Theorem 3. It is assumed mainly for technical simplicity.*

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 $\delta' z_T$. Let $\hat{y}_{T+h|T} = \hat{\delta}'\hat{z}_T$ be the forecast of $y_{T+h|T}$.

Theorem 4 (Forecasting) *Under Assumptions 1-8, as* $(N,T) \to \infty$, $(\hat{y}_{T+h|T} - y_{T+h|T})/B_T \xrightarrow{d} \mathcal{N}(0,1)$, *where* $B_T^2 = T^{-1}z_T'\Sigma_{zz}^{-1}\Sigma_{zz\epsilon}\Sigma_{zz}^{-1}z_T + N^{-1}\alpha'\Sigma_{T\Lambda}^{-1}\alpha$. *A consistent estimator of* B_T^2 *is* $\hat{B}_T^2 = T^{-1}\hat{z}_T'\hat{\Sigma}_\delta\hat{z}_T + N^{-1}\hat{\alpha}'\hat{\Sigma}_{T\Lambda}^{-1}\hat{\alpha}$.

Theorem 4 generalizes Theorem 3 of Bai and Ng (2006) to allow factors to be extracted from discrete or mixed-measurement data. The variance of the estimated conditional mean has two components, one from the estimated parameters $\hat{\delta}$ and the other one from the estimated factors $\hat{f}$. 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 consistency 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.

6 Algorithms

We shall introduce two algorithms to calculate the maximum likelihood estimator numerically. Standard optimization algorithms are not applicable since the likelihood function $L(X\,|f, \lambda)$ is not concave with respect to (f, λ).

6.1 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 $k = 0, ...,$ *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* $i = 1, ..., N$ *and* $t = 1, ..., T$, *then* $(\hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)}) = \arg\max \sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{x}_{it}^{(k)} - f_t'\lambda_i)^2$. *Iterate until* $L(X\,\big|\,\hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)}) - L(X\,\big|\,\hat{f}^{(k)}, \hat{\lambda}^{(k)}) \leq error$, *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*

$\hat{F}^{(s)} = (\hat{f}_1^{(s)}, ..., \hat{f}_T^{(s)})'$ *and* $\hat{\Lambda}^{(s)} = (\hat{\lambda}_1^{(s)}, ..., \hat{\lambda}_N^{(s)})'$. *Let* $\hat{V}^{(s)}$ *be the diagonal matrix of eigenvalues of* $(\hat{\Lambda}^{(s)\prime}\hat{\Lambda}^{(s)})^{\frac{1}{2}}\hat{F}^{(s)\prime}\hat{F}^{(s)}(\hat{\Lambda}^{(s)\prime}\hat{\Lambda}^{(s)})^{\frac{1}{2}}$ *and* $\hat{\Upsilon}^{(s)}$ *be the corresponding matrix of eigenvectors and let* $\hat{G}^{(s)} = (\hat{\Lambda}^{(s)\prime}\hat{\Lambda}^{(s)})^{\frac{1}{2}}\hat{\Upsilon}^{(s)}(\hat{V}^{(s)})^{-\frac{1}{4}}$. *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 modified from de Leeuw (2006). Similar algorithm is also used by Chen (2016) for nonlinear panel models. Minorization maximization is a class of algorithm more general than the expectation maximization (EM). A function $h(x|y)$ is said to minorize a function $l(x)$ at y if $h(x|y) \leq l(x)$ for all x and $h(y|y) = l(y)$, i.e., $h(x|y)$ lies below $l(x)$ and is tangent to $l(x)$ at the point y. To maximize $l(x)$, the MM algorithm starts from an initial value $x^{(0)}$ and iteratively maximizes $h(x|x^{(k)})$ until convergence. By definition of $h(x|y)$, it is not difficult to see that $l(x^{(k)}) = h(x^{(k)}|x^{(k)}) \leq h(x^{(k+1)}|x^{(k)}) \leq l(x^{(k+1)})$. Thus convergence to local maximum is guaranteed. In applications, how to choose $h(x|y)$ mainly depends on computational simplicity. If there exists a function $w(y)$ such that $l(x) - l(y) \geq l'(y)(x-y) + \frac{1}{2}w(y)(x-y)^2$ for all x and y, a popular choice is $h(x|y) = l(y) + l'(y)(x-y) + \frac{1}{2}w(y)(x-y)^2$. For more details on the MM algorithm, see Bohning and Lindsay (1988), Hunter and Lange (2004) and Lange, Hunter and Young (2000), to name a few.

In current context, in view of the fact $\partial_{\pi^2} l_{it}(\pi_{it}) \geq -b_U$, we choose $h_{it}(x|y) = l_{it}(y) + l'_{it}(y)(x-y) - \frac{1}{2}b_U(x-y)^2$ for each (i,t). Let $\hat{\pi}_{it}^{(k)} = \hat{f}_t^{(k)\prime}\hat{\lambda}_i^{(k)}$, it follows that

$$\begin{aligned} l_{it}(\hat{\pi}_{it}^{(k+1)}) &\geq l_{it}(\hat{\pi}_{it}^{(k)}) + \partial_\pi l_{it}(\hat{\pi}_{it}^{(k)})(\hat{\pi}_{it}^{(k+1)} - \hat{\pi}_{it}^{(k)}) - \frac{1}{2}b_U(\hat{\pi}_{it}^{(k+1)} - \hat{\pi}_{it}^{(k)})^2 \\ &= l_{it}(\hat{\pi}_{it}^{(k)}) - \frac{1}{2}b_U(\hat{\pi}_{it}^{(k+1)} - \hat{\pi}_{it}^{(k)} - \frac{\partial_\pi l_{it}(\hat{\pi}_{it}^{(k)})}{b_U})^2 + \frac{(\partial_\pi l_{it}(\hat{\pi}_{it}^{(k)}))^2}{2b_U}. \end{aligned}$$

Take sum over i and t, then $L(X \left| \hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)} \right.) - L(X \left| \hat{f}^{(k)}, \hat{\lambda}^{(k)} \right.)$ is not smaller than

$$-\frac{1}{2}b_U \sum_{i=1}^N \sum_{t=1}^T (\hat{x}_{it}^{(k)} - \hat{\pi}_{it}^{(k+1)})^2 + \frac{1}{2b_U} \sum_{i=1}^N \sum_{t=1}^T (\partial_\pi l_{it}(\hat{\pi}_{it}^{(k)}))^2.$$

If $\hat{\pi}_{it}^{(k+1)} = \hat{\pi}_{it}^{(k)}$, this term is zero. Since $\hat{f}_t^{(k+1)}$ and $\hat{\lambda}_i^{(k+1)}$ minimizes $\sum_{i=1}^N \sum_{t=1}^T (\hat{x}_{it}^{(k)} - f_t'\lambda_i)^2$, this term must be nonnegative, and consequently $L(X \left| \hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)} \right.)$ is not

smaller than $L(X \left| \hat{f}^{(k)}, \hat{\lambda}^{(k)} \right.)$. This guarantees convergence of the MM algorithm in current context.

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 Alternating Maximization (AM)

Algorithm:

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

Step 2 (Iterate): For $k = 0, ...,$ *calculate*

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

Iterate until $L(X \left| \hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)} \right.) - L(X \left| \hat{f}^{(k)}, \hat{\lambda}^{(k)} \right.) \leq error$, *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 MM 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.*

This algorithm is not totally new. In machine learning literature, similar algorithm has been proposed in Collins, Dasgupta and Schapire (2001) and Schein, Saul and Ungar (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 $L(X \left| f, \lambda \right.)$ is globally concave with respect to λ for given f and vice versa. Because the likeli-

hood is maximized alternately, we have $L(X \left| \hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)} \right.) \geq L(X \left| \hat{f}^{(k+1)}, \hat{\lambda}^{(k)} \right.) \geq L(X \left| \hat{f}^{(k)}, \hat{\lambda}^{(k)} \right.)$. Thus convergence of step 2 to a local maximum is guaranteed. Step 3 and Step 4 are the same as the MM algorithm discussed above.

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 the case with one factor. For $i = 1, ..., N$ and $t = 1, ..., T$, f_t and λ_i are $i.i.d.\mathcal{N}(0, 1)$ and once generated, they are normalized to f_t^0 and λ_i^0 such that $\sum_{t=1}^{T}(f_t^0)^2 = \sum_{i=1}^{N}(\lambda_i^0)^2$. f_t^0 and λ_i^0 are fixed down for each simulation. For the given f_t^0 and λ_i^0, we consider three data generating processes (DGPs) for x_{it}.

DGP 1 (Logit): For $i = 1, ..., N$ and $t = 1, ..., T$, x_{it} is a binary random variable and $P(x_{it} = 1) = \Psi(f_t^0 \lambda_i^0)$, where $\Psi(z) = 1/(1 + e^{-z})$.

DGP 2 (Probit): For $i = 1, ..., N$ and $t = 1, ..., T$, x_{it} is a binary random variable and $P(x_{it} = 1) = \Phi(f_t^0 \lambda_i^0)$, where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution.

DGP 3 (Mixed): For $i = 1, ..., 2N/5$ and $t = 1, ..., T$, x_{it} is a binary random variable and $P(x_{it} = 1) = \Psi(f_t^0 \lambda_i^0)$; for $i = 2N/5 + 1, ..., 4N/5$ and $t = 1, ..., T$, x_{it} is binary random variable and $P(x_{it} = 1) = \Phi(f_t^0 \lambda_i^0)$; for $i = 4N/5 + 1, ..., N$ and $t = 1, ..., T$, x_{it} is normally distributed with mean $f_t^0 \lambda_i^0$ and variance 1.

Once $\{x_{it}; i = 1, ..., N, t = 1, ..., T\}$ is generated, we use the MM algorithm[11] to calculate the maximum likelihood estimators, $\{\hat{f}_t, t = 1, ..., T\}$ and $\{\hat{\lambda}_i, i = 1, ..., N\}$. For step 1, the initial values of the factors and loadings, $(\hat{f}_t^{(0)}, \hat{\lambda}_i^{(0)})$ are randomly generated from standard normal distribution for DGP1 and $Uniform(-2, 2)$ for DGP2 and DGP3[12]. 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

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

[12] We choose $U(-2, 2)$ for DGP2 and DGP3 partly because Matlab's default computational accuracy is limited.

Probit case and equals 1 for the Gaussian case. For step 3, the maximum number of iteration is 20. In 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.

According to Theorem 2, $N^{\frac{1}{2}}\Sigma_{t\Lambda}^{\frac{1}{2}}(\hat{f}_t - f_t^0)$ follows standard normal distribution for each t and so does $T^{\frac{1}{2}}\Sigma_{iF}^{\frac{1}{2}}(\hat{\lambda}_i - \lambda_i^0)$ for each i. Figure 1 displays the histograms of $N^{\frac{1}{2}}\Sigma_{T/2,\Lambda}^{\frac{1}{2}}(\hat{f}_{T/2} - f_{T/2}^0)$ for the three DGPs. Figure 2 displays the histograms of $T^{\frac{1}{2}}\Sigma_{N/2,F}^{\frac{1}{2}}(\hat{\lambda}_{N/2} - \lambda_{N/2}^0)$ for DGP1 and DGP2. For DGP3, Figure 3 displays the histograms of $T^{\frac{1}{2}}\Sigma_{i,F}^{\frac{1}{2}}(\hat{\lambda}_i - \lambda_i^0)$ for $i = N/5$, $3N/5$ and $9N/10$. Due to limited space, we only display histograms for $(N, T) = (50, 50)$ and $(100, 100)$. The histograms are normalized to be a density function and 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., $\hat{f}_t - f_t^0$ and $\hat{\lambda}_i - \lambda_i^0$, varies with N, T and DGP of x_{it}. But once normalized, the estimation errors always approximately follow the standard normal distribution. Also, the approximation is better as N and T increases from 50 to 100. These together lend strong support to the theoretical results.

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. $\{\epsilon_{t+1}, t = 1, ..., T\}$ 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. Figure 4 displays its histograms for the three DGPs. As Figures 1-3, the standard normal density curve is overlaid on the normalized histograms. On the whole, standard normal distribution provides reasonable approximation. The slight skewness of the histograms for the logit case disappears if we further increase N and T. 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 for $y_{T+1|T}$ is $(\hat{y}_{T+1|T} - 1.96B_T, \hat{y}_{T+1|T} + 1.96B_T)$ 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 1: Coverage Rates of Confidence Intervals

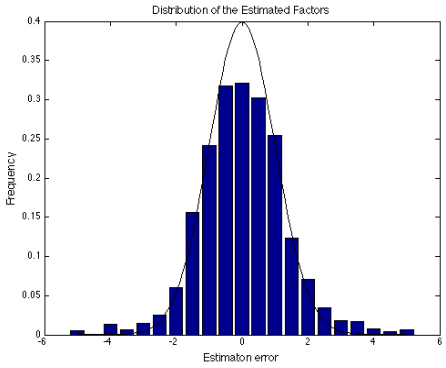
		Logit		Probit		Mixed	
N	T	$\hat{y}_{T+h\|T}$	$\hat{y}_{T+h}$	$\hat{y}_{T+h\|T}$	$\hat{y}_{T+h}$	$\hat{y}_{T+h\|T}$	$\hat{y}_{T+h}$
50	50	0.954	0.947	0.946	0.948	0.959	0.950
50	100	0.955	0.951	0.961	0.950	0.943	0.952
100	50	0.931	0.943	0.961	0.951	0.954	0.952
100	100	0.962	0.944	0.941	0.950	0.948	0.951

Table 1 reports the coverage rates for the three DGPs. In all cases, the coverage rate is close to the nominal level 95%.

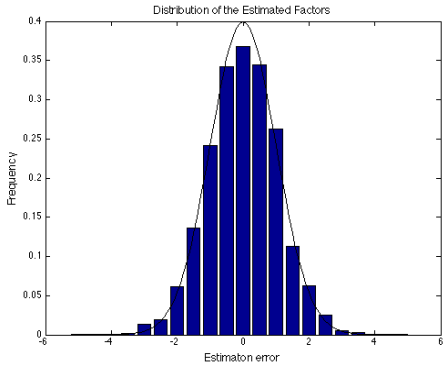
8 Conclusions

This paper studies maximum likelihood estimation of factor models with high dimensional categorical/mixed-measurement data. Convergence rate of the estimated factor space, consistency, convergence rate and asymptotic normality of the estimated factors and loadings are established under mild conditions that allows for linear models, Logit, Probit, some other nonlinear models and mixed-measurement 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 results provide a rigorous treatment of high dimensional categorical/mixed-measurement data in factor analysis and factor-augmented regressions. Given the prevalence of categorical/mixed-measurement data, empirical applications of the results developed in this paper should be fairly fruitful, especially to the topics discussed in the Introduction. The techniques developed in this paper may also provide a new perspective for solving the incidental parameter problem. For example, with these techniques asymptotic analysis for nonlinear panel models with multiple interactive fixed effects would no longer be difficult. This is currently under the author's investigation. We hope this paper would trigger further developments in the analysis of high dimensional discrete data.

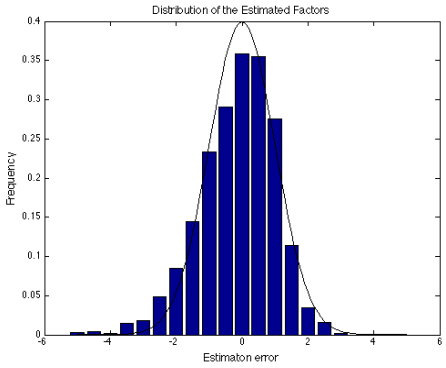
Figure 1: Distribution of the Estimated Factors



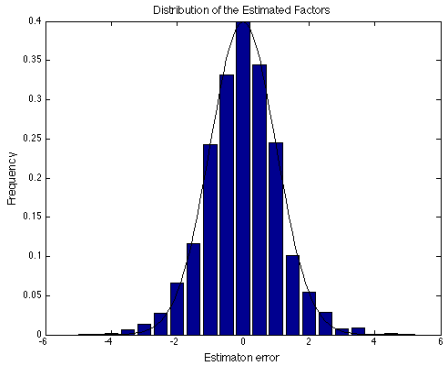
Logit, $N = 50, T = 50$.



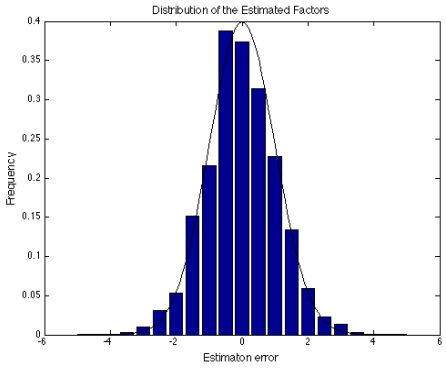
Logit, $N = 100, T = 100$.



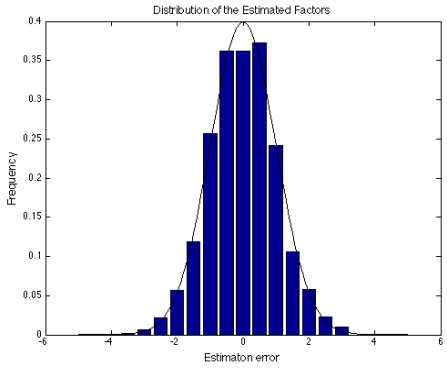
Probit, $N = 50, T = 50$.



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



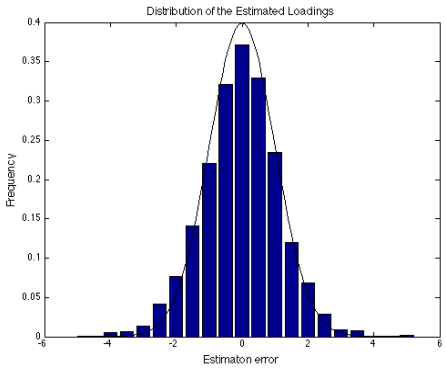
Mixed, $N = 50, T = 50$.



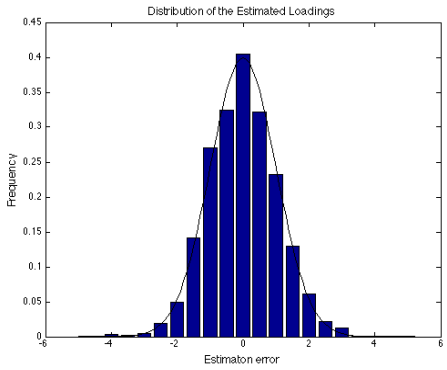
Mixed, $N = 100, T = 100$.

Notes: These histograms are for the standardized estimated factors. The curve overlaid on the histograms is the standard normal density function.

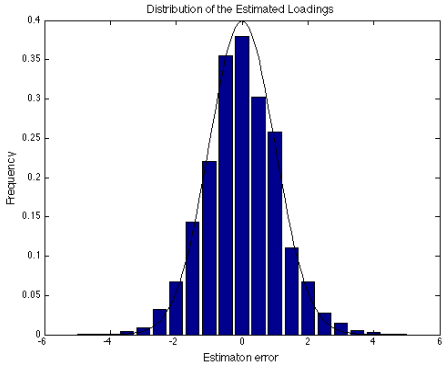
Figure 2: Distribution of the Estimated Loadings (Logit and Probit)



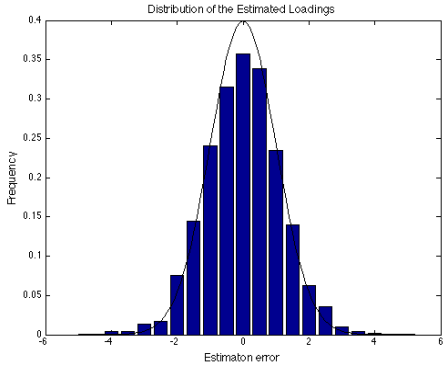
Logit, $N = 50$, $T = 50$.



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



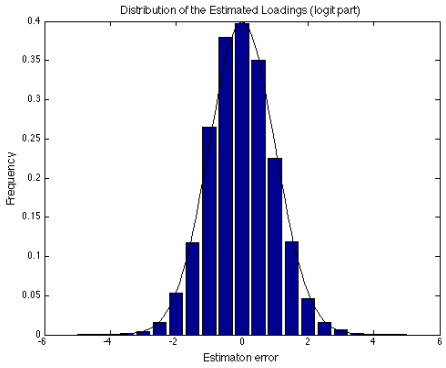
Probit, $N = 50$, $T = 50$.



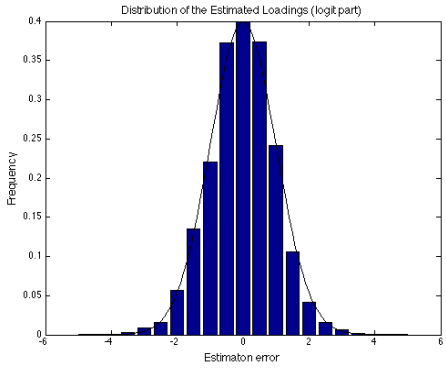
Probit, $N = 100, T = 100$.

Notes: These histograms are for the standardized estimated loadings. The curve overlaid on the histograms is the standard normal density function.

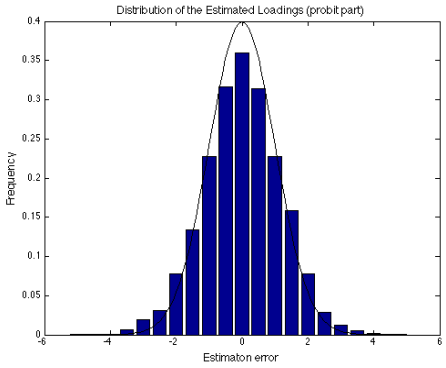
Figure 3: Distribution of the Estimated Loadings (Mixed)



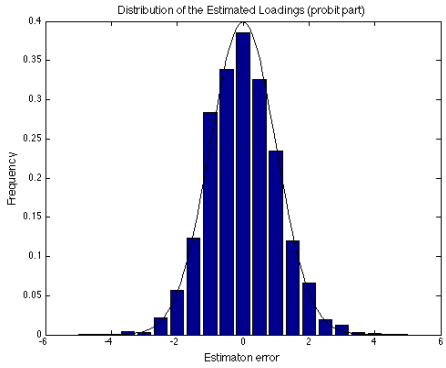
Mixed (logit part), $N = 50, T = 50$.



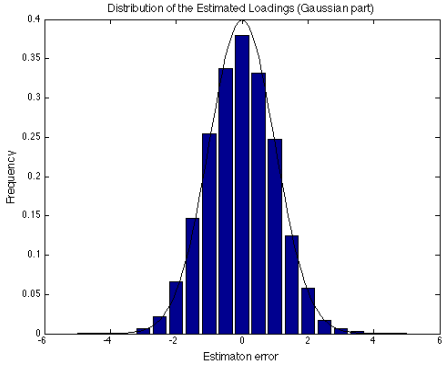
Mixed (logit part), $N = 100, T = 100$.



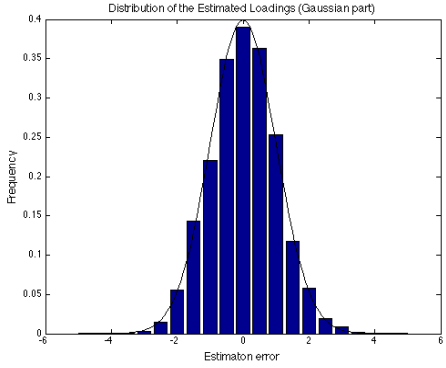
Mixed (probit part), $N = 50, T = 50$.



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



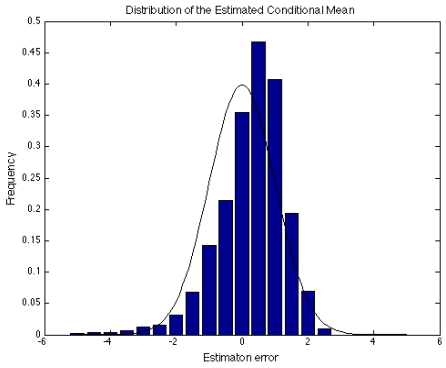
Mixed (Gaussian part), $N = 50, T = 50$.



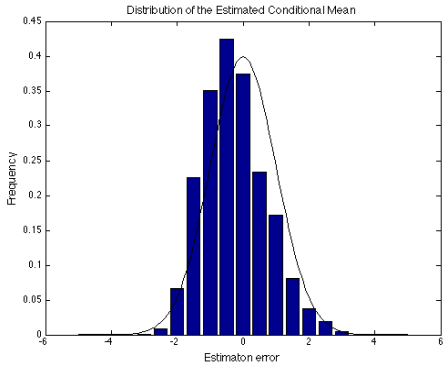
Mixed (Gaussian part), $N = 100, T = 100$.

Notes: These histograms are for the standardized estimated loadings. The curve overlaid on the histograms is the standard normal density function.

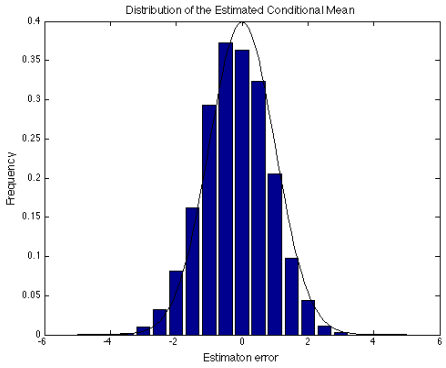
Figure 4: Distribution of the Estimated Conditional Mean



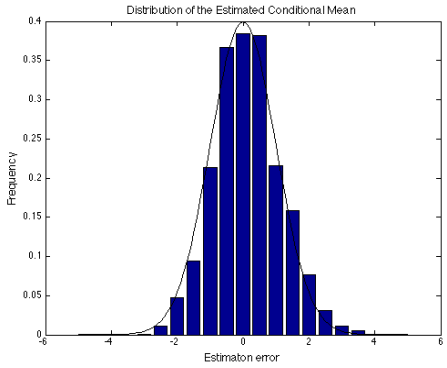
Logit, $N = 50$, $T = 50$.



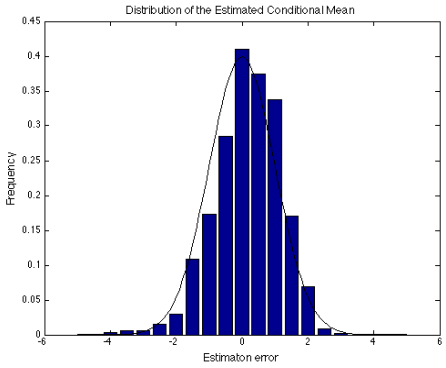
Logit, $N = 100, T = 100$.



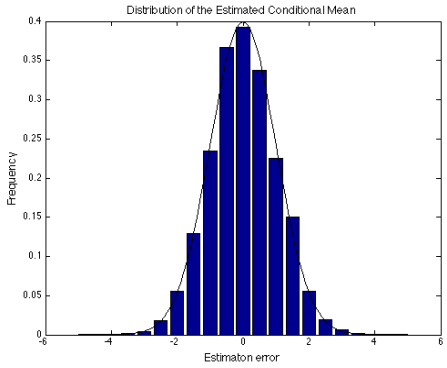
Probit, $N = 50, T = 50$.



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



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



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

Notes: These histograms are for the standardized estimated conditional mean. The curve overlaid on the histograms is the standard normal density function.

References

[1] Bai (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135-171.

[2] Bai and Li (2012): "Statistical Analysis of Factor Models of High Dimension," *Annals of Statistics,* 436-465.

[3] Bai and Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

[4] Bai and Ng (2006): "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-augmented Regressions," *Econometrica*, 74, 1133-1150.

[5] Bartholomew (1980): "Factor Analysis for Categorical Data," *Journal of the Royal Statistical Society. Series B*, 293-321.

[6] Bartholomew and Knott (1999): Latent Variable Models and Factor Analysis. Edward Arnold.

[7] Bernanke, Boivin and Eliasz (2005): "Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) approach," *Quarterly Journal of Economics*, 120, 387–422.

[8] Bohning and Lindsay (1988): "Monotonicity of Quadratic-approximation Algorithms," *Annals of the Institute of Statistical Mathematics*, 40, 641–663.

[9] Campbell, Lo and Mackinlay (1997): The Econometrics of Financial Markets. New Jersey: Princeton University Press.

[10] Chen (2016): "Estimation of Nonlinear Panel Models with Multiple Unobserved Effects," Warwick Economics Research Paper Series.

[11] Chen, Fernandez-Val and Weidner (2014): "Nonlinear Panel Models with Interactive Effects," arXiv preprint arXiv:1412.5647.

[12] Collins, Dasgupta and Schapire (2001): "A Generalization of Principal Component Analysis to the Exponential Family," *Advances in Neural Information Processing System,* Vol. 13.

[13] Cox and Reid (1987): "Parameter Orthogonality and Approximate Conditional Inference," *Journal of the Royal Statistical Society. Series B*, 1-39.

[14] Creal, Koopman and Lucas (2013): "Generalized Autoregressive Score Models with Applications," *Journal of Applied Econometrics*, 28, 777-795.

[15] Creal, Schwaab, Koopman and Lucas (2014): "Observation-driven Mixed-measurement Dynamic Factor Models with an Application to Credit Risk," *Review of Economics and Statistics*, 96, 898-915.

[16] de Leeuw (2006): "Principal Component Analysis of Binary Data by Iterated Singular Value Decomposition," *Computational Statistics & Data Analysis,* 50, 21-39.

[17] Fernandez-Val and Weidner (2016): "Individual and Time Effects in Nonlinear Panel Models with Large N, T," *Journal of Econometrics,* 192, 291-312.

[18] Filmer and Pritchett (2001): "Estimating Wealth Effects without Expenditure Data—or Tears: An Application to Educational Enrollments in States of India," *Demography*, 38, 115-132.

[19] Forni and Reichlin (1998): "Let's Get Real: A Factor-Analytic Approach to Disaggregated Business Cycle Dynamics," *Review of Economic Studies*, 65, 453–473.

[20] Greene (2004): "The Behavior of the Fixed Effects Estimator in Nonlinear Models," *Econometrics Journal*, 7, 98–119.

[21] Hahn and Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72, 1295–1319.

[22] Heckman (1981): "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete time-discrete data Stochastic Process," *Structural Analysis of Discrete Data with Econometric Applications*, 179–195.

[23] Hunter and Lange (2004): "A Tutorial on MM Algorithms," *American Statistician*, 58, 30-37.

[24] Joreskog and Moustaki (2001): "Factor Analysis of Ordinal Variables: A Comparison of Three Approaches," *Multivariate Behavioral Research*, 36, 347-387.

[25] Lancaster (2000): "The Incidental Parameter Problem since 1948," *Journal of Econometrics*, 95, 391-413.

[26] Lancaster (2002): "Orthogonal Parameters and Panel Data," *Review of Economic Studies*, 69, 647-666.

[27] Lange, Hunter and Young (2000): "Optimization Transfer Using Surrogate Objective Functions," *Journal of Computational and Graphical Statistics*, 9, 1-20.

[28] Moustaki (1996): "A Latent Trait and a Latent Class Model for Mixed Observed Variables," *British Journal of Mathematical and Statistical Psychology*, 49, 313-334.

[29] Moustaki (2000): "A Latent Variable Model for Ordinal Variables," *Applied Psychological Measurement*, 24, 211-223.

[30] Moustaki and Knott (2000): "Generalized Latent Trait Models," *Psychometrika*, 65, 391-411.

[31] Newey and McFadden (1994): "Large sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, Vol. IV, 2111-2245.

[32] Neyman and Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32.

[33] Ng (2015): "Constructing Common Factors from Continuous and Categorical Data," *Econometric Reviews*, 34, 1141-1171.

[34] Ross (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Finance*, 13, 341–360.

[35] Schein, Saul and Ungar (2003): "A Generalized Linear Model for Principal Component Analysis of Binary Data," *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*.

[36] Schonbucher (2000): "Factor Models for Portfolio Credit Risk," Bonn Econ Discussion Papers 16.

[37] Stock and Watson (2002): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of American Statistical Association*, 97, 1167–1179.

[38] Stock and Watson (2016): "Factor Models and Structural Vector Autoregressions in Macroeconomics," *Handbook of Macroeconomics*, forthcoming.