Maximum Likelihood Estimation for Factor-Augmented Binary Response Model

Jiachen Cong

School of Finance
Nankai University

2024 International Workshop on Econometrics
with Its Application and Practice in Finance

2024.7.2, Tianjin

Outline

1. Introduction
2. Estimation
3. Asymptotic Theory
4. Simulation
5. Conclusion
6. Empirical analysis

Introduction

Motivation: binary variable

- In finance and economics fields, binary variables are frequently used to indicate or predict whether a certain event will happen.
- Binary response models have a very wide range of applications. For instance, they can be used to resolve the following problems:
  - Predicting the probability of corporate failure (Caggiano et al.,2014).
  - Credit rating analysis (Jones et al.,2015).
  - Portfolio optimization (Anagnostopoulos & Mamanis,2010).
  - Financial assets selection (Ferreira et al.,2020).
- When analyzing binary variables, forecasting the probability of certain events is highly significant.
- In the era of big data, how can we predict binary variables efficiently?

Motivation: factor model

- Factor model paradigm provides us an efficient method to extract information from high-dimensional data.
- Factor models also have a wide range of applications, for instance:
  - Macroeconomic indicators prediction (Bernanke, Boivin and Eliasz, 2005).
  - Public policy analysis (Chong, He and Leung, 2013).
  - Financial risk factors analysis and forecasting (Fan, Ke and Liao, 2021).
- The factor model framework helps us choose appropriate number of regressors, avoiding problems of over-fitting and intensive computation that traditional regression models have when there are too many explanatory variables.
- With help of the factor model framework, we can conduct causal inference and forecasting when we dealing with large scale data.

Motivation: Existing models

Combining two features above, here comes the following question:

- How to use factor model to process high dimensional predictors to predict binary response variable?

However, few papers have addressed this question. The existing studies mainly focus on factor-augmented models with continuous response variable. For example,

- Bai & Ng (2006) introduced a factor-augmented forecasting regression model with continuous response variable.

$$y_{t+h} = \alpha^\top F_t + \beta^\top \omega_t + \epsilon_{t+h} \tag{1}$$

  where $t = 1, 2, \cdots, T-h$, F_t and ω_t are unobservable factors and observable predictors, respectively. The factors F_t can be recovered from a group of observable variables X_{it}.

- Yan & Cheng (2022) added threshold effect into factor-augmented model:

$$y_{t+h} = \alpha^\top F_t + \beta^\top \omega_t + \theta_T^\top F_t I(q_t \le r_0) + \eta_T^\top \omega_t I(q_t \le r_0) + \epsilon_{t+h} \tag{2}$$

  where r_0 is an unknown threshold parameter.

Motivation: Existing models

Besides, some papers investigate binary response models with common factors:

- Wang (2020) proposes an approach that can be used to estimate the common factor structure in binary response models:

$$y_{it} = \begin{cases} 1, & \text{if } \lambda_i^\top F_t - \epsilon_{it} \geq 0 \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

  for $i = 1, \cdots, N$ and $t = 1, \cdots, T$, where λ_i are corresponding factor loadings.

- Gao et al. (2023) provide analysis for a binary response panel data model with interactive fixed effects:

$$y_{it} = \begin{cases} 1, & \text{if } \beta_i^\top \omega_{it} + \lambda_i^\top F_t - \epsilon_{it} \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Factor-augmented binary response model

In this paper, we introduce the following factor-augmented binary response model, for $t = 1, 2, \cdots, T$ and a positive integer h,

$$y_{t+h} = \begin{cases} 1, & \text{if } \alpha^\top F_t + \beta^\top \omega_t - \epsilon_{t+h} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where w_t and F_t are p-dimensional and k-dimensional vectors of observable predictors and unobservable factors, respectively; α and β are unknown parameters; ϵ_{t+h} is the idiosyncratic error term with known PDF and CDF. Additionally, the factors F_t can be recovered from a group of observable variables X_{it}:

$$X_{it} = \lambda_i^\top F_t + e_{it}, \quad i = 1, 2, \cdots, N, \tag{6}$$

where λ_i is a vector of factor loadings and e_{it} is another idiosyncratic error.

Our contributions

- We introduce a factor-augmented binary response model, which can complement Bai & Ng(2006)'s original model under binary response variable condition. This can be considered as a generalization of classical factor-augmented model.
- We propose a maximum likelihood estimation method to estimate the unknown parameters and establish the corresponding asymptotic properties.
- We examine the finite-sample performance of the estimators through simulation studies and perform an empirical application to forecast the rise and fall of the gold futures prices .

Estimation

Estimation method

Step 1: Estimate factors F_t.
Based on (6), we estimate $F = (F_1, \ldots, F_T)^\top$ using PCA method:

$$\frac{1}{NT} X X^\top \tilde{F} = \tilde{F} V_{NT}, \tag{7}$$

where $X = (X_1, \ldots, X_N)$ with $X_i = (X_{i1}, \ldots, X_{iT})^\top$, V_{NT} is a $k \times k$ diagonal matrix that contains the first k largest eigenvalues of $\frac{1}{NT} X X^\top$ in descending order, $\tilde{F}$ is a $T \times k$ matrix that contains the corresponding eigenvectors and satisfies the following identification condition:

$$\frac{1}{T} \tilde{F}^\top \tilde{F} = I_k. \tag{8}$$

Estimation method

Step 2: Estimate α and β.
Let $z_t = (F_t^\top, \omega_t^\top)^\top$, $r = (\alpha^\top, \beta^\top)^\top$, simple algebra yields

$$P(y_{t+h} = 1 | z_t) = P(\epsilon_{t+h} \leq r^\top z_t) = \Phi_\epsilon(r^\top z_t), \tag{9}$$

where $\Phi_\epsilon(\cdot)$ is the CDF of ϵ_{t+h}. Then for y=0 or 1, we have:

$$P(y_{t+h} = y | z_t) = \Phi_\epsilon(r^\top z_t)^y [1 - \Phi_\epsilon(r^\top z_t)]^{1-y}. \tag{10}$$

Drawing upon this fact, we construct the following (infeasible) likelihood function:

$$L_0(r) = \prod_{t=1}^{T-h} \Phi_\epsilon(r^\top z_t)^{y_{t+h}} [1 - \Phi_\epsilon(r^\top z_t)]^{1-y_{t+h}}. \tag{11}$$

Estimation method

By replacing F_t with $\tilde{F}_t$, we obtain

$$L(r) = \prod_{t=1}^{T-h} \Phi_\epsilon(r^\top \tilde{z}_t)^{y_{t+h}} [1 - \Phi_\epsilon(r^\top \tilde{z}_t)]^{1-y_{t+h}}, \tag{12}$$

where $\tilde{z}_t = (\tilde{F}_t{}^\top, \omega_t^\top)^\top$.
The log-likelihood function is given as follows:

$$\log[L(r)] = \sum_{t=1}^{T-h} [y_{t+h} \log \Phi_\epsilon(r^\top \tilde{z}_t) + (1 - y_{t+h}) \log(1 - \Phi_\epsilon(r^\top \tilde{z}_t))]. \tag{13}$$

Therefore, we can construct the MLE estimators:

$$\hat{r} = \arg\max_r \log[L(r)]. \tag{14}$$

where we define $\hat{r} = (\hat{\alpha}^\top, \hat{\beta}^\top)^\top$.

Asymptotic Theory

Assumptions

Let $\Lambda = (\lambda_1, \ldots, \lambda_N)^\top$ and $E_t = (e_{1t}, \ldots, e_{Nt})^\top$.

Assumption 1

(1) Let $\{\epsilon_t, E_t, F_t, w_t\}$ be strictly stationary and α-mixing across t with the α-mixing coefficient $\alpha(|t-s|)$ satisfying that $\sum_{t=0}^{\infty}(\alpha(t))^{\delta/(4+\delta)} = O(1)$. Additionally, $E[\|v_{it}\|^{4+\delta}] < \infty$, for $v_{it} \in \{\epsilon_t, e_{it}, F_t, w_t, \lambda_i\}$.

(2) For each i, j, s, t, let $\sigma_{ij,ts} = E(e_{it}e_{js})$. Assume that $E\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(e_{it}e_{is} - \sigma_{ii,ts})\right]^4 = O(1)$. Additionally, $\max_{s,t}\frac{1}{N}\sum_{i,j=1}^{N}\sigma_{ij,st} = O(1)$ and $\frac{1}{NT}\sum_{i,j;s,t}|\sigma_{ij,st}| = O(1)$.

(3) Assume that F_t and w_t are uniformly bounded.

(4) ϵ_t are independent with $\{e_{it}, F_t, w_t, \lambda_i\}$; $\{e_{it}\}$ are independent with $\{F_t, \lambda_i\}$.

(5) $N^{-1}\Lambda^\top\Lambda \to_P \Sigma_\lambda$, where Σ_λ is positive definite.

(6) There exists a set $\Xi_T = [\Xi_T^l, \Xi_T^u]$ such that all ξ_t's belong to Ξ_T with probability approaching 1, and $0 < \Phi_\epsilon(\Xi_T^l) < \Phi_\epsilon(\Xi_T^u) < 1$, where $\xi_t = z_t^\top r$.

(7) There exists a positive constant c such that $\inf_{z\in\Xi_T}|\phi_\epsilon(z)| > c$, where $\phi_\epsilon(\cdot)$ is the PDF of ϵ_t. $\phi_\epsilon(\cdot)$ is twice differentiable on Ξ_T.

(8) $\sup_{z\in\Xi_T}(|l_t^{(2)}(z)| + |l_t^{(3)}(z)|) < \infty$ uniformly, where $l_t^{(k)}(z)$ is the k^{th} derivative of $l_t(z) = y_{t+h}\log\Phi_\epsilon(z) + (1-y_{t+h})\log(1-\Phi_\epsilon(z))$ for $k = 2, 3$.

Assumptions

Assumption 2

(1) Assume that $\Sigma_r = E\left[\frac{\phi_\epsilon^2(r^\top z_t)}{[1-\Phi_\epsilon(r^\top z_t)]\Phi_\epsilon(r^\top z_t)} z_t z_t^\top\right]$ and Σ_r is positive definite.

(2) Assume that $\Omega_r = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T-h}\sum_{s=1}^{T-h} E\left[m_{t+h}m_{s+h}z_t z_s^\top\right]$, where $m_{t+h} = \frac{(y_{t+h}-\Phi_\epsilon(r^\top z_t))\phi_\epsilon(r^\top z_t)}{(1-\Phi_\epsilon(r^\top z_t))\Phi_\epsilon(r^\top z_t)}$, and Ω_r is positive definite.

Asymptotic Theory

Theorem 1 (Consistency) Let Assumption 1 hold. As $N, T \to \infty$,

(1) $\|\hat{\alpha} - H^{-1}\alpha\| = o_P(1)$, where $H = (NT)^{-1}(\Lambda^\top \Lambda)(F^\top \tilde{F})V_{NT}^{-1}$;

(2) $\|\hat{\beta} - \beta\| = o_P(1)$.

Theorem 2 (CLT) Let Assumptions 1 and 2 hold. As $N, T \to \infty$ and $\sqrt{T}/N \to 0$,

$$\sqrt{T}(\hat{r} - \tilde{H}r) \xrightarrow{D} \mathcal{N}(0, H_0 \Sigma_r^{-1} \Omega_r \Sigma_r^{-1} H_0^\top), \tag{15}$$

where $\tilde{H} = \text{diag}(H^{-1}, I)$, $H_0 = \text{diag}(VQ^{-1}\Sigma_\lambda^{-1}, I)$ with $V = \text{plim}V_{NT}$, $Q = \text{plim}T^{-1}F^\top \tilde{F}$, and Σ_r and Ω_r are defined in Assumption 2.

Simulation

Data generating processes

- In this section, we conduct a simulation experiment to examine the finite sample performance of our estimation method.
- Following Bai & Ng (2006) and Yan & Cheng (2022), we generate factors F_t as follows:

$$X_{it} = \lambda_i^\top F_t + e_{it} \tag{16}$$

  where $i = 1, 2, \cdots, N$, $t = 1, 2, \cdots, T$, N is the number of observable variables.

$$F_{jt} = \rho_j F_{j,t-1} + (1 - \rho_j^2)^{\frac{1}{2}} \mu_{jt} \tag{17}$$

  where $j = 1, 2, \cdots, m$, $\rho_j = 0.8^j$, m is the number of factors in factor model.
- λ_i is generated from the uniform distribution $U(0, 6)$, e_{it} and μ_{jt} are independently and identically distributed (i.i.d.) from $N(0, 1)$.

Data generating processes

- In the simulations, we consider the model:

$$y_{t+2} = \begin{cases} 1, & \text{if } \alpha_1 F_{t1} + \alpha_2 F_{t2} + \beta_0 + \beta_1 \omega_{t1} + \beta_2 \omega_{t2} - \epsilon_{t+2} \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

- Observable regressors ω_{t1} and ω_{t2} are generated from uniform distributions $U(0,2)$ and $U(-3,3)$, respectively; we assume m=2 so there are 2 factor regressors F_{t1} and F_{t2}; ϵ_{t+2} is generated from normal distribution $N(0,1)$.
- We set $\beta_0 = -1$, $\beta_1 = 1$, $\beta_2 = 1$, $\alpha_1 = 1$, $\alpha_2 = 1$. In vector form, we have $\alpha = (1,1)^\top$, $\beta = (-1,1,1)^\top$.

Data generating processes

- We set N to be 100,200,300 and T to be 100,200,400. For each pair of (N, T), we conduct 500 replications.
- We measure the accuracy of the estimates by computing

$$P_{error} = \|(\hat{\alpha}^\top, \hat{\beta}^\top)^\top - ((H^{-1}\alpha)^\top, \beta^\top)^\top\| \tag{19}$$

  Where $\|\cdot\|$ is the frobenius norm of a certain vector or matrix and $H = V_{NT}^{-1}(\tilde{F}^\top F/T)(\Lambda^\top \Lambda/N)$.
- We evaluate the results of the simulation based on the sample mean $\overline{P_{error}}$ and sample variance $\widehat{Var(P_{error})}$ of P_{error}.

Simulation results

The simulation results are presented as follows:

Table: Simulation results

N	T	$\overline{P_{error}}$	$Var\widehat{(P_{error})}$
100	100	0.8125	0.4470
100	200	0.5219	0.2943
100	400	0.3434	0.1957
200	100	0.8176	0.4614
200	200	0.5006	0.2622
200	400	0.3342	0.1738
300	100	0.7930	0.4840
300	200	0.5233	0.2966
300	400	0.3290	0.1766

Simulation results

- We can find that keeping the number of factors N constant, as T increases, the values of $\overline{P_{error}}$ and $\widehat{Var(P_{error})}$ decrease, which are consistent with the asymptotic theories we established.
- We can also find that $\overline{P_{error}}$ and $\widehat{Var(P_{error})}$ and their convergence rate is more sensitive to changes in T than changes in N. Changes in N have slight effect on values of those indicators. This result also corresponding to our asymptotic theories.

Conclusion

Conclusion

- In this paper, we have introduced a factor-augmented model with binary response variable. We have developed a maximum likelihood estimation for regression parameters and established the corresponding asymptotic theories for slope coefficients.
- Besides, identification restrictions provided in Bai & Ng(2002,2006), Wang(2020), Yan & Cheng(2022) are readily applicable in our model with slight modification. When it comes to resolving heteroscedasticity and serial correlation issues, we just need to modify the maximum likelihood function in case we know the exact class of distribution and conditional probability density function.
- Finally, we examine the usefulness of the proposed model through an application to forecasting the changing trend of gold futures prices.

Thank you!

Empirical analysis

Empirical Analysis:a brief introduction

- Current research have shown that in predicting the price movement of gold futures, it is possible to utilize the corresponding leading indicators to achieve the relevant price prediction.
- However, if too many predictors are used, this can lead to problems of over-fitting or requiring too much computation.
- In this section, we employ our model to forecast the price changing trend of gold futures. Our prediction have 2 results,"price goes up" & "price goes down".
- If our model's output is greater or equal to 0.5, we consider it as "the price goes up"; otherwise, we consider it as "the price goes down".

Empirical Analysis:steps & data

- Steps we take to conduct this experiment are to train the model using data from the training set and to test the predictive accuracy of the model on the test set.
- Training set is from 2020.09.01 to 2023.08.04 (707 days); we conduct out-of-sample forecasting on data set from 2023.08.04 to 20232.12.29 (100 days).
- We take following 10 predictors into consideration: CSI 300 Index Yield,Silver Futures Price Change Rate,Platinum Futures Price Change Rate,Crude Oil Futures Price Change Rate,Gold Futures Volume Change Rate,Gold Futures Position Change Rate,Change rate of USD-RMB exchange rate,US Dollar Index Change Rate,Maximum intraday spread of gold futures,Gold futures price change rate.
- Based on 10 predictors, we construct 2 factor estimators and then put them into our model as predictors.

Empirical Analysis:models & results

- We consider factor-augmented forecasting model as follows:

$$y_{t+1} = \alpha + \Sigma_{i=1}^{2} \lambda_i * F_{it} + \epsilon_{t+1} \tag{20}$$

- We consider a moving average model as follows:

$$y_{t+1} = \frac{y_t + y_{t-1} + y_{t-2} + y_{t-3} + y_{t-4}}{5} \tag{21}$$

- Similar to results of the factor model, if the result of moving average model is greater or equal to 0,we consider it as "the price goes up"; otherwise, we consider it as "the price goes down".
- In test set, our accuracy is 64 percent while traditional method is 52 percent.
- Notice: We assume that the market structure is relatively stable, so the weights of PCA estimators and parameters of the factor model are only determined by the training set.