Statistics and Its Interface Volume 13 (2020) 65–75

Adaptive LASSO regression against heteroscedastic idiosyncratic factors in the covariates

Kaimeng Zhang and Chi Tim Ng

Recent studies suggest that by including the principal components of the covariates, LASSO regression achieves certain consistency properties when the idiosyncratic factors are homoscedastic. In this paper, it is shown that if the principal components are replaced by the common factors obtained based on the maximum likelihood estimation of factor model and the covariates are replaced by the estimated idiosyncratic factors, selection consistency holds even in the heteroscedestic cases. The new results hold for both LASSO and adaptive LASSO under the high-dimensional settings with $p \to \infty$ but $p = o(n)$, where p and n are the number of components of the covariates and the number of observations respectively. Simulation studies suggest that when the idiosyncratic factors are heteroscedastic, penalized regression based on factor analysis outperforms that based on principal component analysis. To illustrate the ideas, real data examples of international economic input-output data and international stock indexes data are studied in particular.

AMS 2000 subject classifications: Primary 62J05, 62F12; secondary 62H25.
Keywords and phrases: Factor analysis, Global economic interaction, Irrespentable condition, Adaptive LASSO, Penalized regression, Selection consistency.

1. INTRODUCTION

Penalized least square estimation methods has been extensively studied for variable selection since the introduction of least absolute shrinkage and selection operator (LASSO) in [20] and the subsequent work of adaptive LASSO in [24]. Going beyond LASSO, a number of alternative penalties are proposed, for example, [5], [11], [22], [13], and [18] propose alternative penalties to LASSO. The selection consistency of the penalized regression methods widely studied in the literature, to name a few, [5], [9], [6], [14], [15], [16], and [18].

In spite of the remarkable attention among the statisticians on the topic of penalized regression, serious discussion on the impacts of the dependence structure of the covariates on the variable selection is limited. An exception is the work of [10] that introduce the so-called augmented model by including common factors on top of the covariates in the regression models. Here, the common factors are estimated as the principal components of the covariates. Decomposing the covariate into common factors and idiosyncratic factors allow one to perform variable selection under a more general situation where the response is generated from a linear model involving common factors and idiosyncratic factors. This encompasses the usual regression model against the covariates as a special case. In the usual regression settings without factor analysis, selection consistency of LASSO estimation is established in [23] under the so-called "irrepresentable condition" and similarly for adaptive LASSO in [24]. The crucial idea is that the dependence between the relevant covariates and the irrelevant covariates cannot be too strong. If the regression model is used without considering common and idiosyncratic factors, the "irrepresentable condition" can be too stringent in many practical situations. Fortunately, common and idiosyncratic factors are independent of each other. Roughly speaking, if the estimation error in the factor model is small, selection consistency can be satisfied easily.

Common factors and idiosyncratic factors can be obtained by either principal component analysis or maximum likelihood estimation of the factor model. As noted in [1], estimation based on principal component analysis entails homoscedasticity of the idiosyncratic factors that is restrictive to hold. Therefore, the results of [10] are applicable in the homoscedasticity cases only. To allow heteroscedasticity of the idiosyncratic factor, the principal components are replaced by the common factors obtained based on the maximum likelihood estimation of factor model and the covariates are replaced by the estimated idiosyncratic factors. In this paper, selection consistency, see [5] and [23] is formally established under the heteroscedasticity settings. In addition, new definition of "irrepresentable condition" is provided so as to take the estimation error into account. It is also illustrated through simulation that the K-fold cross validation (see [7]) can be used to select the tuning parameter in the LASSO penalty.

This paper is organized as follows. In section 2, the model and assumptions are presented. The penalized regression method against idiosyncratic factors ($PRAIF$) is described.

The selection consistency of $PRAIF$ is established in section 3. In section 4, simulation studies are given. Though the idiosyncratic factors can be obtained from the principal components of the covariates as suggested in [10], it is illustrated that under the strong heteroscedasticity settings on the idiosyncratic factors, the $PRAIF$ performs better if the idiosyncratic factors are estimated based on maximum likelihood estimation of the factor model instead. Section 5 presents empirical data examples of international economic input/output and global financial data of stock indexes. In particular, we study the impacts of foreign capital and labor inputs on the domestic economic output, that is the gross domestic production. Here, the capital and labor inputs of all foreign countries can be correlated to each other due to global systematic factors. Concluding remarks are given in section 6.

2. PENALIZED REGRESSION AGAINST IDIOSYNCRATIC FACTORS

In this section, the factor model for the covariates is described. The response is regressed against the common factors and idiosyncratic factors estimated based on maximum likelihood estimation of factor model.

2.1 Modeling covariates with factor model

For $t = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$, let y_t be the response and $x_t = (x_{1t}, x_{2t}, \ldots, x_{pt})^T$ be the $p \times 1$ vector of observed covariates. Consider the model

$$\begin{aligned} y_t &= \boldsymbol{b}^T \boldsymbol{f}_t + \boldsymbol{\gamma}^T \boldsymbol{\eta}_t + \epsilon_t, && (1)\\ \boldsymbol{x}_t &= \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\eta}_t\,, && (2) \end{aligned}$$

where the common factors $\boldsymbol{f}_{t1}, ..., \boldsymbol{f}_{tm}$ are independent $N(0, \boldsymbol{I}_m)$ random variables and the idiosyncratic factors $\boldsymbol{\eta}_t$ are independent $N(0, \boldsymbol{\Phi})$ random vectors with $\boldsymbol{\Phi} = diag\left(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2\right)$ and $C^{-1} \leq \sigma_j^2 \leq C$ for all $j = 1, 2, \ldots, p$ for some sufficiently large positive constant C. The errors $\epsilon_t = (\epsilon_{1t}, \epsilon_{2t}, \ldots, \epsilon_{pt})^T$ are $N(0, \sigma^2)$ random variables. For all $t = 1, 2, \ldots, n$, $\boldsymbol{f}_t$, $\boldsymbol{\eta}_t$, and ϵ_t are independent. The factor loading $\boldsymbol{\Lambda}$ is a $p \times m$ matrix. For model identification, $\boldsymbol{\Lambda}$ is rotated so that $\frac{1}{n}\boldsymbol{\Lambda}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\Lambda}$ is diagonal. In addition, suppose that all conditions as described in [1] for the "average consistency" of the maximum likelihood estimations of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ hold. $\boldsymbol{b}$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)$ are the coefficients vectors against the common factors and the idiosyncratic factors respectively.

It is interesting to note that when $\boldsymbol{b}^T = \boldsymbol{\gamma}^T \boldsymbol{\Lambda}$, the model (1)-(2) reduces to the usual linear regression model that the response is regressed against the covariates, otherwise, the usual linear regression model is misspecified. Under such a misspecification case, the optimal predictor as defined in [16] is the conditional expectation

$$\begin{aligned} & E\left(y_t | \boldsymbol{x}_t\right) \\ = \; & \boldsymbol{\gamma}^T \boldsymbol{x}_t + \left(\boldsymbol{b}^T - \boldsymbol{\gamma}^T \boldsymbol{\Lambda}\right) E(f_t | \boldsymbol{x}_t) \\ = \; & \boldsymbol{\gamma}^T \boldsymbol{x}_t \\ & + \left(\boldsymbol{b}^T - \boldsymbol{\gamma}^T \boldsymbol{\Lambda}\right) \left(\boldsymbol{I}_m + \boldsymbol{\Lambda}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\Lambda}\right)^{-1} \boldsymbol{\Lambda}^T \boldsymbol{\Phi}^{-1} \boldsymbol{x}_t \\ = \; & \boldsymbol{x}_t^T \left[\boldsymbol{\gamma} + \boldsymbol{\Phi}^{-1} \boldsymbol{\Lambda} \left(\boldsymbol{I}_m + \boldsymbol{\Lambda}^T \boldsymbol{\Phi}^{-1} \boldsymbol{\Lambda}\right)^{-1} (\boldsymbol{b} - \boldsymbol{\Lambda}\boldsymbol{\gamma})\right] \\ = \; & \boldsymbol{x}_t^T \boldsymbol{\gamma}^\dagger\,. \end{aligned}$$

It can be seen that when $\boldsymbol{b}^T \neq \boldsymbol{\gamma}^T \boldsymbol{\Lambda}$, the sets $\{i = 1, 2, \ldots, p : \boldsymbol{\gamma}_i = 0\}$ and $\{i = 1, 2, \ldots, p : \boldsymbol{\gamma}_i^\dagger = 0\}$ are different in general.

2.2 Penalized likelihood estimation

Let $\hat{\boldsymbol{f}}_t$ and $\hat{\boldsymbol{\eta}}_t$, $t = 1, 2, \ldots, n$ be the estimated common factors and idiosyncratic factors obtained by the expectation maximization algorithm described in Appendix C. The adaptive LASSO estimator is defined as

$$\begin{aligned} & \left(\hat{\boldsymbol{b}}, \hat{\boldsymbol{\gamma}}\right) = \\ (3) \quad & \underset{b,\gamma}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{t=1}^{n} \left(y_t - \boldsymbol{b}^T \hat{\boldsymbol{f}}_t + \boldsymbol{\gamma}^T \hat{\boldsymbol{\eta}}_t\right)^2 + \lambda \sum_{j=1}^{p} \omega_j |\gamma_j| \right\}, \end{aligned}$$

where $\hat{\boldsymbol{\omega}}$ is a weight vector. When $\hat{\boldsymbol{\omega}} = (1, 1, \ldots, 1)^T$ is chosen, the adaptive LASSO estimation reduces to the LASSO estimation. Alternatively, one can choose $\hat{\boldsymbol{\omega}} = \left(1/\hat{\gamma}_1^*, 1/\hat{\gamma}_2^*, \ldots, 1/\hat{\gamma}_p^*\right)^T$. Here, for $j = 1, 2, \ldots, p$, $\hat{\gamma}_j^*$ is an estimator of γ_j, for example, the ordinary least squares estimator or LASSO estimator.

The number of factors m is chosen so that the first m factors explains 95% of the total variation in the covariates.

The tuning parameter λ can be chosen based on the K-fold cross-validation method as described below. Let K be an integer. The sample $I = \{1, 2, \ldots, n\}$ is randomly partitioned into K equal-sized subsamples I_k, $k \in 1, \ldots, K$. Let $n_k = |I_k|$ be the size of the subset I_k. Define

$$\begin{aligned} & \left(\hat{\boldsymbol{b}}_{-k}(\lambda), \hat{\boldsymbol{\gamma}}_{-k}(\lambda)\right) = \underset{b,\gamma}{\operatorname{argmin}} \\ (4) \quad & \left\{ \frac{1}{n - n_k} \sum_{t \notin I_k} \left(y_t - \boldsymbol{b}^T \hat{\boldsymbol{f}}_t + \boldsymbol{\gamma}^T \hat{\boldsymbol{\eta}}_t\right)^2 + \lambda \sum_{j=1}^{p} \omega_j |\gamma_j| \right\}. \end{aligned}$$

Then, λ is chosen by minimizing

$$\sum_{k=1}^{K} \sum_{t \in I_k} \left(y_t - \hat{\boldsymbol{b}}_{-k}^T(\lambda) \hat{\boldsymbol{f}}_t + \hat{\boldsymbol{\gamma}}_{-k}^T(\lambda) \hat{\boldsymbol{\eta}}_t\right)^2 . \qquad (5)$$

3. MAIN RESULTS

The theory of selection consistency under the heteroscedasticity assumptions on the idiosyncratic factors is established in this section.

3.1 Notation

Some notations are introduced here. Define $p \times n$ design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T$ and $n \times 1$ response vector $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^T$. Let $\boldsymbol{F}$ and $\boldsymbol{E}$ be $m \times n$ and $p \times n$ matrices containing $\boldsymbol{f}_t$ and $\boldsymbol{\eta}_t$, $t = 1, 2, \ldots, n$. Let $\boldsymbol{G} = (\boldsymbol{F}, \boldsymbol{E})$ and $\boldsymbol{\alpha} = (\boldsymbol{b}, \boldsymbol{\gamma})^T$. Denote by $\hat{\boldsymbol{F}}$, $\hat{\boldsymbol{E}}$, and $\hat{\boldsymbol{G}}$ the estimated values of $\boldsymbol{F}$, $\boldsymbol{E}$, and $\boldsymbol{G}$.

The true coefficient vector $\boldsymbol{\gamma}$ is allowed to be sparse and d is the number of relevant covariates. Let I_0 be the subset of indexes corresponding to the non-zero coefficients for $\{1, 2, \cdots, d\}$ and the I_c be the compliment set of I_0 which includes $\{d+1, d+2, \cdots, p\}$. The sub-matrices $\boldsymbol{X}_{I_0}$ and $\boldsymbol{X}_{I_c}$ contain the columns of $\boldsymbol{X}$ corresponding to relevant and irrelevant covariates respectively. Similarly, define $\boldsymbol{E}_{I_0}$, $\boldsymbol{E}_{I_c}$, $\hat{\boldsymbol{E}}_{I_0}$, and $\hat{\boldsymbol{E}}_{I_c}$. Set $\boldsymbol{G}_{I_0} = (\boldsymbol{F}, \boldsymbol{E}_{I_0})$, $\boldsymbol{G}_{I_c} = \boldsymbol{E}_{I_c}$, $\boldsymbol{\alpha}_{I_0} = (\boldsymbol{b}, \boldsymbol{\gamma}_{I_0})$, and $\boldsymbol{\alpha}_{I_c} = \boldsymbol{\gamma}_{I_c}$.

3.2 Revised irrepresentable conditions

There is a huge literature devoted to studying the statistical properties of the adaptive LASSO method. Selection consistency of the adaptive LASSO estimation can be established under the so-called "strong irrepresentable condition" and some regularity conditions as described in [9], [23], and [24]. "Strong irrepresentable condition" means that

$$\left\| \boldsymbol{X}_{I_c}^T \boldsymbol{X}_{I_0} (\boldsymbol{X}_{I_0}^T \boldsymbol{X}_{I_0})^{-1} sign(\boldsymbol{\beta}_{I_0}) \right\|_\infty \leqslant \zeta \,, \tag{6}$$

where ζ is a positive constant and $0 < \zeta < 1$. The crucial idea is to restrict the dependence between the relevant covariates and irrelevant covariates. Under the following conditions,

(A1) $p = o(n)$ and $p \to \infty$,
(A2) $d = o(p)$,
(A3) $\lambda = o(d^{-1})$ and $(2n^{-1}\log(d))^{1/2} = o_p(\lambda \min_{j \in I_c} |\omega_j|)$,
(A4) $m = O(1)$.

For the penalized regression against the idiosyncratic factors, we establish the following proposition.

Proposition 1. *Let $\boldsymbol{H} = \hat{\boldsymbol{G}}_{I_0}(\hat{\boldsymbol{G}}_{I_0}^T \hat{\boldsymbol{G}}_{I_0})^{-1} \hat{\boldsymbol{G}}_{I_0}^T$ be the hat matrix. Under Conditions (A1) to (A4), the selection consistency holds if the following conditions are satisfied,*

(IR1) $\|\hat{\boldsymbol{G}}_{I_c} \hat{\boldsymbol{G}}_{I_0}^T (\hat{\boldsymbol{G}}_{I_0}^T \hat{\boldsymbol{G}}_{I_0})^{-1} sign(\boldsymbol{\alpha}_{I_0})\|_\infty \leqslant v$ *for some constant* $0 < v < 1$,
(IR2) the equation $\frac{\partial f(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_{I_0}} = 0$ *admits a solution* $\hat{\boldsymbol{\alpha}}_{I_0} = (\hat{\boldsymbol{b}}, \hat{\boldsymbol{\gamma}}_{I_0})$ *so that all entries are non-zero and* $sign(\hat{\boldsymbol{\gamma}}_{I_0}) = sign(\boldsymbol{\gamma}_{I_0})$, *where* $f(\cdot)$ *is the penalized sum-of-squares function,*
(IR3) $\|\hat{\boldsymbol{G}}_{I_c}^T (\boldsymbol{I} - \boldsymbol{H})\epsilon\|_\infty = o(n\lambda \min_{j \in I_c} |\omega_j|)$, *and*
(IR4) $\|\hat{\boldsymbol{G}}_{I_c}^T (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{G}_{I_0} \boldsymbol{\alpha}_{I_0}\|_\infty = o(n\lambda \min_{j \in I_c} |\omega_j|)$.

Condition (IR1) is similar to the "irrepresentable condition" (6) excepted that the covariates $\boldsymbol{X}$ are replaced by the estimated idiosyncratic factors $\boldsymbol{E}$ and common factors $\boldsymbol{F}$. The intuition is that if the estimation error in $\boldsymbol{G} = (\boldsymbol{F}, \boldsymbol{E}_{I_0})$ is negligible, the new covariates $\hat{\boldsymbol{G}}_{I_0}$ and $\hat{\boldsymbol{G}}_{I_c}$ are less correlated than the original covariates $\boldsymbol{X}_{I_0}$ and $\boldsymbol{X}_{I_c}$. As a result, (IR1) is easier to satisfy than the strong irrepresentable condition (6). Condition (IR2) are needed to guarantee selection consistency even in the usual regression cases. Conditions (IR3)-(IR4) are new conditions used to guarantee that the error in the factor analysis is negligible.

The validity of the revised irrepresentable conditions are discussed in the following theorem.

Theorem 3.1. *Suppose that (A1) to (A4) hold. Then, (IR1) to (IR4) holds with probability going to one.*

4. SIMULATION STUDIES

Table 1. Simulation results comparing LASSO and PRAIF method

			LASSO		$PRAIF_{lasso}$	
n	p	d	FN	FP	FN	FP
500	50	5	0.12	3.98	0	8.86
1000			0.16	4.05	0.03	4.48
1500			0.13	3.82	0.02	2.19
2000			0.08	4.21	0.01	1.65
500	100	10	0.48	10.89	0.09	7.67
1000			0.52	10.74	0.09	6.47
1500			0.50	11.11	0.08	2.37
2000			0.36	10.53	0.06	0.84
500	150	15	0.89	17.2	0.16	1.06
1000			1.02	17.35	0.27	1.85
1500			1.01	16.19	0.19	0.65
2000			0.93	16.22	0.22	0.23

Table 2. Simulation results comparing adaptive LASSO and PRAIF with adaptive LASSO method

			adaLASSO		$PRAIF_{ada}$	
n	p	d	FN	FP	FN	FP
500	50	5	0.73	0	0.15	0
1000			0.69	0	0.21	0
1500			0.57	0	0.21	0
2000			0.72	0	0.14	0
500	100	10	2.28	0	0.31	0
1000			2.2	0	0.44	0
1500			2.48	0	0.38	0
2000			2.34	0	0.39	0
500	150	15	4.08	0	0.59	0
1000			4.08	0	0.51	0
1500			4.06	0	0.67	0
2000			4.03	0	0.48	0

Table 3. Simulation results comparing PCA-LASSO and PRAIF method on different heteroscedasticity

			LC		PRAIF		LC		PRAIF	
			$s = 10$				$s = 20$			
n	p	d	FN	FP	FN	FP	FN	FP	FN	FP
500	50	5	0.05	2.67	0.02	4.59	0.13	2.56	0.05	1.09
1000			0.09	2.84	0.01	2.41	0.11	2.32	0.08	0.43
1500			0.06	2.75	0.01	0.7	0.09	2.25	0.09	0.22
2000			0.14	2.87	0.02	0.68	0.08	2.2	0.06	0.26
500	100	10	0.29	8.26	0.11	1.25	0.3	6.59	0.12	0.01
1000			0.37	7.41	0.09	0.79	0.35	6.65	0.23	0.18
1500			0.3	7.37	0.11	0.29	0.28	6.17	0.18	0.06
2000			0.32	7.44	0.1	0.29	0.27	6.28	0.16	0
500	150	15	0.72	12.12	0.18	0.44	0.74	11.07	0.28	0
1000			0.82	12.07	0.29	0.02	0.61	10.36	0.31	0
1500			0.65	12.08	0.17	0	0.6	10.14	0.26	0
2000			0.83	11.67	0.27	0	0.67	10.75	0.28	0

In this section, the finite-sample properties of the following methods are compared regarding the correct identification of relevant covariates,

1. LASSO: LASSO regression,
2. adaLASSO: adaptive LASSO regression,
3. LC: LASSO regression against covariates and principal components in [10],
4. ALC: adaptive LASSO regression against covariates and principal components,
5. PRAIF: PRAIF with LASSO, and
6. $PRAIF_{ada}$: PRAIF with adaptive LASSO.

In the simulation, the number of factors is chosen as $m = 2$. The d non-zero elements in $\boldsymbol{\gamma}$ is chosen at random. The sizes of the non-zero coefficients in $\boldsymbol{\gamma}$ are generated randomly from $Unif(0, 10)$. The non-zero entries of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$ are chosen independently from $Unif(2, 6)$ and $Unif(2, 2 + s)$. Here, s is used to describe the heteroscedasticity of the idiosyncratic factors. The upper triangle of $\boldsymbol{\Lambda}$ is first set to zero. Then, $\boldsymbol{\Lambda}$ is rotated so as to fulfill the model identification condition. The error variance $\sigma^2 = 1$ is chosen. The common factors $\boldsymbol{f}_t$ and idiosyncratic factors $\boldsymbol{\eta}_t$ are generated from Normal distributions $N(0, \boldsymbol{I}_m)$ and $N(0, \boldsymbol{\Phi})$ respectively. The covariates $\boldsymbol{x}_t$ are then generated from $\boldsymbol{f}_t$ and $\boldsymbol{\eta}_t$ using Equation (2).

All computer programs are implemented in R language. To estimate the coefficients, the R package named "parcor" [12] is used. Here, the optimal value of the tuning parameter λ is selected via 10-fold cross validation. For the methods involving "adaptive" LASSO, the LASSO counterparts are first obtained and the weights are set as the reciprocals of the LASSO estimators.

To evaluate the performances of different methods, the following measures are used. False negative refers to a deselected relevant covariate that is not chosen; similarly, false positive refers to a selected irrelevant covariate. The false positive rate (FP) and false negative rate (FN) are obtained from 100 replicates.

To compare the performance of $PRAIF$ and LASSO, consider the model with $\boldsymbol{b} = 0$. In this example, $\boldsymbol{b}^T \neq \boldsymbol{\gamma}^T \boldsymbol{\Lambda}$. Therefore, the usual linear regression model $y_t = \beta^T \boldsymbol{x}_t + \epsilon_t$ is misspecified. The simulation results of the two methods are shown in Table 1. It can be seen that PRAIF method tends to give smaller FN and FP than LASSO excepting the case of ($n = 500$, $p = 50$) and ($n = 1000$, $p = 50$). Table 2 shows the FN and FP of adaptive LASSO and $PRAIF_{ada}$. PRAIF method gives smaller FN than adaptive lasso and PRAIF method performs better as p and d increases. Comparing Tables 1 and 2, PARIF with adaptive LASSO method in general has smaller FP but larger FN values than PARIF method. FP are all zero in Table 2, so, the $PRAIF_{ada}$ performs better than $PRAIF$.

Table 3 compares LC and $PRAIF$ under different heteroscedasticity settings. The model with $\boldsymbol{b} = 0$ is considered. In both settings with heteroscedasticity $s = 10$ and $s = 20$, $PRAIF$ has smaller FN and FP than LC. The performance of $PRAIF_{ada}$ and ALC are shown in Table 4. $PRAIF_{ada}$ has smaller FN and FP in both $s = 10$ and $s = 20$ cases. Comparing Tables 3 and 4, $PARIF_{ada}$ has smaller FP but larger FN than $PARIF$.

To study the effects of dependence strengths between the relevant covariates and the irrelevant covariates, consider the communality ρ, that means the ratio between the contributions of the common factors and idiosyncratic factors to the variance of the covariates. In the simulation, $\boldsymbol{\Lambda}$ is generated as before and $\boldsymbol{\Phi}_{ii}$ are determined by $\boldsymbol{\Phi}_{ii} = (1 - \rho)^{-1} \rho \boldsymbol{\Lambda}_i^T \boldsymbol{\Lambda}_i$, where $\boldsymbol{\Lambda}_i$ is the i-th row of $\boldsymbol{\Lambda}$. The model with $\boldsymbol{b} = 0$ is considered. Table 5 compares ALC and PRAIF with Adaptive LASSO under different communality setting with $\rho = 0.1$ and $\rho = 0.35$. In both cases, $PRAIF_{ada}$ has smaller FN and FP than ALC.

To conclude, as shown in the above simulation results, $PRAIF_{ada}$ method outperforms other methods in general, particularly in the presence of strong heteroscedasticity and communality in the idiosyncratic factors.

Table 4. Simulation results comparing PCA-adaLASSO and PRAIF with Adaptive LASSO method on different heteroscedasticity

			ALC		$PRAIF_{ada}$		ALC		$PRAIF_{ada}$	
			$s = 10$				$s = 20$			
n	p	d	FN	FP	FN	FP	FN	FP	FN	FP
500	50	5	0.5	0	0.21	0	0.48	0	0.24	0
1000			0.41	0	0.17	0	0.48	0	0.17	0
1500			0.55	0	0.14	0	0.52	0	0.12	0
2000			0.64	0	0.22	0	0.5	0	0.2	0
500	100	10	1.68	0	0.3	0	1.47	0	0.62	0
1000			1.56	0	0.32	0	1.39	0	0.44	0
1500			1.57	0	0.39	0	1.35	0	0.41	0
2000			1.72	0	0.44	0	1.48	0	0.46	0
500	150	15	3.23	0	0.69	0	2.83	0	0.74	0
1000			3.17	0	0.66	0	2.58	0	0.88	0
1500			3.06	0	0.71	0	2.7	0	0.86	0
2000			3.35	0	0.89	0	2.65	0	0.81	0

Table 5. Simulation results comparing PCA-adaLASSO and PRAIF with Adaptive LASSO method on different communality

			ALC		$PRAIF_{ada}$		ALC		$PRAIF_{ada}$	
			$\rho = 0.1$				$\rho = 0.35$			
n	p	d	FN	FP	FN	FP	FN	FP	FN	FP
500	50	5	0.54	0	0.26	0	0.44	0	0.25	0
1000			0.67	0	0.23	0	0.42	0	0.20	0
1500			0.60	0	0.13	0	0.43	0	0.17	0
2000			0.57	0	0.19	0	0.37	0	0.25	0
500	100	10	2.29	0	0.42	0	1.47	0	0.31	0
1000			2.13	0	0.32	0	1.18	0	0.52	0
1500			2.10	0	0.43	0	1.29	0	0.47	0
2000			2.17	0	0.40	0	0.98	0	0.47	0
500	150	15	4.02	0	0.74	0	2.54	0	0.71	0
1000			3.87	0	0.55	0	2.37	0	0.64	0
1500			3.95	0	0.52	0	2.20	0	0.68	0
2000			3.91	0	0.52	0	2.21	0	0.77	0

5. EMPIRICAL DATA EXAMPLES

In this section, two real econometric data examples are studied, namely international economic input/output data and global stock index data. Both datasets are strongly affected by common systematic factors due to the globalization. The number of factors m is chosen so that the cumulative proportion of variance explained by the common systematic factors is higher than 97%.

5.1 International economic input/output data

It is common to study the relationship between the economic inputs and outputs (measured as gross domestic production, GDP) of a country via the Cobb-Douglas model. Due to the globalisation, all economies in the world become unprecedentedly closely tied to each other. As more regional cooperation organisations are found, and more economic cooperation agreements are signed, domestic economic output is increasingly influenced by both domestic and international economic inputs.

To study the international impacts on the domestic economy, we consider the data from the The World Bank website (http://www.worldbank.org/). The dataset contains the capital inputs, labor inputs, and nominal gross domestic production of 79 countries and regions over the period from 1990 to 2017. For each country or region, both GDP and the capital input are measured using the current prices (in millions of domestic currency) in each year. The labor input is measured regarding thousands of persons. Before analyzing the data, the monetary unit of both GDP and capital inputs are standardized to US dollars.

When only one country is studied, the Cobb-Douglas production function [4] widely used among economists is

$$Y = AK^{\alpha}L^{\beta}\,, \tag{7}$$

where Y is the GDP of the country, K is the capital input, and L is labor input. In addition, α and β are the unknown

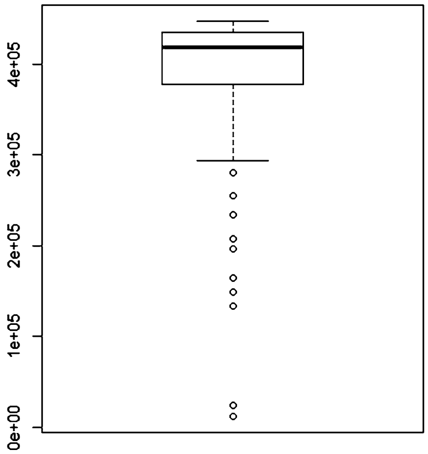

Figure 1. Boxplot of variance distribution of idiosyncratic factors in GDP dataset.

coefficients. A is a constant describing the technology of a country. Equivalently, the model can be written as

$$\log Y = \log A + \alpha \cdot \log K + \beta \cdot \log L \,. \tag{8}$$

To allow international impacts on the domestic economy, take $\log Y_0$ as the response and

$$(\log K_0, \log K_1, \cdots, \log K_n, \log L_0, \log L_1, \cdots, \log L_n)$$

as the covariates, where Y_0 is the GDP of a country, K_0 and L_0 are the capital and labour input of the country. $K_1, \cdots, K_n$ are the capital inputs of countries $1, \cdots, n$ and $L_1, \cdots, L_n$ are the labour inputs of countries $1, \cdots, n$.

In this example, Canada is the country labeled zero. The GDP of Canada is analyzed using the adaptive $LASSO$, ALC and $PRAIF_{ada}$ methods. The distribution variance of the idiosyncratic factors $diag(\mathbf{\Phi})$ is plotted in Figure 1. The box-plot shows heteroscedasticity.

To study the prediction performances, choose the observations of the first 20 years as the training dataset and that of the remaining 8 years as the testing dataset. The adaptive $LASSO$, ALC and $PRAIF_{ada}$ methods are applied to the training dataset and predictions are made on the GDP of Canada in the testing dataset. The performance is then evaluated via the relative root mean square error (RRMSE),

$$RRMSE = \sqrt{\frac{1}{7}\sum_{t=21}^{28}\left(\frac{\hat{y}_t - y_t}{y_t}\right)^2}\,, \tag{9}$$

where $\hat{y}_t$ is the predicted value and y_t is the realized value of $\log Y_0$ in the t-th Year. Table 6 shows that the RRMSE values of predicted and real GDP values by using adaptive $LASSO$, ALC and $PRAIF_{ada}$. From Table 6, $PRAIF_{ada}$ gives the smallest RRMSE among the three model selection methods.

Table 6. The average RRMSE comparing original adaptive LASSO, PCA-adaLASSO and $PRAIF$ with Adaptive LASSO method of global GDP dataset

	adaLasso	ALC	$PRAIF_{ada}$
RRMSE	2.684	0.0387	0.0191

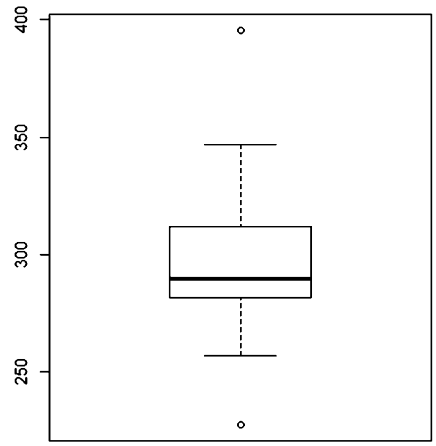

Figure 2. Box-plot of variance distribution of idiosyncratic factors in global technology services sector stock price dataset.

5.2 Stock return data in global technology services sector

The PRAIF method can be used to study the interactions between stocks in the financial market. Due to the systematic risk factors, the stock returns of all companies are correlated.

To demonstrate the benefits of using $PRAIF_{ada}$ in the financial data, we study the interactions between Intel Corporation and other 21 companies in the technology services sectors from 6 different countries. The data is obtained from Yahoo Finance website (https://finance.yahoo.com/), covering the period of 1296 trading days from 11/04/2014 to 11/04/2019. The stock prices are standardized so that the monetary unit is in US dollars. Figure 2 shows the heteroscedasticity in the idiosyncratic factors.

Take the rate of return of Intel company as the response and the rates of return of the remaining 21 companies as the covariates. Choose the observations of the first 1000 trading days as the training dataset and that of the remaining as the testing dataset. The adaptive $LASSO$, ALC and $PRAIF_{ada}$ methods are applied to the training dataset and predictions are made on the rate of return of Intel Corporation. Table 7 shows the RRMSE of the adaptive $LASSO$, ALC and $PRAIF_{ada}$ variable selection methods. $PRAIF_{ada}$ outperforms other two methods.

Table 7. The average RRMSE comparing original adaptive LASSO, PCA-adaLASSO and $PRAIF$ with Adaptive LASSO method of stock price dataset

	adaLasso	ALC	$PRAIF_{ada}$
RRMSE	2.879	3.657	1.459

6. CONCLUSION

As shown in Theorem 3.1, penalized regression against idiosyncratic factors allows selection consistency even under the factor model assumptions on the covariates. Comparing to the existing work of [10], the new results allow heteroscedasticity of the idiosyncratic factors. The theoretical results are well-supported from the simulation examples.

APPENDIX A. PROOFS OF MAIN RESULTS

Proof of Proposition 1. Let

$$sign(\boldsymbol{\alpha}_{I_0}) \quad = \quad \begin{pmatrix} 0 \\ sign(\boldsymbol{\gamma}_{I_0}) \end{pmatrix} . \tag{10}$$

Suppose that (IR2) holds. With a little bit abuse of notation, we write $\boldsymbol{\omega}_{I_0} sign(\boldsymbol{\alpha}_{I_0}) = (0, \omega_j sign(\gamma_j))_{j\in I_0}$. Choose

$$\hat{\boldsymbol{\alpha}}_{I_0} \quad = \quad (\hat{\boldsymbol{G}}_{I_0}{}^T\hat{\boldsymbol{G}}_{I_0})^{-1}(\hat{\boldsymbol{G}}_{I_0}^T\boldsymbol{Y} + n\lambda\boldsymbol{\omega}_{I_0} sign(\boldsymbol{\alpha}_{I_0}))\,. \tag{11}$$

Then, $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_{I_0}, 0)$ is a local solution if the following KKT (Karush-Kuhn-Tucker) conditions are satisfied at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$,

$$\frac{\boldsymbol{d}\left\|\boldsymbol{Y} - \hat{\boldsymbol{G}}_{I_0}\hat{\boldsymbol{\alpha}}_{I_0}\right\|_2^2}{\boldsymbol{d}\hat{\boldsymbol{\alpha}}_{I_0}} \quad = \quad n\lambda\boldsymbol{\omega}_{I_0} sign(\boldsymbol{\alpha}_{I_0})\,, \tag{12}$$

$$\left|\frac{\boldsymbol{d}\left\|\boldsymbol{Y} - \hat{\boldsymbol{G}}_{I_0}\hat{\boldsymbol{\alpha}}_{I_0}\right\|_2^2}{\boldsymbol{d}\hat{\boldsymbol{\alpha}}_j}\right| \quad \leq \quad n\lambda\omega_j\,, for\ all\ j \in I_c\,. \tag{13}$$

They are equivalent to

$$-\hat{\boldsymbol{G}}_{I_0}^T(\boldsymbol{Y} - \hat{\boldsymbol{G}}_{I_0}\hat{\boldsymbol{\alpha}}_{I_0}) + n\lambda\boldsymbol{\omega}_{I_0} sign(\boldsymbol{\alpha}_{I_0}) \quad = 0\,, \tag{14}$$

and

$$\left|-\hat{\boldsymbol{G}}_j^T(\boldsymbol{Y} - \hat{\boldsymbol{G}}_{I_0}\hat{\boldsymbol{\alpha}}_{I_0})\right| \quad \leq n\lambda\omega_j\,, for\ all\ j \in I_c\,. \tag{15}$$

(14) holds trivially under condition (IR2). Substituting (11) into (15) and rewriting $\boldsymbol{Y} = \boldsymbol{G}_{I_0}\boldsymbol{\alpha}_{I_0} + \epsilon$, inequality (15) becomes

$$\begin{aligned} n\lambda\omega_j \quad &\geq \quad \Big|-\hat{\boldsymbol{G}}_j^T\Big(\boldsymbol{G}_{I_0}\boldsymbol{\alpha}_{I_0} + \epsilon - \hat{\boldsymbol{G}}_{I_0}(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1} \\ &\qquad \Big(\hat{\boldsymbol{G}}_{I_0}^T(\boldsymbol{G}_{I_0}\boldsymbol{\alpha}_{I_0} + \epsilon) + n\lambda\boldsymbol{\omega}_{I_0} sign(\boldsymbol{\alpha}_{I_0})\Big)\Big)\Big| \\ &= \quad \Big|-\hat{\boldsymbol{G}}_j^T(\boldsymbol{I} - \boldsymbol{H})\epsilon - \hat{\boldsymbol{G}}_{I_c}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{G}_{I_0}\boldsymbol{\alpha}_{I_0} \\ &\qquad -n\lambda\hat{\boldsymbol{G}}_j^T\hat{\boldsymbol{G}}_{I_0}(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\boldsymbol{\omega}_{I_0} sign(\boldsymbol{\alpha}_{I_0})\Big|\,. \end{aligned} \tag{16}$$

Here, $\boldsymbol{\alpha_{I_0}}$ refer to the true values. (IR1), (IR3), and (IR4) guarantee that (16) holds. □

Proof of Theorem 3.1. By Proposition 1, it suffices to establish (IR1)-(IR4). □

Proof of (IR1). Consider

$$\begin{aligned} &\left\|\hat{\boldsymbol{G}}_{I_c}^T\hat{\boldsymbol{G}}_{I_0}(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1} sign(\boldsymbol{\alpha}_{I_0})\right\|_\infty \\ \leq \quad &(m+d)^{1/2}\left\|\left(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0}\right)^{-1}\right\|_2 \\ &\cdots\left\|\hat{\boldsymbol{G}}_{I_c}^T\hat{\boldsymbol{G}}_{I_0}\right\|_\infty \cdot \|sign(\boldsymbol{\alpha}_{I_0})\|_\infty\,. \end{aligned} \tag{17}$$

Note that from Condition (A4), $m + d = O(d)$. The term $\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0}$ can be handled using Lemma 5. The term $\hat{\boldsymbol{G}}_{I_c}^T\hat{\boldsymbol{G}}_{I_0}$ can be rewritten as

$$\begin{aligned} &\left\|\hat{\boldsymbol{G}}_{I_c}^T\boldsymbol{G}_{I_0}\right\|_\infty \\ = \quad &\left\|\left[\hat{\boldsymbol{E}}_{I_c}^T\hat{\boldsymbol{F}}, \hat{\boldsymbol{E}}_{I_c}^T\hat{\boldsymbol{E}}_{I_0}\right]\right\|_\infty \\ \leq \quad &\left\|\hat{\boldsymbol{E}}_{I_c}^T\hat{\boldsymbol{F}}\right\|_\infty + \left\|\hat{\boldsymbol{E}}_{I_c}^T\hat{\boldsymbol{E}}_{I_0}\right\|_\infty\,. \end{aligned} \tag{18}$$

The second term on the right-hand-side of (18) can be bounded using Lemma 1, 2, and 5,

$$\begin{aligned} &\left\|\hat{\boldsymbol{E}}_{I_c}^T\hat{\boldsymbol{E}}_{I_0}\right\|_\infty \\ \leq \quad &\left\|\left(\hat{\boldsymbol{E}}_{I_c} - \boldsymbol{E}_{I_c}\right)^T\right\|_2\left\|\left(\hat{\boldsymbol{E}}_{I_0} - \boldsymbol{E}_{I_0}\right)\right\|_2 \\ &+\left\|\left(\hat{\boldsymbol{E}}_{I_c} - \boldsymbol{E}_{I_c}\right)^T\right\|_2\|\boldsymbol{E}_{I_0}\|_\infty \\ &+\left\|\left(\hat{\boldsymbol{E}}_{I_0} - \boldsymbol{E}_{I_0}\right)^T\right\|_2\|\boldsymbol{E}_{I_c}\|_\infty + \left\|\boldsymbol{E}_{I_0}^T\boldsymbol{E}_{I_c}\right\|_\infty \\ \leq \quad &O_p\left(\max\left(p + n^{1/2}, (np)^{1/2} + np^{-1/2}\right)\right) \\ &+O_p\left(\max\left((n\log n)^{1/2}, (p\log n)^{1/2}\right)\right) \\ &+\sqrt{\left\|\hat{\boldsymbol{E}}_{I_0} - \boldsymbol{E}_{I_0}\right\|_2}\left(\max_{i\in I_c}\sqrt{\|\boldsymbol{E}_i\|_2}\right) \\ &+\left\|\boldsymbol{E}_{I_0}^T\boldsymbol{E}_{I_c}\right\|_\infty \\ = \quad &O_p\left((np)^{1/2} + np^{-1/2}\right)\,. \end{aligned} \tag{19}$$

Similar to (19), the first term on the right-hand-side of (18) can be bounded as

$$\begin{aligned} \left\|\hat{\boldsymbol{E}}_{I_c}^T\hat{\boldsymbol{F}}\right\|_\infty \quad \leq \quad &\left\|\left(\hat{\boldsymbol{E}_{I_c}} - \boldsymbol{E}_{I_c}\right)^T\right\|_2\left\|\left(\hat{\boldsymbol{F}} - \boldsymbol{F}\right)\right\|_2 \\ &+\left\|\left(\hat{\boldsymbol{E}}_{I_c} - \boldsymbol{E}_{I_c}\right)^T\right\|_2\|\boldsymbol{F}\|_\infty \end{aligned}$$

$$+\left\|\left(\hat{\boldsymbol{F}}-\boldsymbol{F}\right)^T\right\|_2\left\|\hat{\boldsymbol{E}}_{I_c}\right\|_\infty+\left\|\boldsymbol{E}_{I_c}^T\boldsymbol{F}\right\|_\infty$$

$$\leq\ O_p\left(np^{-1/2}\right). \tag{20}$$

From (19), (20) and Lemma 3,

$$\left\|\hat{\boldsymbol{G}}_{I_c}\hat{\boldsymbol{G}}_{I_0}^T(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}sign(\boldsymbol{\alpha}_{I_0})\right\|_\infty$$

$$=\ O_p\left(dp^{1/2}n^{-1/2}+dp^{-1/2}\right). \tag{21}$$

Under Condition $A1$ and $A2$ the irrepresentable condition (IR1) holds. □

Proof of (IR2). Define $sign(\alpha_{I_0})$ as in (10). It suffices to show that

$$\hat{\boldsymbol{\alpha}}_{I_0}=(\hat{\boldsymbol{b}},\hat{\boldsymbol{\gamma}}_{I_0})=(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\left(\hat{\boldsymbol{G}}_{I_0}^T\boldsymbol{Y}+n\lambda\boldsymbol{\omega}_{I_0}sign(\boldsymbol{\alpha}_{I_0})\right)$$

fulfills $sign(\hat{\boldsymbol{\alpha}}_{I_0})=sign(\boldsymbol{\alpha}_{I_0})$ with probability going to one, where α_{I_0} is the true value. Rewriting $\boldsymbol{Y}=\boldsymbol{G}_{I_0}\boldsymbol{\alpha}_{\boldsymbol{I_0}}+\epsilon$, we have

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_{I_0}&=(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\hat{\boldsymbol{G}}_{I_0}^T(\boldsymbol{G}_{I_0}\boldsymbol{\alpha}_{I_0}+\epsilon)\\&\quad+n\lambda(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\boldsymbol{\omega}_{I_0}sign(\boldsymbol{\alpha}_{I_0})\\&=\boldsymbol{\alpha}_{I_0}+(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\hat{\boldsymbol{G}}_{I_0}^T\epsilon\\&\quad+n\lambda(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\boldsymbol{\omega}_{I_0}sign(\boldsymbol{\alpha}_{I_0})\\&\quad+(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\hat{\boldsymbol{G}}_{I_0}(\hat{\boldsymbol{G}}_{I_0}-\boldsymbol{G}_{I_0})\boldsymbol{\alpha}_{I_0}.\end{aligned}\tag{22}$$

To guarantee that $sign(\hat{\boldsymbol{\alpha}}_{I_0})=sign(\boldsymbol{\alpha}_{I_0})$, the quantity $\min_{j\in I_0}|\alpha_j|$ must dominates

$$\begin{aligned}&A=\|\hat{\alpha}_{I_0}-\alpha_{I_0}\|_\infty\\&\ =\Big\|(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\hat{\boldsymbol{G}}_{I_0}^T\epsilon+n\lambda(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\boldsymbol{\omega}_{I_0}sign(\boldsymbol{\alpha}_{I_0})\\&\qquad+(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\hat{\boldsymbol{G}}_{I_0}^T(\hat{\boldsymbol{G}}_{I_0}-\boldsymbol{G}_{I_0})\boldsymbol{\alpha}_{I_0}\Big\|_\infty.\end{aligned}\tag{23}$$

The quantity A can further be bounded as

$$\begin{aligned}A&\leq\left\|(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\begin{pmatrix}\hat{\boldsymbol{E}}_{I_0}^T\epsilon\\\hat{\boldsymbol{F}}^T\epsilon\end{pmatrix}\right\|_\infty\\&\quad+\left\|n\lambda(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\boldsymbol{\omega}_{I_0}sign(\boldsymbol{\alpha}_{I_0})\right\|_\infty\\&\quad+\left\|(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\hat{\boldsymbol{G}}_{I_0}(\hat{\boldsymbol{G}}_{I_0}-\boldsymbol{G}_{I_0})\boldsymbol{\alpha}_{I_0}\right\|_\infty\\&=A_1+A_2+A_3.\end{aligned}$$

Using Lemmas 1 and 4, we have

$$\begin{aligned}\left\|\hat{\boldsymbol{E}}_{I_0}\epsilon\right\|_\infty&\leq\left\|(\hat{\boldsymbol{E}}_{I_0}-\boldsymbol{E}_{I_0})^T\epsilon\right\|_\infty+\left\|\boldsymbol{E}_{I_0}^T\epsilon\right\|_\infty\\&\leq\left\|(\hat{\boldsymbol{\Lambda}}_{I_0}-\boldsymbol{\Lambda}_{I_0})^T\boldsymbol{F}\epsilon+\hat{\boldsymbol{\Lambda}}_{I_0}(\boldsymbol{F}-\hat{\boldsymbol{F}})^T\epsilon\right\|_\infty\\&\quad+\left\|\boldsymbol{E}_{I_0}^T\epsilon\right\|_\infty\\&\leq\ O_p(p^{-1/2}(n\log n)^{1/2})\end{aligned}\tag{24}$$

and

$$\begin{aligned}\left\|\hat{\boldsymbol{F}}^T\epsilon\right\|_\infty&\leq\left\|\boldsymbol{F}^T\epsilon\right\|_\infty+\left\|(\hat{\boldsymbol{F}}-\boldsymbol{F})^T\epsilon\right\|_\infty\\&\leq\ O_p((2n\log p)^{1/2}).\end{aligned}\tag{25}$$

Then, $A_1=O_p(n^{-1/2}p^{-1/2}(\log n)^{1/2})$. Lemma 4 suggests that $A_2=O_p(\lambda d)$. The last term in (23) can be bounded as

$$\begin{aligned}A_3&=\left\|(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\hat{\boldsymbol{G}}_{I_0}(\hat{\boldsymbol{G}}_{I_0}-\boldsymbol{G}_{I_0})\boldsymbol{\alpha}_{I_0}\right\|_\infty\\&\leq\left\|(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\right\|_\infty\cdot\\&\quad\left\|(\hat{\boldsymbol{G}}_{I_0}-\boldsymbol{G}_{I_0})^T(\hat{\boldsymbol{G}}_{I_0}-\boldsymbol{G}_{I_0})\boldsymbol{\alpha}_{I_0}\right\|_\infty\\&\quad+\left\|(\hat{\boldsymbol{G}}_{I_0}^T\hat{\boldsymbol{G}}_{I_0})^{-1}\right\|_\infty\left\|(\hat{\boldsymbol{G}}_{I_0}-\boldsymbol{G}_{I_0})^T\boldsymbol{G}_{I_0}\boldsymbol{\alpha}_{I_0}\right\|_\infty\\&\leq\ O_p(dp^{-1}).\end{aligned}\tag{26}$$

Under Condition $A2$ and $A3$, results (24) to (26) suggest that $sign(\hat{\boldsymbol{\alpha}}_{I_0})=sign(\boldsymbol{\alpha}_{I_0})$. □

Proof of (IR3) and (IR4). Note that H is idempotent and the eigenvalues are either 0 or 1 with rank $d+m$. Then, H can also be rewritten using orthogonal projections as $\boldsymbol{P}\boldsymbol{P}^T$, where $\boldsymbol{P}$ is $n\times(d+m)$ matrix of orthogonal vectors. By Condition $A1$ to $A4$, *Lemma 2* and *Lemma 3*

$$\begin{aligned}&\left\|-\hat{\boldsymbol{G}}_{I_c}^T(\boldsymbol{I}-\boldsymbol{H})\epsilon\right\|_\infty\\&=\Big\|-\boldsymbol{G}_{I_c}^T(\boldsymbol{I}-\boldsymbol{H})\epsilon-(\boldsymbol{G}_{I_c}-\hat{\boldsymbol{G}}_{I_c})^T\boldsymbol{H}\epsilon+\boldsymbol{G}_{I_c}^T\epsilon\\&\quad+(\boldsymbol{G}_{I_c}-\hat{\boldsymbol{G}}_{I_c})\epsilon\Big\|_\infty\\&\leq\left\|\boldsymbol{G}_{I_c}^T\boldsymbol{P}\boldsymbol{P}^T\epsilon\right\|_\infty+\left\|(\boldsymbol{G}_{I_c}-\hat{\boldsymbol{G}}_{I_c})^T\boldsymbol{P}\boldsymbol{P}^T\epsilon\right\|_\infty\\&\quad+\left\|(\boldsymbol{G}_{I_c}-\hat{\boldsymbol{G}}_{I_c})^T\epsilon\right\|_\infty+\left\|\boldsymbol{G}_{I_c}^T\epsilon\right\|_\infty\\&\leq\left\|\boldsymbol{G}_{I_c}^T\boldsymbol{P}\right\|_\infty\left\|\boldsymbol{P}^T\epsilon\right\|_\infty+\left\|(\boldsymbol{G}_{I_c}-\hat{\boldsymbol{G}}_{I_c})^T\boldsymbol{P}\right\|_\infty\left\|\boldsymbol{P}^T\epsilon\right\|_\infty\\&\quad+\left\|(\boldsymbol{E}_{I_c}-\hat{\boldsymbol{E}}_{I_c})^T\epsilon\right\|_\infty+O_p\left((n\log(p-d))^{1/2}\right)\\&=O_p\left(\max((p\log d)^{1/2},(n\log d)^{1/2})\right)\\&\quad+O_p\left(d^{1/2}d\log(p+d))^{1/2}\right)\\&\quad+O_p\left(\max((p\log n)^{1/2},(n\log n)^{1/2})\right)\\&\quad+O_p\left((n\log(p-d))^{1/2}\right)\\&=O_p\left((n\log n)^{1/2}\right)\end{aligned}\tag{27}$$

and

$$\left\|-\hat{\boldsymbol{G}}_{I_c}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{G}_{I_0}\boldsymbol{\alpha}_{I_0}\right\|_\infty$$

$$
\begin{aligned}
&\leq \left\|-\hat{\boldsymbol{G}}_{I_c}^T \boldsymbol{P}\right\|_\infty \cdot \left\|\boldsymbol{P}^T \boldsymbol{G}_{I_0}\boldsymbol{\alpha}_{I_0}\right\|_\infty + \left\|\hat{\boldsymbol{G}}_{I_c}^T \boldsymbol{G}_{I_0}\boldsymbol{\alpha}_{I_0}\right\|_\infty \\
&\leq \left(\left\|(\boldsymbol{G}_{I_c} - \hat{\boldsymbol{G}}_{I_c})^T \boldsymbol{P}\right\|_\infty + \left\|\boldsymbol{G}_{I_c}^T \boldsymbol{P}\boldsymbol{\alpha}_{I_0}\right\|_\infty\right) \\
&\quad + \left(\left\|(\hat{\boldsymbol{G}}_{I_c} - \boldsymbol{G}_{I_c})^T \boldsymbol{G}_{I_0}\right\|_\infty + \left\|\boldsymbol{G}_{I_c}^T \boldsymbol{G}_{I_0}\right\|_\infty\right) \|\alpha\|_\infty \\
&\leq O_p(\max(p^{1/2}, n^{1/2})) \\
&\quad + O_p((2n\log d)^{1/2}) + \left(\left\|(\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}})\right\|_\infty\right. \\
&\quad \left.\left\|\boldsymbol{F}^T \boldsymbol{G}_{I_0}\right\|_\infty + \left\|\hat{\boldsymbol{\Lambda}}\right\|_\infty \left\|(\boldsymbol{F} - \hat{\boldsymbol{F}})\boldsymbol{G}_{I_0}\right\|_\infty\right) \|\boldsymbol{\alpha}_{I_0}\|_\infty \\
&\leq O_p((2n\log d)^{1/2}) + \left(\left\|(\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}})\right\|_\infty \left\|\boldsymbol{F}^T \boldsymbol{G}_{I_0}\right\|_\infty\right) \cdot \\
&\quad \|\boldsymbol{\alpha}_{I_0}\|_\infty + \left\|\hat{\boldsymbol{\Lambda}}\right\|_\infty \left\|(\boldsymbol{F} - \hat{\boldsymbol{F}})\boldsymbol{G}_{I_0}\right\|_\infty . \qquad (28)
\end{aligned}
$$

The term $\left\|(\boldsymbol{F} - \hat{\boldsymbol{F}})\boldsymbol{G}_{I_0}\right\|_\infty$ in (28) can be expanded as

$$
\begin{aligned}
&\left\|(\boldsymbol{F} - \hat{\boldsymbol{F}})\boldsymbol{G}_{I_0}\right\|_\infty \\
\leq &\left\|(\left(\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right)^{-1} \hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \left(\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}}\right)^T)\right\|_\infty \cdot \\
&\left\|\boldsymbol{F}^T \boldsymbol{G}_{I_0}\right\|_\infty \\
&+ \left\|\left(\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right)^{-1}\right\|_\infty \left\|\hat{\boldsymbol{\Lambda}}_{I_c}^T \hat{\boldsymbol{\Phi}}_{I_c}^{-1} \boldsymbol{E}_{I_c}^T \boldsymbol{G}_{I_0}\right\|_\infty \\
&+ \left\|\left(\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right)^{-1}\right\|_\infty \left\|\hat{\boldsymbol{\Lambda}}_{I_0}^T \hat{\boldsymbol{\Phi}}_{I_0}^{-1} \boldsymbol{E}_{I_0}^T \boldsymbol{E}_{I_0}\right\|_\infty \\
&+ \left\|\left(\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right)^{-1}\right\|_\infty \left\|\hat{\boldsymbol{\Lambda}}_{I_0}^T \hat{\boldsymbol{\Phi}}_{I_0}^{-1} \boldsymbol{E}_{I_0}^T \boldsymbol{F}\right\|_\infty \\
&\leq O_p(p^{-1} n (\log(p))^{1/2}) . \qquad (29)
\end{aligned}
$$

Substituting (29) into (28) yields

$$
\begin{aligned}
&\left\|-\hat{\boldsymbol{G}}_{I_c}^T (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{G}_{I_0}\boldsymbol{\alpha}_{I_0}\right\|_\infty \\
\leq\ & O_p((2n\log d)^{1/2}) \\
&+ O_p\left(p^{-1} n (\log(p))^{1/2}\right) \\
&+ \left(\left\|(\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}})\right\|_\infty \left\|\boldsymbol{F}^T \boldsymbol{G}_{I_0}\right\|_\infty\right) \|\boldsymbol{\alpha}_{I_0}\|_\infty \\
\leq\ & O_p\left((2n\log d)^{1/2}\right) . \qquad (30)
\end{aligned}
$$

This completes the proof. □

APPENDIX B. TECHNICAL LEMMAS

Lemma 1. *(See Theorem 5.1 and Equation (A.8) of Bai and Li [1].) We have*

$$
\left\|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\right\|_2 = \left(O_p\left(n^{-1/2} p^{1/2}\right)\right) \quad \textit{and} \quad \left(\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}} \hat{\boldsymbol{\Lambda}}\right)^{-1} = O_p(p^{-1}) .
$$

Lemma 2. *(See Chow and Teugal, 1978.) Assume that $A_1, A_2, \cdots, A_n$ are independent $N(0,1)$ random variables and $A = (A_1, A_2, \cdots, A_n)$. Then,*

$$
\|\boldsymbol{A}\|_\infty = O_p(\log(n)^{1/2}) . \qquad (31)
$$

Lemma 3. *(See Ng and Lee (2016) [18].) Assume that $\boldsymbol{B} = (b_{ij})$ is $n \times p$ matrix of independent $N(0,1)$ random variables. $\boldsymbol{v} = (v_t)_{t=1,2,\cdots,n}$ is a vector of independent and identically distributed with mean zero and variance one. Then,*

$$
\left\|\boldsymbol{B}^T v\right\|_\infty = O_p((2n\log p)^{1/2}) . \qquad (32)
$$

Lemma 4. *The following holds,*

(a) $\|\boldsymbol{E}\|_2 = O_p(n^{1/2})$,
(b) $\left\|\hat{\boldsymbol{F}} - \boldsymbol{F}\right\|_2 = O_p(n^{1/2} p^{-1/2})$,
(c) $\left\|\hat{\boldsymbol{E}}_{I_0} - \boldsymbol{E}_{I_0}\right\|_2 = O_p(p^{1/2}) + O_p(n^{1/2} p^{-1/2})$,
(d) $\left\|\hat{\boldsymbol{E}}_{I_c} - \boldsymbol{E}_{I_c}\right\|_2 = O_p\left(\max(n^{1/2}, p^{1/2})\right)$,
(e) $\left\|\hat{\boldsymbol{G}}_{I_0} - \boldsymbol{G}_{I_0}\right\|_2 = O_p(p^{1/2}) + O_p(n^{1/2} p^{-1/2})$,
(f) $\left\|\hat{\boldsymbol{G}}_{I_c} - \boldsymbol{G}_{I_c}\right\|_2 = O_p\left(\max(n^{1/2}, p^{1/2})\right)$.

Proof. Result (a) is a direct consequence of random matrix theory, see [2]. The bound of (b) can be obtained by Lemma 1 and Cauchy-Schwartz inequality as follows,

$$
\begin{aligned}
&\left\|\hat{\boldsymbol{F}} - \boldsymbol{F}\right\|_2 \\
= &\left\|\left(\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right)^{-1} \hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \left(\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}}\right)^T \boldsymbol{F} + \right. \\
&\left.\left(\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right)^{-1} \hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \boldsymbol{E}\right\|_2 \\
= &\left\|\left(\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right)^{-1} \hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \left(\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}}\right)^T \boldsymbol{F}\right. \\
&\left. + \left(\frac{1}{p}\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right)^{-1} \frac{1}{p}\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \boldsymbol{E}\right\|_2 \\
\leq &\left\|\left(\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right)^{-1}\right\|_2 \cdot \\
&\sqrt{\left\|\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right\|_2 \left\|\left(\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}}\right)^T \hat{\boldsymbol{\Phi}}^{-1} \left(\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}}\right)\right\|_2} \cdot \\
&\|\boldsymbol{F}\|_2 + \left\|\left(\frac{1}{p}\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right)^{-1}\right\|_2 \left\|\frac{1}{p}\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \boldsymbol{E}\right\|_2 \\
\leq\ & O_p(n^{1/2} p^{-1/2}) \qquad (33)
\end{aligned}
$$

Result (d) can be shown from results *(a)* and *(b)*,

$$
\begin{aligned}
\left\|\hat{\boldsymbol{E}_{I_c}} - \boldsymbol{E}_{I_c}\right\|_2 &\leq \left\|\hat{\boldsymbol{E}} - \boldsymbol{E}\right\|_2 \\
&= \left\|\left(\boldsymbol{\Lambda} - \hat{\boldsymbol{\Lambda}}\right)^T \boldsymbol{F} - \hat{\boldsymbol{\Lambda}}\left(\boldsymbol{F} - \hat{\boldsymbol{F}}\right)\right\|_2
\end{aligned}
$$

$$\leq O_p\left(\max(p^{1/2}, n^{1/2})\right) . \tag{34}$$

Result (c) can be obtained similarly. The bounds of *(e)* and *(f)* can be obtained from *(b)*, *(c)*, and *(d)*,

$$\begin{aligned} \left\|\hat{\boldsymbol{G}}_{I_0} - \boldsymbol{G}_{I_0}\right\|_2 &\leq \left\|\hat{\boldsymbol{E}}_{I_0} - \boldsymbol{E}_{I_0}\right\|_2 + \left\|\hat{\boldsymbol{F}} - \boldsymbol{F}\right\|_2 \\ &\leq O_p(n^{1/2}p^{-1/2}) \end{aligned} \tag{35}$$

$$\begin{aligned} \left\|\hat{\boldsymbol{G}}_{I_c} - \boldsymbol{G}_{I_c}\right\|_2 &\leq \left\|\hat{\boldsymbol{E}}_{I_c} - \boldsymbol{E}_{I_c}\right\|_2 + \left\|\hat{\boldsymbol{F}} - \boldsymbol{F}\right\|_2 \\ &\leq O_p\left(\max(p^{1/2}, n^{1/2})\right) . \end{aligned} \tag{36}$$

□

Lemma 5. *Let* $\boldsymbol{S} = \boldsymbol{G}_{I_0}^T \boldsymbol{G}_{I_0}$ *and* $\hat{\boldsymbol{S}} = \hat{\boldsymbol{G}}_{I_0}^T \hat{\boldsymbol{G}}_{I_0}$. *Then,*

$$\left\|\boldsymbol{S}^{-1}\right\|_2 = O_p(n^{-1}) \quad \text{and} \quad \left\|\hat{\boldsymbol{S}}^{-1}\right\|_2 = O_p(n^{-1}) .$$

Proof. For sufficiently large n, choose a subset I_1 of size zn that includes I_0, where z is an arbitrarily chosen constant between 0 and 1. Using the results of limits of extreme eigenvalues in [2], it can be shown that

$$\frac{1}{n}\lambda_{min}\left(\boldsymbol{G}_{I_0}^T \boldsymbol{G}_{I_0}\right) \geq \frac{1}{n}\lambda_{min}\left(\boldsymbol{G}_{I_1}^T \boldsymbol{G}_{I_1}\right) \xrightarrow{a.s.} \left(1 - \sqrt{z}\right) .$$

Then,

$$\left\|\left(\boldsymbol{G}_{I_0}^T \boldsymbol{G}_{I_0}\right)^{-1}\right\|_2 = \frac{1}{\lambda_{min}\left(\boldsymbol{G}_{I_0}^T \boldsymbol{G}_{I_0}\right)} \xrightarrow{a.s.} \frac{1}{n\left(1 - \sqrt{z}\right)}$$

which is $O_p\left(n^{-1}\right)$. Consider

$$\begin{aligned} \left\|\hat{\boldsymbol{S}}^{-1}\right\|_2 &= \left\|\hat{\boldsymbol{S}}^{-1} - \boldsymbol{S}^{-1} + \boldsymbol{S}^{-1}\right\|_2 \\ &= \left\|\hat{\boldsymbol{S}}^{-1} - \boldsymbol{S}^{-1}\right\|_2 + \left\|\boldsymbol{S}^{-1}\right\|_2 \\ &= \left\|\hat{\boldsymbol{S}}^{-1}(\hat{\boldsymbol{S}} - \boldsymbol{S})\boldsymbol{S}^{-1}\right\|_2 + \left\|\boldsymbol{S}^{-1}\right\|_2 . \end{aligned} \tag{37}$$

Using Lemma 4 and Cauchy-Schwarz inequality, the terms $\hat{\boldsymbol{S}} - \boldsymbol{S}$ and $\hat{\boldsymbol{S}}^{-1}$ on the right-hand-side of (37) can be bounded as

$$\begin{aligned} &\left\|\hat{\boldsymbol{S}} - \boldsymbol{S}\right\|_2 \\ = &\left\|\hat{\boldsymbol{G}}_{I_0}^T \hat{\boldsymbol{G}}_{I_0} - \boldsymbol{G}_{I_0}^T \boldsymbol{G}_{I_0}\right\|_2 \\ = &\left\|(\hat{\boldsymbol{G}}_{I_0} - \boldsymbol{G}_{I_0})^T \boldsymbol{G}_{I_0} + (\hat{\boldsymbol{G}}_{I_0} - \boldsymbol{G}_{I_0})^T (\hat{\boldsymbol{G}}_{I_0} - \boldsymbol{G}_{I_0}) \right. \\ &\left. + \boldsymbol{G}_{I_0}^T (\hat{\boldsymbol{G}}_{I_0} - \boldsymbol{G}_{I_0})\right\|_2 \\ = &\ O_p(np^{-1/2}) \end{aligned} \tag{38}$$

and

$$\begin{aligned} \left\|\hat{\boldsymbol{S}}^{-1}\right\|_2 &= \frac{1}{\lambda_{min}(\hat{\boldsymbol{S}} - \boldsymbol{S} + \boldsymbol{S})} \\ &\leq \frac{1}{\lambda_{min}(\hat{\boldsymbol{S}} - \boldsymbol{S}) + \lambda_{min}(\boldsymbol{S})} \\ &\leq \frac{1}{\lambda_{min}(\boldsymbol{S}) - \lambda_{max}(\hat{\boldsymbol{S}} - \boldsymbol{S})} \\ &\leq \frac{1}{\lambda_{min}(\boldsymbol{S}) - \left\|\hat{\boldsymbol{S}} - \boldsymbol{S}\right\|_2} \\ &\leq O_p(n^{-1}) . \end{aligned} \tag{39}$$

Here, we have used the fact that $\lambda_{min}(\boldsymbol{S})$ dominates $\lambda_{max}(\hat{\boldsymbol{S}} - \boldsymbol{S})$. The lemma then follows immediately. □

APPENDIX C. EM ALGORITHM FOR FACTOR ANALYSIS

The maximum likelihood estimation of the factor model can be implemented via the EM algorithm proposed by Bai and Li [1]. Let $\boldsymbol{\theta} = (\Lambda, \Phi)$. For integer k, denote by $\boldsymbol{\theta}^{(k)}$ the estimation obtained in the k-th iteration. Define $\boldsymbol{M}_{\boldsymbol{xx}} = \frac{1}{n-1}\sum_{t=1}^n (\boldsymbol{x}_t - \bar{\boldsymbol{x}})(\boldsymbol{x}_t - \bar{\boldsymbol{x}})^T$ and $\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Phi}$, where

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{t=1}^n \boldsymbol{x}_t .$$

Step 1. (Initial guess) Construct $\boldsymbol{\Lambda}^{(0)}$ using the first m eigenvectors of $\boldsymbol{M}_{\boldsymbol{xx}}$. Suitable rotation is applied for model identification purpose. Then, compute $diag(\Phi^{(0)}) = diag(\boldsymbol{M}_{\boldsymbol{xx}} - (\boldsymbol{\Lambda}^{(0)})(\boldsymbol{\Lambda}^{(0)})^T)$.

Step 2. (Expectation-Maximization step)

The EM algorithm updates the unknown estimator according to

$$\begin{aligned} \hat{\boldsymbol{\Lambda}}^{(k+1)} &= \left[\frac{1}{n}\sum_{t=1}^n E\left(\boldsymbol{x}_t \boldsymbol{f}_t^T | \boldsymbol{x}_t, \boldsymbol{\theta}^k\right)\right] \times \\ &\quad \left[\frac{1}{n}\sum_{t=1}^n E\left(\boldsymbol{f}_t \boldsymbol{f}_t^T | \boldsymbol{x}_t, \boldsymbol{\theta}^k\right)\right]^{-1} , \\ \hat{\boldsymbol{\Phi}}^{(k+1)} &= diag\left(\boldsymbol{M}_{\boldsymbol{xx}} - (\hat{\boldsymbol{\Lambda}}^k)(\hat{\boldsymbol{\Lambda}}^k)^T (\boldsymbol{\Omega}^k)^{-1} \boldsymbol{M}_{\boldsymbol{xx}}\right) , \end{aligned}$$

where

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^n E\left(\boldsymbol{x}_t \boldsymbol{f}_t^T | \boldsymbol{x}_t, \boldsymbol{\theta}\right) &= \boldsymbol{\Lambda}^T \boldsymbol{\Omega}^{-1} \boldsymbol{M}_{\boldsymbol{xx}} \boldsymbol{\Omega}^{-1} \boldsymbol{\Lambda} + \boldsymbol{I}_m \\ &\quad - \boldsymbol{\Lambda}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\Lambda} , \\ \frac{1}{n}\sum_{t=1}^n E\left(\boldsymbol{f}_t \boldsymbol{f}_t^T | \boldsymbol{x}_t, \boldsymbol{\theta}\right) &= \boldsymbol{M}_{\boldsymbol{xx}} \boldsymbol{\Omega}^{-1} \boldsymbol{\Lambda} . \end{aligned}$$

Step 3. Repeat Step 2 until coverages.

Step 4. The common factor $\hat{\boldsymbol{f}}_t$ and the idiosyncratic factors $\hat{\boldsymbol{\eta}}_t$ are then estimated as

$$\hat{\boldsymbol{f}}_t = E\left(\boldsymbol{f}_t | \boldsymbol{x}_t, \hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Phi}}\right)$$

$$
\begin{aligned}
&= \left(\hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} \hat{\boldsymbol{\Lambda}}\right)^{-1} \hat{\boldsymbol{\Lambda}}^T \hat{\boldsymbol{\Phi}}^{-1} (\boldsymbol{x}_t - \bar{\boldsymbol{x}}) \,, \\
\hat{\boldsymbol{\eta}}_t &= \boldsymbol{x}_t - \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{f}}_t - \bar{\boldsymbol{x}} \,.
\end{aligned}
$$

ACKNOWLEDGMENTS

Chi Tim, Ng's work is supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2017R1C1B2011652) and a research program of Rural Development Administration, Republic of Korea (Project No. PJ01283009).

Received 28 December 2018

REFERENCES

[1] Bai J., Li K. (2012) Statistical analysis of factor models of high dimension, *The Annals of Statistics*, **40** 436–465.

[2] Bai Z., Silverstein J. W. (2005) *Random Matrix Theory*, Science Press. MR0571177

[3] Chow T. L., Teugels J. L. (1978) The sum and the maximum of i.i.d. random variables, In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, North Holland, N.Y.

[4] Cobb, C. W., Douglas, P. H. (1928) A theory of production, *American Economic Review*, **18** 139–165. MR1946581

[5] Fan J., Li R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist.*, **96** 74–99. MR2640659

[6] Fan J., Lv J. (2012) A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, **20** 101–148.

[7] Chetverikov, Z., Liao, Z., and Chernozhukov, V. (2016) On cross-validated Lasso, *ArXiv e-prints*. MR2166557

[8] Hunter D. R. and Li R. (2005) Variable selection using MM algorithms, *The Annals of Statistics*, **33** 1617–1642. MR2682632

[9] Jia J. and Yu B. (2010) On model selection consistency of the elastic net when $p \gg n$, *Statistics Sinica*, **20** 595–611. MR2906873

[10] Kneip, A. and Sarda, P. (2011) Factor models and variable selection in high-dimensional regression analysis, *Annals of Statistics*, **39** 2410–2447 MR1805787

[11] Knight K., Fu W. (2000) Asymptotics for LASSO-type estimators, *The Annals of Statistics*, **28** 1356–1378.

[12] Kraemer N. and Schaefer J. and Boulesteix L. (2014) Regularized estimation of large-scale gene regulatory networks using gaussian graphical models, *BMC Bioinformatics*, **10**, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-384. MR3163832

[13] Lee Y., Oh H. (2014) A new sparse variable selection via random-effect model, *Journal of Multivariate Analysis*, **125** 89–99. MR3015038

[14] Loh P., Wainwright M. J. (2012) High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity, *Annals of Statistics*, **40** 1637–1664. MR3335800

[15] Loh P., Wainwright M. J. Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. http://arxiv.org/abs/1305.2436. MR2758523

[16] Meinshausen N., Bühlmann P. (2010) Stability selection. *Journal of the Royal Statistical Society: Series B*, **72** 417–473. MR3474502

[17] Ma C., Huang J. (2016) Asympotic properties of Lasso in high-dimensional partially linear models, *Science China*, **59** 769–788. MR3529723

[18] Ng C. T., Oh S., Lee Y. (2016) Going beyond oracle property: selection consistency and uniqueness of local solution of the generalized linear model, *Statistical Methodology*, **32** 147–160. MR1744773

[19] Pearl J. (2000) *Causality: Models, Reasoning, and Inference*, Cambridge University Press. MR1379242

[20] Tibshirani R. (1996) Regression shrinkage and selection via lasso, *Journal of the Royal Statistical Society*, **58** 267–288. MR2749913

[21] Wang H. S., Li B., Leng C. L. (2007) Shrinkage tuning parameter selection with a diverging number of parameters, *Journal of the Royal Statistical Society*, **71** 671–683. MR2604701

[22] Zhang C. H. (2010) Nearly unbiased variable selection under minimax concave penalty, *Annals of Statistics*, **38**(2) 894–942. MR2274449

[23] Zhao P., Yu B. (2006) On model selection consistency of Lasso, *Journal of Machine Learning Research*, **7** 2541–2563. MR2279469

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

Kaimeng Zhang
Department of Statistics
Chonnam National University
Gwangju, 61186
Korea
E-mail address: zhangkm215@gmail.com

Chi Tim Ng
Department of Statistics
Chonnam National University
Gwangju, 61186
Korea
E-mail address: easterlyng@gmail.com