Variable Selection in Linear Regressions with Many Highly Correlated Covariates *

Mahrad Sharifvaghefi †

University of Pittsburgh

June 15, 2021

Abstract

This paper is concerned with variable selection in linear high-dimensional framework when the set of covariates under consideration are highly correlated. Existing methods in the literature require that the collinearity among covariates to be weak, yet, often in applied research, covariates could be strongly cross correlated due to common factors. This paper generalizes the One Covariate at a Time Multiple Testing procedure proposed by Chudik et al. (2018) to allow the set of covariates under consideration to be highly correlated. We exploit ideas from latent factor and multiple testing literature to control the probability of selecting the approximating model. The proposed method is shown to be valid under general assumptions and is computationally very fast. Monte Carlo experiments indicate that the newly suggested method have appealing finite-sample performance relative to competing methods, like LASSO, under many different settings. The benefits of the proposed method are also illustrated by an empirical application to selection of risk factors in asset pricing literature.

Keywords: High-dimensionality, Multiple Testing, Variable Selection, Latent Factor Structure, Strong Cross-sectional Dependence

JEL Classifications: C38, C52, C55

*I am indebted to Hashem Pesaran for his invaluable guidance and support. I am also grateful to Yingying Fan, Cheng Hsiao, Jinchi Lv, Marcelo Moreira, and Geert Ridder for their constructive comments. This paper is also benefited from helpful comments by Alex Chudik, Simon Reese, Yoshimasa Uematsu, and seminar participants at University of Southern California.

†4927 Wesley W. Posvar Hall, 230 S Bouquet St., Pittsburgh, PA 15260. Email: sharifvaghefi@pitt.edu.

1 Introduction

Researchers are often interested in selecting a small number of important variables from a large set of covariates, known as the *active set*. Thanks to recent advances in data processing and computing, there has been a significant increase in the data and information available for research over the past decade. However, these large datasets make the model selection problem considerably more difficult. On one hand, the set of possible specifications rises exponentially with the number of covariates in the active set and hence classical model selection criteria such as **A**kaike **I**nformation **C**riterion (AIC) or **B**ayesian **I**nformation **C**riterion (BIC) become impossible to implement. On the other hand, existing variable selection methods in linear high-dimensional settings require the degree of correlations across the covariates in the active set to be sufficiently weak. Yet, often in practice, especially in macroeconomics and finance, the covariates in the active set are strongly correlated. This paper contributes to this literature by proposing a procedure for variable selection when collinearity among the set of covariates under consideration is high and standard penalized regression techniques do not apply.

Before presenting our procedure, we provide a brief summary of a growing body of research on linear high dimensional settings that mostly exploits the penalized regression framework[1]. In this framework the vector of regression coefficients, $\boldsymbol{\beta}$, of a regression of y_t on $\mathbf{x}_{nt} = (x_{1t}, x_{2t}, \cdots, x_{nt})'$, is estimated by $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\{\sum_{t=1}^{T}(y_t - \mathbf{x}_{nt}'\boldsymbol{\beta})^2 + P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})\}$, where $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ is a penalty function that penalizes $\boldsymbol{\beta}$, and $\boldsymbol{\lambda}$ is a vector of tuning parameters to be set by the researcher. Setting $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ proportional to the ℓ_1 norm of $\boldsymbol{\beta}$ yields the famous **L**east **A**bsolute **S**hrinkage and **S**election **O**perator (LASSO) proposed by Tibshirani (1996). Other forms of $P_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$ include the ℓ_q norm of $\boldsymbol{\beta}$ for $0 \leq q \leq 2$ [for example see Fan and Li (2001), Zou and Hastie (2005), Zhang et al. (2010), and Belloni et al. (2012)]. Recently, Fan and Lv (2013) show that the estimation errors and the prediction loss of the ℓ_1-regularization method of LASSO and the concave ones are asymptotically equivalent. Despite considerable progress made in the theory and practice of penalized regression, open questions remain, including the choice of the penalty function and tuning parameters. To avoid some of these issues, Chudik et al. (2018) propose an alternative method to penalized regression procedures called **O**ne **C**ovariate at a Time **M**ultiple **T**esting (OCMT). The authors establish that under general assumptions the suggested procedure asymptotically selects all the relevant covariates and none of the irrelevant ones. Their results additionally show that the estimation errors of coefficients and prediction loss converge to zero. Finally, their Monte Carlo studies show that the suggested method performed

[1] A number of procedures introduced in machine learning literature such as boosting, regression trees, and stepwise regression are also commonly used as an alternative to penalized regression. See, for example, Friedman et al. (2000), Friedman (2001), Buhlmann (2006) and Fan and Lv (2008).

better than penalized regression or boosting procedures under various designs.

As it mention earlier, all these suggested procedures require the collinearity among covariates in the active set to be weak. In particular, Zhao and Yu (2006) show that the Irrepresentable Condition is required for LASSO to asymptotically select only the covariates with non-zero marginal effects. Generally speaking, this condition requires the degree of linear dependence among the covariates under consideration to be sufficiently weak. Moreover, the OCMT procedure requires an upper bound on the degree of correlation among the covariates so that pair-wise correlations across them are absolute summable. This paper exploits the ideas from research on latent factor models and multiple testing to propose a variable selection method which applies even when the set of covariates under consideration are highly correlated. Suppose that the degree of cross-sectional correlation among the covariates in the active set is strong. In this case, following the factor literature, we decompose the covariates into unobserved *common* and *idiosyncratic* components, where the degree of cross sectional correlation across the idiosyncratic component of the covariates is weak. Ideally, if the common and idiosyncratic components were observable, we could have simply conditioned on the common component, which is already low-dimensional, and confine the variable selection problem to the remaining idiosyncratic components. However, because the common and idiosyncratic components are unobservable, they must be estimated from the data. In this paper, we show that the deviation of **P**rincipal **C**omponent (PC) estimators of common and idiosyncratic components from their true values are bounded in probability sufficiently sharply as $N, T \to \infty$. This finding allows us to condition on the estimated common factors in place of the true ones, and still be able to use the OCMT procedure for valid variable selection. We refer to our proposed method as **G**eneralized **O**ne **C**ovariate at a Time **M**ultiple **T**esting (GOCMT) as it generalize the OCMT procedure to allow the set of covariates in the active set to be highly correlated. We generalize the OCMT procedure, rather than the penalized regression methods, since the OCMT approach is based on reasonably mild assumptions and it does not rely on an unknown tuning parameter. However, one could apply a similar idea to generalize the penalized regression procedures. Our theoretical result shows that the GOCMT procedure asymptotically selects the *approximating model*, which contains all the signals and none of the semi-noise/noise variates. Monte Carlo experiments also indicate that the newly suggested method have appealing finite-sample performance relative to competing methods, such as OCMT, LASSO, Adaptive LASSO (A-LASSO) proposed by Zou (2006), and Intertwined Probabilistic Factors Decoupling (IPAD) proposed Fan et al. (2019) by under many different settings.

Fan et al. (2019) exploit the factor structure, as we do, but for a different purpose. They

are interested in asymptotically controlling the false discovery rate (FDR) in high-dimensional settings with a strong degree of cross-sectional correlation. However, they assume that the idiosyncratic terms are independent across both covariates and observations. Moreover, they assume that the idiosyncratic components are generated from an identical known probability distribution function with unknown finite parameters. In the current paper, we provide the theory behind for the GOCMT procedure under fairly general assumptions. In particular, we allow for the idiosyncratic terms to be weakly dependent across both variables and observations and permit the idiosyncratic terms to be generated from an unknown heterogeneous probability distribution.

We illustrate GOCMT procedure with an empirical application to selection of risk factors that can explain risk premia in stock market. Currently, one important concern in asset pricing literature is to evaluate the relative importance of many of the risk factors suggested to explain risk premia in stock market (Feng et al., 2019). As it is shown the risk factors in asset pricing literature are highly correlated. Therefore, the GOCMT procedure can be consider as a proper tool to evaluate the importance of suggested risk factors in explaining risk premia in stock market. Our results suggest that among the 146 risk factors considered recently by Feng et al. (2019), only the excess market return is strong and can be used to estimate the risk premia. The other risk factors are mostly found to be weak and hence could reflect pricing errors and their selection by standard penalized regression techniques could lead to misleading outcomes.

The rest of the paper is organized as follows: Section 2 sets out the model specification. Section 3 explains the basic idea behind the GOCMT procedure, providing a brief overview of the OCMT method. Section 4 discusses the technical assumptions and the asymptotic properties of the GOCMT procedure. Section 5 gives the details of Monte Carlo experiments and a summary of the main simulation results. Section 6 presents the empirical application, and Section 7 concludes.

Notations: Generic finite positive constants are denoted by C_i for $i = 1, 2, \cdots$. $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_F$ denote the spectral and Frobenius norms of matrix $\mathbf{A}$, respectively. $\|\mathbf{x}\|$ denotes the ℓ_2 norm of vector $\mathbf{x}$. If $\{f_n\}_{n=1}^{\infty}$ and $\{g_n\}_{n=1}^{\infty}$ are both positive sequences of real numbers, then $f_n = \ominus(g_n)$ if there exist $n_0 \geq 1$ and positive constants C_0 and C_1, such that $\inf_{n \geq n_0} (f_n/g_n) \geq C_0$ and $\sup_{n \geq n_0} (f_n/g_n) \leq C_1$.

2 Model Setting

We consider the following data generating process (DGP) for the *target variable*, y_t,

$$y_t = \mathbf{a}'\mathbf{z}_t + \sum_{i=1}^{k} \beta_i x_{it} + u_t, \quad \text{for} \quad t = 1, 2, \cdots, T, \tag{1}$$

where $\mathbf{z}_t$ is a vector of preselected covariates; x_{it} for $i = 1, 2, \cdots, k$ are the covariates with $0 < |\beta_i| \leq C < \infty$ which we refer to as *signals*; and u_t is an error term. It is assumed that $\mathbf{z}_t$ and x_{it}, $i = 1, 2, \cdots, k$, are uncorrelated with u_t at time t. The vector $\mathbf{z}_t$ can contains deterministic components such as a constant, dummy variables, and a deterministic time trend; as well as stochastic variables: observable factors and lag values of y_t. The k signals are unknown and an investigator wishes to select them from the active set $\mathcal{S}_{Nt} = \{x_{1t}, x_{2t}, \cdots, x_{Nt}\}$ with N possibly larger than T.

It is assumed that the covariates in the active set are generated as

$$x_{it} = \boldsymbol{\gamma}_i^{0\prime}\mathbf{f}_t^0 + \varepsilon_{it} = c_{it}^0 + \varepsilon_{it}, , \text{ for } i = 1, 2, \cdots, N; \, t = 1, 2, \cdots, T \tag{2}$$

where $c_{it}^0 = \boldsymbol{\gamma}_i^{0\prime}\mathbf{f}_t^0$ is the *common component* of x_{it}, $\mathbf{f}_t^0 = (f_{0,1t}, f_{0,2t}, \cdots, f_{0,m_0t})'$ is an $m_0 \times 1$ vector of *unobserved common factors*, $\boldsymbol{\gamma}_i^0 = (\gamma_{0,1i}, \gamma_{0,2i}, \cdots, \gamma_{0,m_0i})'$ is an $m_0 \times 1$ vector of *factor loadings*, and ε_{it} is the *idiosyncratic component* of x_{it}. Equations in (2) can be written in the N-dimensional time series format:

$$\mathbf{x}_t = \boldsymbol{\Gamma}^0\mathbf{f}_t^0 + \boldsymbol{\varepsilon}_t, \tag{3}$$

where $\mathbf{x}_t = (x_{1t}, x_{2t}, \cdots, x_{Nt})'$ and $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \varepsilon_{2t}, \cdots, \varepsilon_{Nt})'$ are $N \times 1$ vectors, and $\boldsymbol{\Gamma}^0 = (\boldsymbol{\gamma}_1^0, \boldsymbol{\gamma}_2^0, \cdots, \boldsymbol{\gamma}_N^0)'$ is an $N \times m_0$ matrix. Alternatively, the equations in (2) can be written as T-dimensional cross section format:

$$\mathbf{x}_i = \mathbf{F}^0\boldsymbol{\gamma}_i^0 + \boldsymbol{\varepsilon}_i, \tag{4}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iT})'$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{iT})'$ are $T \times 1$ vectors, and $\mathbf{F}^0 = (\mathbf{f}_1^0, \mathbf{f}_2^0, \cdots, \mathbf{f}_T^0)'$ is a $T \times m_0$ matrix. It is also convenient to write (2) in the matrix format:

$$\mathbf{X} = \mathbf{F}^0\boldsymbol{\Gamma}^{0\prime} + \mathbf{E}, \tag{5}$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ and $\mathbf{E} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \cdots, \boldsymbol{\varepsilon}_N)$ are $T \times N$ matrices.

In this setting, all x_{it}'s can possibly be correlated with each other through the unobserved common factors, $\mathbf{f}_t^0$, with the degree of cross-sectional dependence determined by the matrix of factor loadings, $\boldsymbol{\Gamma}^0$. Also following the factor literature we allow the idiosyncratic components to be weakly cross-correlated such that $\sup_i \sum_{j=1}^{N} |\mathbb{E}(\varepsilon_{it}\varepsilon_{jt})| \leq M < \infty$. This condition is required so that the common components, c_{it}, can be identified (distinguished) from the idiosyncratic components, ε_{it}.

3 Generalized One Covariate at a Time Multiple Testing Method

3.1 GOCMT: The Basic Idea

To highlight the basic idea behind GOCMT initially suppose that the common factors and idiosyncratic components of the covariates were observable. Then by substituting (2) into (1), we would have

$$y_t = \mathbf{a}'\mathbf{z}_t + \boldsymbol{\delta}'\mathbf{f}_t^0 + \sum_{i=1}^{k} \beta_i \varepsilon_{it} + u_t, \tag{6}$$

where $\boldsymbol{\delta} = \sum_{i=1}^{k} \beta_i \boldsymbol{\gamma}_i^{0\prime}$. In (6), the marginal effect of ε_{it} on y_t; β_i; is equal to that of x_{it} on y_t in (1). Additionally, the idiosyncratic terms are weakly cross correlated, namely $\sup_i \sum_{j=1}^{N} |\mathbb{E}(\varepsilon_{it}\varepsilon_{jt})| \leq M < \infty$. Therefore, by conditioning on the common factors, and focusing on ε_{it} instead of x_{it}, we can use existing methods such as penalized regression or OCMT for the purpose of variable selection. However, the common factors and idiosyncratic component of x_{it} are unobservable and hence need to be estimated from the data. In this paper, we establish condition under which that the deviation of PC estimators of common and idiosyncratic components from its true values are bounded in probability sufficiently sharply as $N, T \to \infty$. This result allows us to condition on the estimated common factors in place of the true ones, and still be able to use the OCMT procedure for valid variable selection. We generalize the OCMT procedure, rather than the penalized regression methods, since the OCMT approach is based on reasonably mild assumptions and it does not require calibration of unknown tuning parameters. However, one could use a similar idea to extend the penalized regression methods to the case of highly correlated covariates.

3.2 An Overview of the OCMT Procedure

Chudik et al. (2018) categorize the covariates in the active set into three groups: *signals*, *pseudo-signals* and *noise variates*. As mentioned in Section 2, *signals* are the covariates with non-zero slope coefficient; $\beta_i \neq 0$; in DGP (1). *Pseudo-signals* are the covariates that do not enter the DGP but have non-zero correlations with the signals once the effect of $\mathbf{z}_t$ is filtered out, and hence can be falsely viewed as signals. *Noise variates* are those covariates that conditional on $\mathbf{z}_t$ have zero correlation with signals, and hence are uncorrelated with the target variable.

For each covariate x_{it}, $i = 1, 2, \cdots, N$, the OCMT procedure considers the following linear regression of y_t on x_{it} conditional on $\mathbf{z}_t$:

$$y_t = \boldsymbol{\lambda}_i'\mathbf{z}_t + \phi_{i,T} x_{it} + \eta_{it}, \ t = 1, 2, \cdots, T \tag{7}$$

where $\phi_{i,T} = [\mathbb{E}(\mathbf{x}_i'\mathbf{M}_{\mathbf{z}}\mathbf{x}_i)]^{-1}[\mathbb{E}(\mathbf{x}_i'\mathbf{M}_{\mathbf{z}}\mathbf{y})]$, $\mathbf{y} = (y_1, y_2, \cdots, y_T)'$, $\mathbf{M}_{\mathbf{z}} = \mathbf{I}_T - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, and

$\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T)'$. Furthermore, by substituting $\mathbf{y}$ from (1), we have

$$\phi_{i,T} = \frac{\sum_{j=1}^{k} \beta_j \sigma_{ij,T}(\mathbf{z})}{\sigma_{ii,T}(\mathbf{z})} = \frac{\theta_{i,T}(\mathbf{z})}{\sigma_{ii,T}(\mathbf{z})} \tag{8}$$

where $\sigma_{ij,T}(\mathbf{z}) = \mathbb{E}\left(T^{-1}\mathbf{x}_i'\mathbf{M}_{\mathbf{z}}\mathbf{x}_j\right)$, and $\theta_{i,T}(\mathbf{z}) = \sum_{j=1}^{k} \beta_j \sigma_{ij,T}(\mathbf{z})$ is the *net impact* of x_{it} on y_t. As it is clear from (8), the OCMT procedure focuses on the net impact of x_{it} on y_t, $\phi_{i,T}$, rather than the marginal effect, β_i, for variable selection. Due to correlation between the covariates, knowing $\phi_{i,T}$ does not imply that we can determine β_i. There are four possibilities.

	$\phi_{i,T} \neq 0$	$\phi_{i,T} = 0$
$\beta_i \neq 0$	(I) Unhidden Signals	(II) Hidden Signals
$\beta_i = 0$	(III) Pseudo-signals	(IV) Noises

The goal of the OCMT procedure is to use the t-ratio of the estimated $\phi_{i,T}$ to select all the signals and none of the noise variables, the selected model is referred to as an *approximating model* since it can include pseudo-signals. The approximating model can be used for forecasting or as a basis for separating signals from pseudo-signals by application of standard model selection techniques. To deal with the multiple testing nature of the problem, Chudik et al. (2018) adjust the critical value of the tests so that they are monotonically increasing function of N. They show that the probability of selecting the approximating model by the OCMT procedure tends to one as N and T go to infinity, so long as the number of pseudo-signals rise at an order sufficiently lower than N and T. Clearly, if the signals have a common factor shared with the other covariates in the active set, as it will be the case under the common factor representation, (2), then, the number of pseudo-signals will rise at the same order as N, and hence the OCMT procedure fails to apply. The main propose of this paper is to use the basic idea described in Section 3.1 to generalize the OCMT procedure to the case where the number of pseudo-signals can rise at the same order as N.

3.3 GOCMT Procedure

To simplify the exposition, from now on, we assume that the set of preselected covariates is empty, in which case the DGP (1) simplifies to

$$\mathbf{y} = \sum_{i=1}^{k} \beta_i \mathbf{x}_i + \mathbf{u}, \tag{9}$$

where $\mathbf{u} = (u_1, u_2, \cdots, u_T)'$ is a $T \times 1$ vector of error terms. For each covariate x_{it}, $i = 1, 2, \cdots, N$, the GOCMT procedure considers the regression of y_t on x_{it} conditional on $\mathbf{f}_t^0$, $y_t = \boldsymbol{\phi}_{f,i}'\mathbf{f}_t^0 + \phi_{i,x}x_{it} + \eta_{it}$, where $\phi_{i,x} = [\mathbb{E}(\mathbf{x}_i'\mathbf{M}_{\mathbf{f}^0}\mathbf{x}_i)]^{-1}[\mathbb{E}(\mathbf{x}_i'\mathbf{M}_{\mathbf{f}^0}\mathbf{y})]$, and $\mathbf{M}_{\mathbf{f}^0} = \mathbf{I}_T - \mathbf{F}^0(\mathbf{F}^{0\prime}\mathbf{F}^0)^{-1}\mathbf{F}^{0\prime}$.

Furthermore, by substituting $\mathbf{y}$ from (9), we have

$$\phi_{i,x} = \frac{\sum_{j=1}^{k} \beta_j \sigma_{ij,T}(\mathbf{f}^0)}{\sigma_{ii,T}(\mathbf{f}^0)} = \frac{\theta_{i,T}(\mathbf{f}^0)}{\sigma_{ii,T}(\mathbf{f}^0)} \tag{10}$$

where $\sigma_{ij,T}(\mathbf{f}^0) = \mathbb{E}\left(T^{-1}\mathbf{x}_i'\mathbf{M}_{\mathbf{f}^0}\mathbf{x}_j\right) = \mathbb{E}\left(T^{-1}\boldsymbol{\varepsilon}_i'\mathbf{M}_{\mathbf{f}^0}\boldsymbol{\varepsilon}_j\right)$, and $\theta_{i,T}(\mathbf{f}^0) = \sum_{j=1}^{k} \beta_j \sigma_{ij,T}(\mathbf{f}^0)$. As it is clear from (10), the GOCMT procedure focuses on $\phi_{i,x}$, the net impact of ε_{it} on y_t for variable selection. As it is discussed in section 3.2, knowing $\phi_{i,x}$ does not imply that we can determine β_i. For a signal to be hidden, we need $\phi_{i,x}$, the net impact of ε_{it} on y_t, to be exactly equal to zero which is very unlikely. In what follows, we first assume that there exist no hidden signals and propose a single stage variable selection procedure. Then, to deal with the possibility of existence of hidden signals, we extend the procedure to possibly have multiple stages.

3.3.1 No Hidden Signals and Single Stage GOCMT

Given the number of factors, m_0, the single stage GOCMT procedure is as follows:

1. *The $T \times m_0$ matrix of PC estimator of factors, $\tilde{\mathbf{F}} = (\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \cdots, \tilde{\mathbf{f}}_T)'$, is computed by*

$$\tilde{\mathbf{F}} = \sqrt{T}\tilde{\mathbf{P}}, \tag{11}$$

where $\tilde{\mathbf{P}}$ is a $T \times m_0$ matrix of orthonormal eigenvectors corresponding to the m_0 largest eigenvalues of the $T \times T$ matrix, $\mathbf{X}\mathbf{X}'$.

2. *For $i = 1, 2, \cdots, N$, regress $\mathbf{y}$ on $\tilde{\mathbf{F}}$ and $\mathbf{x}_i$; $\mathbf{y} = \tilde{\mathbf{F}}\boldsymbol{\phi}_{f,i} + \phi_{x,i}\mathbf{x}_i + \boldsymbol{\eta}_i$; and compute the t-ratio of $\phi_{x,i}$, given by*

$$t_{i,T} = \frac{\hat{\phi}_{x,i}}{s.e.\left(\hat{\phi}_{x,i}\right)} = \frac{\mathbf{x}_i'\mathbf{M}_{\tilde{\mathbf{f}}}\mathbf{y}}{\hat{\sigma}_{i,(1)}\sqrt{\mathbf{x}_i'\mathbf{M}_{\tilde{\mathbf{f}}}\mathbf{x}_i}},$$

where $\mathbf{M}_{\tilde{\mathbf{f}}} = \mathbf{I} - \tilde{\mathbf{F}}(\tilde{\mathbf{F}}'\tilde{\mathbf{F}})^{-1}\tilde{\mathbf{F}}'$, $\hat{\phi}_{x,i} = \left(\mathbf{x}_i'\mathbf{M}_{\tilde{\mathbf{f}}}\mathbf{x}_i\right)^{-1}\left(\mathbf{x}_i'\mathbf{M}_{\tilde{\mathbf{f}}}\mathbf{y}\right)$ is the Least Square (LS) estimator of $\phi_{x,i}$, $\hat{\sigma}_i^2 = \hat{\boldsymbol{\eta}}_i'\hat{\boldsymbol{\eta}}_i/T$, and $\hat{\boldsymbol{\eta}}_i$ is a $T \times 1$ vector of regression residuals.

3. *Consider the critical value function, $c_p(N, \delta)$, defined by*

$$c_p(N, \delta) = \Phi^{-1}\left(1 - \frac{p}{2N^{\delta}}\right), \tag{12}$$

where $\Phi^{-1}(.)$ is the inverse of a standard normal distribution function, δ is a finite positive constant, and p $(0 < p < 1)$ is the nominal size of the individual tests to be set by the investigator.

4. *Given $c_p(N, \delta)$, the selection indicator is given by*

$$\hat{\mathcal{J}}_i = \boldsymbol{I}\left[|t_{i,T}| > c_p(N, \delta)\right], \text{ for } i = 1, 2, \cdots, N.$$

The covariates x_{it} is selected if $\hat{\mathcal{J}}_i = 1$.

If all the signals were unhidden, $\phi_{i,x} \neq 0$ for all i's with $\beta_i \neq 0$, It is reasonable to expect that the single step procedure could potentially select them all. However, if a signal were hidden, $\phi_{i,x} = 0$ while $\beta_i \neq 0$, it would not be detected by the single stage procedure. In the following section, we first discuss a possible solution to this problem and then provide the multi-stage GOCMT procedure to deal with selection of hidden signals.

3.3.2 Hidden Signals Possibility and Multi-Stage GOCMT

As it is formally discussed later on in proposition 1, it is impossible for all the signals to be hidden in one stage. Therefore, there exists at least one signal with $\phi_{i,x} \neq 0$ which can be selected by the single stage GOCMT procedure. Now, after conditioning on the selected signals with non-zero net effect, $\phi_{i,x} \neq 0$, and the common factors, there exists at least one more signal whose net effect, $\phi_{i,x}$, becomes non-zero. Therefore, by repeating the similar exercise as in the single stage procedure where the conditioning set is now augmented by the selected covariates in the previous stages, we can detect all the hidden signals. Following this intuition, for a give number of factors, m_0, we extent the proposed procedure to potentially have multiple stages as it is described below.

1. *Run the single stage GOCMT procedure and select the covariates x_{it} if $\hat{\mathcal{J}}_{i,(1)} = 1$.*

2. *If $\sum_{i=1}^{N} \hat{\mathcal{J}}_{i,(1)} = 0$, stop the procedure without selecting any covariates. Otherwise, continue to the next stage.*

3. *Let $\mathcal{S}^o_{(j)}$ denote the index set of all selected covariates up to and including stage j of GOCMT. Then given $\mathcal{S}^o_{(j-1)}$ and $\tilde{\mathbf{F}}$, at any stages $j \geq 2$:*

   (a) *Let $\mathfrak{S}_j = \{1, 2, \cdots, N\} \backslash \mathcal{S}^o_{(j-1)}$ denote the stage j active index set. For $i \in \mathfrak{S}_j$, regress $\mathbf{y}$ on $\tilde{\mathbf{Q}}_{(j)}$ and $\mathbf{x}_i$, where $\tilde{\mathbf{Q}}_{(j)} = \left[\tilde{\mathbf{F}}, \mathbf{X}_{\mathcal{S}^o_{(j-1)}}\right]$ and $\mathbf{X}_{\mathcal{S}^o_{(j-1)}}$ is a matrix including covariates selected up to stage j, $\mathbf{y} = \tilde{\mathbf{Q}}_{(j)}\boldsymbol{\phi}_{\tilde{q},i,(j)} + \phi_{x,i,(j)}\mathbf{x}_i + \boldsymbol{\eta}_{i,(j)}$. Compute the t-ratio of $\phi_{x,i,(j)}$, which is*

   $$t_{i,T,(j)} = \frac{\hat{\phi}_{x,i,(j)}}{s.e.\left(\hat{\phi}_{x,i,(j)}\right)} = \frac{\mathbf{x}_i' \mathbf{M}_{\tilde{\mathbf{q}}_{(j)}} \mathbf{y}}{\hat{\sigma}_{i,(j)} \sqrt{\mathbf{x}_i' \mathbf{M}_{\tilde{\mathbf{q}}_{(j)}} \mathbf{x}_i}},$$

   where $\mathbf{M}_{\tilde{\mathbf{q}}_{(j)}} = \mathbf{I} - \tilde{\mathbf{Q}}_{(j)}(\tilde{\mathbf{Q}}'_{(j)}\tilde{\mathbf{Q}}_{(j)})^{-1}\tilde{\mathbf{Q}}'_{(j)}$, $\hat{\phi}_{x,i,(j)} = \left(\mathbf{x}_i' \mathbf{M}_{\tilde{\mathbf{q}}_{(j)}} \mathbf{x}_i\right)^{-1} \left(\mathbf{x}_i' \mathbf{M}_{\tilde{\mathbf{q}}_{(j)}} \mathbf{y}\right)$ is the LS estimator of $\phi_{x,i,(j)}$, $\hat{\sigma}^2_{i,(j)} = \hat{\boldsymbol{\eta}}'_{i,(j)}\hat{\boldsymbol{\eta}}_{i,(j)}/T$, and $\hat{\boldsymbol{\eta}}_{i,(j)}$ is a $T \times 1$ vector of regression residuals..

   (b) *Consider critical value function $c_p(N, \delta^*)$ where $\delta^* > \delta > 0$.*

(c) Given $c_p(N,\delta^*)$, the stage j selection indicator is given by

$$\hat{\mathcal{J}}_{i,(j)} = \boldsymbol{I}\left[\left|t_{i,T,(j)}\right| > c_p(N,\delta^*)\right], \text{ for } i \in \mathfrak{S}_j.$$

Covariate x_{it} with $\hat{\mathcal{J}}_i,(j)=1$ is selected.

(d) If $\sum_{i=1}^{N}\hat{\mathcal{J}}_{i,(j)}=0$, stop the procedure and consider

$$\hat{\mathcal{J}}_i = \sum_{j=1}^{\hat{\mathbb{S}}}\hat{\mathcal{J}}_{i,(j)}, \tag{13}$$

where $\hat{\mathbb{S}}$ denotes the number of stages at completion of GOCMT, formally defined as

$$\hat{\mathbb{S}} = \min_j\{j : \textstyle\sum_{i=1}^{N}\hat{\mathcal{J}}_{i,(j)}=0\} - 1. \tag{14}$$

Otherwise, continue to the next stage.

In the rest of this paper we refer to the proposed multi-stage GOCMT procedure as the GOCMT procedure. In the worst case scenario, we would have only one signal with non-zero net effect, $\theta_{iT,(j)} \neq 0$, at each stage of the GOCMT procedure. In this case, if the procedure selects the signal with non-zero net effect at each stage, the number of stages will be equal to the number of signals, k. In section 4.2, it is formally shown that the probability of the event that the number of GOCMT stages be greater than k, approaches to zero as $N,T \to \infty$. In practice, it is very unlikely for a signal to be hidden and hence the GOCMT procedure does not go beyond the first stage.

Before providing our theoretical results in the next section, we would like to highlight some points regarding the GOCMT procedure. Firstly, the suggested method assumes that at least one common factor exists among all the covariates in the active set $\mathcal{S}_{Nt}$. But in practice, to check this assumption, we can first use the Pesaran (2015) test for the degree of cross-sectional correlation (CD test) which is

$$\text{CD} = \sqrt{\tfrac{2}{N(N-1)}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sqrt{T}\hat{\rho}_{ij},$$

where $\hat{\rho}_{ij}$ is the sample correlation between variables i and j. To understand the intuition behind the CD test, we can write (2) as $x_{it} = \sum_{\ell=1}^{m_0}\gamma_{0,\ell i}f_{0,\ell t}+\varepsilon_{it}$, and define the degree of cross-sectional dependence due to the ℓ^{th} factor by $\alpha_\ell = \frac{M_\ell}{N}$ where $M_\ell = \sum_{i=1}^{N}\boldsymbol{I}(|\gamma_{0,\ell i}|>0)$. Further, we can define the overall degree of cross-sectional dependence by $\alpha = \max_\ell \alpha_\ell$. Bailey et al. (2016) refer to α as the *exponent of cross-sectional dependence* and Pesaran (2015) shows that the average pair-wise correlation coefficient given as $\bar{\rho} = \frac{2}{N(N-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\rho_{ij}$ is $O(N^{2\alpha-2})$, where ρ_{ij} is the correlation between variable i and j. Therefore, if $\alpha < 1/2$, $\bar{\rho}\to 0$ as $N\to\infty$. So, roughly speaking the CD test allows us to test whether covariates under consideration are highly

correlated or not.

Secondly, signals and pseudo-signals/noise variables may not share all the factors. For example, consider the following special case of (3).

$$\begin{pmatrix} \mathbf{x}_{1t} \\ \mathbf{x}_{2t} \end{pmatrix} = \begin{pmatrix} \mathbf{\Gamma}_{11} & 0 \\ \mathbf{\Gamma}_{21} & \mathbf{\Gamma}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{f}_{1t} \\ \mathbf{f}_{2t} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{1t} \\ \boldsymbol{\varepsilon}_{2t} \end{pmatrix},$$

where $\mathbf{x}_{1t}$ is a $k \times 1$ vector of signals, $\mathbf{x}_{2t}$ is a $N-k$ vectors of pseudo-signals, $\mathbf{f}_{1t}$ and $\mathbf{f}_{2t}$ are $m_1 \times 1$ and $m_2 \times 1$ vector of factors with $m_1 + m_2 = m_0$, and $\mathbf{\Gamma}_{11}$, $\mathbf{\Gamma}_{21}$, and $\mathbf{\Gamma}_{22}$ are $k \times m_1$, $N-k \times m_1$ and $N-k \times m_2$ matrices of factor loadings. In this case, while $\mathbf{f}_{1t}$ is common among signals and pseudo-signals, $\mathbf{f}_{2t}$ is only common among the pseudo-signals. By substituting $\mathbf{x}_{1t}$ into (9), we have

$$y_t = \boldsymbol{\beta}_1' \mathbf{\Gamma}_{11} \mathbf{f}_{1t} + \boldsymbol{\beta}_1' \boldsymbol{\varepsilon}_{1t} + u_t = \boldsymbol{\delta}_1' \mathbf{f}_{1t} + \boldsymbol{\beta}_1' \boldsymbol{\varepsilon}_{1t} + u_t. \tag{15}$$

As it can be seen clearly in (15), we only need the common factors that link the signals to pseudo-signals, $\mathbf{f}_{1t}$, in the conditioning set and there is no need to control the remaining factors, $\mathbf{f}_{2t}$, that are only common across pseudo-signals. In practice, after estimation of factors $\mathbf{f}_t = (\mathbf{f}_{1t}, \mathbf{f}_{2t})'$, $\tilde{\mathbf{f}}_t$, we can run a regression of y_t on $\tilde{\mathbf{f}}_t$, and select the ones with significant coefficients using standard t tests.

Finally, note that the GOCMT procedure assumes that the true number of factors is known. In practice we can use procedures suggested in the literature to estimate the number of factors consistently, such as those of Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013). Alternatively, we can set the number of factors to $m_{\max}$ known to be greater than m_0.

4 Asymptotic Properties of GOCMT Procedure

We now provide the theoretical justifications for the proposed GOCMT procedure. In the rest of this paper, we will refer to the covariates whose `idiosyncratic component` have non-zero correlation with the `idiosyncratic component` of `signals` once the effect of $\mathbf{z}_t$ is filtered out as *strong-pseudo-signals*. Moreover, we will refer to the covariates whose `idiosyncratic component` conditional on $\mathbf{z}_t$ have zero correlation with `idiosyncratic component` of `signals` as *semi-noise variables*. Further, we will denote the number of strong-pseudo-signals by k^*. It is assumed that both k and k^* are unknown finite fixed integers. Finally, we definite an *approximating model* to be a model that contains all the signals; $\{x_{it} : i = 1, 2, \cdots, k\}$; and none of the semi-noise variables; $\{x_{it} : k + k^* + 1, k + k^* + 2, \cdots, N\}$. Clearly, such models can contain one or more of the strong-pseudo-signals; $\{x_{it} : k + 1, k + 2, \cdots, k + k^*\}$. In (2), if no factors exist among the covariates in the active set, then, strong-pseudo-signals and semi-noise

variables are equivalent to pseudo-signals and noise variables, respectively. Hence, our modified definition of approximating models will match the original definition of Chudik et al. (2018). We start with some technical assumptions in Section 4.1 and then provide the asymptotic properties of the GOCMT procedure in Section 4.2.

4.1 Technical Assumptions

In what follows we make use of the following filtrations: $\mathcal{F}_t^u = \sigma(u_t, u_{t-1}, \cdots)$, $\mathcal{F}_{\ell t}^f = \sigma(f_{0,\ell t}, f_{0,\ell t-1}, \cdots)$ for $\ell = 1, 2, \cdots, m_0$, and $\mathcal{F}_{it}^\varepsilon = \sigma(\varepsilon_{it}, \varepsilon_{i,t-1}, \cdots)$ for $i = 1, 2, \cdots, N$. Moreover, we set $\mathcal{F}_t^f = \cup_{\ell=1}^{m_0} \mathcal{F}_{\ell t}^f$, $\mathcal{F}_t^\varepsilon = \cup_{i=1}^{N} \mathcal{F}_{it}^\varepsilon$, and $\mathcal{F}_t = \mathcal{F}_t^f \cup \mathcal{F}_t^\varepsilon \cup \mathcal{F}_t^u$.

Assumption 1 (Factors)
$T^{-1} \sum_{t=1}^{T} \mathbf{f}_t^0 \mathbf{f}_t^{0\prime} \to \mathbf{\Sigma_F}$ *as* $T \to \infty$ *for some* $m_0 \times m_0$ *positive definite matrix* $\mathbf{\Sigma_F}$.

Assumption 2 (Factor loadings)
$\|\boldsymbol{\gamma}_i^0\|_F \leq \bar{\gamma} < \infty$ *for all* $i = 1, 2, \cdots, N$, *and* $\|\mathbf{\Gamma}^{0\prime}\mathbf{\Gamma}^0/N - \mathbf{\Sigma_\Gamma}\|_F \to 0$, *as* $N \to \infty$ *for some* $m_0 \times m_0$ *positive definite matrix* $\mathbf{\Sigma_\Gamma}$.

Assumption 3 (Idiosyncratic components)

(i) $\mathbb{E}(\varepsilon_{it}) = 0$.

(ii) $\mathbb{E}(\varepsilon_{it}\varepsilon_{jt}) = \sigma_{ij,t}$ *with* $\sup_t |\sigma_{ij,t}| \leq |\sigma_{ij}|$ *for some* σ_{ij}; *in addition,* $\sup_i \sum_{j=1}^{N} |\sigma_{ij}| \leq M$ *where* M *is a finite positive number.*

(iii) ε_{it} *is independent of* ε_{js} *for* $i = 1, 2, \cdots, k + k^*$ *and* $j = k + k^* + 1, k + k^* + 2, \cdots, N$, *and for all* t *and* s.

(iv) Let $\mathbf{E}_{k,k^*} = (\mathbf{E}_k, \mathbf{E}_{k^*})$, *where* $\mathbf{E}_k = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \cdots, \boldsymbol{\varepsilon}_k)$ *and* $\mathbf{E}_{k^*} = (\boldsymbol{\varepsilon}_{k+1}, \cdots, \boldsymbol{\varepsilon}_{k+k^*})$ *are* $T \times k$ *and* $T \times k^*$ *matrices of idiosyncratic terms for signals and strong-pseudo-signals. There exists* T_0 *such that for all* $T > T_0$, $T^{-1}\mathbf{E}'_{k,k^*}\mathbf{E}_{k,k^*}$ *is nonsingular with its eigenvalues uniformly bounded away from 0, and* $\mathbf{\Sigma}_{k,k^*} = \mathbb{E}(T^{-1}\mathbf{E}'_{k,k^*}\mathbf{E}_{k,k^*})$ *is nonsingular for all* T.

(v) The number of strong-pseudo-signals, k^*, *is a finite fixed integer.*

Assumption 4 (Exponential decaying probability tails)
There exist sufficiently large positive constants C_0 *and* C_1, *and* $s > 0$ *such that*

(i) $\sup_{\ell,t} \Pr(|f_{0,\ell t}| > \alpha) \leq C_0 \exp(-C_1 \alpha^s)$, *for all* $\alpha > 0$.

(ii) $\sup_{i,t} \Pr(|\varepsilon_{it}| > \alpha) \leq C_0 \exp(-C_1 \alpha^s)$, *for all* $\alpha > 0$.

(iii) $\sup_t \Pr(|u_t| > \alpha) \leq C_0 \exp(-C_1 \alpha^s)$, *for all* $\alpha > 0$.

Assumption 5 (Martingale difference processes)

(i) $\mathbb{E}\left[f_{0,\ell t} f_{0,\ell' t} - \mathbb{E}(f_{0,\ell t} f_{0,\ell' t})|\mathcal{F}_{t-1}\right] = 0$, *for* $\ell, \ell' = 1, 2, \cdots, m_0$ *and all* t.

(ii) $\mathbb{E}\left[\varepsilon_{it}\varepsilon_{jt} - \mathbb{E}(\varepsilon_{it}\varepsilon_{jt})|\mathcal{F}_{t-1}\right] = 0$, *for* $i, j = 1, 2, \cdots, N$, *and all* t.

(iii) $\mathbb{E}\left(u_t^2 - \mathbb{E}\left(u_t^2\right)|\mathcal{F}_{t-1}\right) = 0$ *for all* t.

(iv) $\mathbb{E}\left(f_{0,\ell t}\varepsilon_{it}|\mathcal{F}_{t-1}\right) = 0$ *for* $i = 1, 2, \cdots, N$; $\ell = 1, 2, \cdots, m_0$; *and for all* t.

(v) $\mathbb{E}\left(f_{0,\ell t} u_t|\mathcal{F}_{t-1}\right) = 0$ *for* $\ell = 1, 2, \cdots, m_0$ *and all* t.

(vi) $\mathbb{E}\left(\varepsilon_{it} u_t|\mathcal{F}_{t-1}\right) = 0$ *for* $i = 1, 2, \cdots, N$ *and all* t.

Assumption 6 (Coefficients of signals)

The number of signals, k, *is a finite fixed integer, and their slope coefficients could change with* T, *such that for* $i = 1, 2, \cdots, k$, $\beta_{i,T} = \ominus(T^{-\vartheta})$, *for some* $0 \leq \vartheta < 1/2$.

Before presenting our theoretical results, we should mention pros and cons of our assumptions and compare them with the assumptions typically made in the high-dimensional linear regression and factor models literature.

Assumptions 1, 2, 3(i), and 3(ii) are common in the factor literature, for example, see Bai and Ng (2002). Assumption 3(iv) corresponds to Assumption 1 of Chudik et al. (2018). This assumption ensures that the regression coefficients in the model, which contains all signals and strong-pseudo-signals, and no semi-noise variables, are identified.

The exponentially decaying probability tails for $f_{\ell t}$, ε_{it}, and u_t, in Assumptions 4 ensures that all moments of $f_{\ell t}$, ε_{it}, and u_t exist. The exponentially decaying probability tails for $f_{\ell t}$, ε_{it} are beyond those needed for estimation of unobserved factors, but are required for deriving the exponential decaying rate of convergence of estimated factors in Lemma 1. It is common in the high-dimensional linear literature to assume some form of exponentially decaying probability bound for the variables. For example, see Zheng et al. (2014), Fan et al. (2019) and Chudik et al. (2018).

Assumption 5 allows $f_{\ell t}$, ε_{it}, and u_t to follow martingale difference processes, which is weaker than the iid assumptions typically made in the literature. Following a similar line of argument as in Section 4.2 of Chudik et al. (2018), we can relax these assumptions to allow for some weak serial correlation in the errors, factors and idiosyncratic components. Assumption 5(iv) also restricts $f_{\ell t}$ and ε_{it} to be uncorrelated with each other, which is common in the factor

literature, for example, see Bai and Ng (2008). It is also common in regression analysis to assume x_{it} and u_t are uncorrelated, which is implied by Assumptions 5(v), and 5(vi). We can relax these assumptions to allow for x_{it} to be possibly correlated with past values of u_t by using an argument similar to Section 4.2 of Chudik et al. (2018).

Assumption 6 allows signals to be weak, such that $\beta_{i,T}$, for $i = 1, 2, \cdots, k$ can decline with the sample size, T, at a sufficiently slow rate. To simplify the notation, subscript T is dropped subsequently, and it is understood that the slope coefficients can change with the sample size according to this assumption.

4.2 Theoretical Findings

As it is discussed in section 3.1, the GOCMT procedure relies on deviation of PC estimators of unobserved common factors from its underling true values to be bounded in probability sufficiently sharply as $N, T \to \infty$. Therefore, we start our theoretical findings by providing such the probability bound for the PC estimators.

Lemma 1 *Let $\tilde{\mathbf{F}} = (\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \cdots, \tilde{\mathbf{f}}_T)'$ be a $T \times m_0$ matrix of estimated factors given by $\tilde{\mathbf{F}} = \sqrt{T}\tilde{\mathbf{P}}$, where $\tilde{\mathbf{P}}$ is a $T \times m_0$ matrix of orthonormal eigenvectors corresponding to the m_0 largest eigenvalues of a $T \times T$ matrix $\mathbf{X}\mathbf{X}'$. Suppose that $N = \ominus(T^{\kappa})$ for some $\kappa > 0$, and $d_T = \ominus\left(T^{\lambda}\right)$ for some $\lambda > \max\{1/2, 1 - \kappa\}$. Then under Assumptions 1, 2, 3(i)-(ii), 4(i)-(ii), and 5(i)-(ii) there exists a finite positive constant C_0 such that if $\lambda \leq (s+2)/(s+4)$,*

$$\Pr\left(\|\tilde{\mathbf{F}} - \mathbf{F}^*\|_F^2 > d_T\right) \leq N^2 \exp\left[-C_0 T^{-1} d_T^2\right],$$

if $(s+2)/(s+4) < \lambda < 1$,

$$\Pr\left(\|\tilde{\mathbf{F}} - \mathbf{F}^*\|_F^2 > d_T\right) \leq N^2 \exp\left[-C_0 d_T^{s/(s+2)}\right],$$

and if $\lambda \geq 1$,

$$\Pr\left(\|\tilde{\mathbf{F}} - \mathbf{F}^*\|_F^2 > d_T\right) \leq N^2 \exp\left[-C_0 (T d_T)^{s/2(s+2)}\right],$$

where $\mathbf{F}^* = \mathbf{F}^0\mathbf{G}$ is a $T \times m_0$ matrix of rotated unobserved common factors with $\mathbf{G} = \left(\mathbf{\Gamma}^{0\prime}\mathbf{\Gamma}^0/N\right)\left(\mathbf{F}^{0\prime}\tilde{\mathbf{F}}/T\right)\mathbf{V}_{NT}^{-1}$ and $\mathbf{V}_{NT}$ is an $m_0 \times m_0$ diagonal matrix of the m_0 largest eigenvalues of $\mathbf{X}'\mathbf{X}/(NT)$.

Remark 1 *Note that in Lemma 1, we need both N and T to go to infinity to control the deviation of PC estimators, $\tilde{\mathbf{F}}$, from its rotated true values. Moreover, if N grow at the rate greater than or equal to $\sqrt{T}$, $\kappa \geq 1/2$, we can control the sum of square of all estimation errors, $\|\tilde{\mathbf{F}} - \mathbf{F}^*\|_F^2$ at the order of $\ominus(T)^{(1/2)+\epsilon}$ were ϵ is an arbitrary small positive number. Finally, in Lemma 1, we provide exponential decaying probability bound for the deviation of PC estimators, $\tilde{\mathbf{F}}$, from its rotated true values with respect to number of observations, T.*

Remark 2 *It is also interesting to note that we can always write*

$$\mathbf{X} = \mathbf{F}^0\mathbf{\Gamma}^{0\prime} + \mathbf{E} = \mathbf{F}^0\mathbf{G}\mathbf{G}^{-1}\mathbf{\Gamma}^{0\prime} + \mathbf{E} = \mathbf{F}^*\mathbf{\Gamma}^{*\prime} + \mathbf{E},$$

where $\mathbf{F}^* = \mathbf{F}^0\mathbf{G}$, $\mathbf{\Gamma}^* = \mathbf{\Gamma}^0\mathbf{G}'^{-1}$ *and* $\mathbf{G}$ *is a* $m_0 \times m_0$ *invertible matrix. Since the GOCMT procedure focuses on variation in the idiosyncratic component;* $\mathbf{E}$*; of* $\mathbf{X}$*, for the purpose of variable selection, we are not required to identify* $\mathbf{F}^0$ *and only need to have an accurate estimator of* $\mathbf{F}^*$*.*

In section 3.3, it is discussed that the GOCMT procedure focuses on net effect of ε_{it} on y_t, ϕ_i, rather than the marginal effect, β_i, for variable selection. However, in this case, it is possible for a signal to be hidden, $\phi_{x,i} = 0$ while $\beta_i \neq 0$. Assuming that there exists at least one signal with non-zero net effect, $\phi_{x,i} \neq 0$ at each stage of GOCMT, we extend the procedure to include multiple stages to deal with selection of hidden signals. In proposition 1, we check the validity of this assumption. We also show that in the population level, where $\Pr\left[\left|t_{i,T,(j)}\right| > c_p(N,\delta)|\phi_{x,i,(j)} \neq 0\right] = 1$ and $\Pr\left[\left|t_{i,T,(j)}\right| > c_p(N,\delta)|\phi_{x,i,(j)} = 0\right] = 0$ for all covariates indexed by i and stages indexed by j, the number of GOCMT stages cannot exceed the number of signals, k.

Proposition 1 *Let* x_{it} *for* $i = 1, 2, \cdots, N$ *and* $t = 1, 2, \cdots, T$ *be generated by (2) and* y_t *for* $t = 1, 2, \cdots, T$ *be generated by (9). Under Assumption 3(iv), there exists* j *in range* $1 \leq j < k$*, for which* $\theta_{iT,(j)} \neq 0$*, and the population value of the number of stages required to select all the signals, denoted as* $\mathbb{S}_0$*, satisfies* $1 \leq \mathbb{S}_0 \leq k$*.*

Of course, these probabilities do not take a value of 1 and 0 in a finite sample. In Theorem 1, we show the probability of the event that the number of GOCMT stages exceeds the number of signals, $\hat{\mathbb{S}} > k$, tends to zero as both N and T go to infinity.

As mentioned in Section 3.2, an *approximating model* is a model that contains all the signals; $\{x_{it} : i = 1, 2, \cdots, k\}$; and none of the semi-noise variables; $\{x_{it} : k + k^* + 1, k + k^* + 2, \cdots, N\}$. Clearly, such models can contain one or more of the strong-pseudo-signals; $\{x_{it} : k + 1, k + 2, \cdots, k + k^*\}$. Therefore, the event of choosing the approximating model is define by

$$\mathcal{A}_0 = \left\{\textstyle\sum_{i=1}^{k} \hat{\mathcal{J}}_i = k\right\} \cap \left\{\textstyle\sum_{i=k+k^*+1}^{N} \hat{\mathcal{J}}_i = 0\right\}. \tag{16}$$

In Theorem 1, we also provide the conditions under which $\Pr(\mathcal{A}_0) \to 1$ as $N, T \to \infty$.

Theorem 1 *Let* x_{it} *for* $i = 1, 2, \cdots, N$ *and* $t = 1, 2, \cdots, T$ *be generated by (2) and* y_t *for* $t = 1, 2, \cdots, T$ *be generated by (9). Moreover let* $T = \ominus(N^{\kappa_1})$ *with* $0 < \kappa_1 \leq 2$*, and* $\mathcal{S}_{Nt} = \{x_{1t}, x_{2t}, \cdots, x_{Nt}\}$ *which contains* k *signals,* k^* *strong-pseudo-signals and* $N - k - k^*$ *semi-noise*

variables. Then under Assumptions 1-6, there exist finite positive constants C_0, and C_1 such that,

(i) the probability that the number of stages in the GOCMT procedure, $\hat{\mathbb{S}}$, given by (14), exceeds k is

$$\Pr\left(\hat{\mathbb{S}} > k\right) = O(N^{4-2C_0\delta}) + O(N^{4-2C_0\delta^*}) + O[\exp(-N^{C_1\kappa_1})], \text{ and} \tag{17}$$

(ii) the probability of selecting the approximating model, $\mathcal{A}_0$, defined by (16), is

$$\Pr\left(\mathcal{A}_0\right) = 1 - O(N^{4-2C_0\delta}) - O(N^{5-2C_0\delta^*}) - O[\exp(-N^{C_1\kappa_1})]. \tag{18}$$

Remark 3 *Based on the results from Theorem 1, for any δ and δ^* greater than $\frac{4}{2C_0}$, the probability of the event that the number of GOCMT stages exceeds the number of signals, $\hat{\mathbb{S}} > k$, goes to zero as $T, N \to \infty$. However, to ensure that the probability of selecting the approximating model tends to one as $T, N \to \infty$, we need $\delta > \frac{4}{2C_0}$ and $\delta^* > \delta > \frac{5}{2C_0}$. Our extensive Monte Carlo Studies suggest that choosing $\delta = 1$ and $\delta^* = 2$ preform well in practice. It is also worth noticing that in Theorem 1, we require T to grow at the order less than N^2. This condition is needed for accurate estimation of latent factors and it is discussed in Remark 1.*

In the rest of this section we focus on coefficients' estimation error and mean square error of the post GOCMT selected model. The model can be written as

$$y_t = \sum_{i=1}^{N} \hat{\mathcal{J}}_i x_{it}\varphi_i + \xi_t \tag{19}$$

where $\hat{\mathcal{J}}_i$ is the selection indicator defined by (13). Also $\sum_{i=1}^{N} \hat{\mathcal{J}}_i = \hat{k}_T$, where $\hat{k}_T$ is the number of covariates selected by GOCMT. By Theorem 1 the probability that the selected model contains the signals tends to unity as $N, T \to \infty$. We can further write

$$y_t = \sum_{i=1}^{N} \hat{\mathcal{J}}_i x_{it}\varphi_i + \xi_t = \sum_{\ell=1}^{\hat{k}_T} w_{\ell t} b_\ell + \xi_t, \tag{20}$$

where $\mathbf{w}_t = \left(w_{1t}, w_{2t}, \cdots, w_{\hat{k}_T t}\right)'$. The least squares (LS) estimator of the selected coefficients, $\mathbf{b}_T = \left(b_1, b_2, \cdots, b_{\hat{k}_T}\right)'$, is given by

$$\hat{\mathbf{b}}_T = \left(T^{-1}\sum_{t=1}^{T} \mathbf{w}_t \mathbf{w}_t'\right)^{-1} \left(T^{-1}\sum_{t=1}^{T} \mathbf{w}_t y_t\right). \tag{21}$$

Moreover, the mean square error (MSE) of the selected model is given by

$$\text{MSE} = T^{-1}\sum_{t=1}^{T} \hat{\xi}_t^2, \tag{22}$$

where $\hat{\xi}_t = y_t - \mathbf{w}_t'\hat{\mathbf{b}}_T$. Theorem 2 shows that the estimation error of the coefficients tends to zero as $N, T \to \infty$. It also establishes the limiting property of MSE of the selected model.

Theorem 2 *Let x_{it} for $i = 1, 2, \cdots, N$ and $t = 1, 2, \cdots, T$ be generated by (2), and y_t for $t = 1, 2, \cdots, T$ be generated by (9). Consider the LS estimator of the selected coefficients and*

the MSE of the selected model given by (21) and (22), respectively. Suppose Assumptions 1-6 holds and $T = \ominus(N^{\kappa_1})$ *with* $0 < \kappa_1 \leq 2$. *Then,*

$$\left\| \hat{\mathbf{b}}_T - \mathbf{b}_T^* \right\|_2 = O_p(N^{-\kappa_1/2}), \tag{23}$$

where $\mathbf{b}_T^* = (b_1^*, b_2^*, \cdots, b_{k_T}^*)'$, *and*

$$\begin{cases} b_\ell^* \in (\beta_1, \beta_2, \cdots, \beta_k)', & \text{if } w_{\ell t} \in (x_{1t}, x_{2t}, \cdots, x_{kt})' \\ b_\ell^* = 0, & \text{otherwise.} \end{cases} \tag{24}$$

Moreover,

$$MSE \equiv T^{-1} \sum_{t=1}^{T} \hat{\xi}^2 = \bar{\sigma}_u^2 + O_p(N^{-\kappa_1/2}), \tag{25}$$

where $\bar{\sigma}_u^2 = T^{-1} \sum_{t=1}^{T} \mathbb{E}(u_t^2)$.

As can be seen in Theorem 2, if $\kappa_1 = 1$ and therefore both T and N grow at the same order, we have the LS estimator of the selected coefficients and the MSE of the selected model approaches to their limiting values at the oracle rate of $\sqrt{T}$. It should also be highlighted that the results of Theorem 2 are built upon the results from Theorem 1 which shows that the GOCMT procedure selects the approximating model with probability tends to one and by assumption the number of the strong-pseudo-signals, k^*, which can be included in the approximating model, is fixed.

5 Monte Carlo Studies

We now investigate the finite-sample performance of GOCMT numerically, using synthetic data sets. We compare GOCMT with OCMT, LASSO, A-LASSO, and IPAD. Monte Carlo (MC) simulation designs and settings are explained in Section 5.1. The implementation of the aforementioned procedures are described in Section 5.2. We discuss the performance evaluation criteria in Section 5.3 and finally summarize the comparative results in Section 5.4.

5.1 Simulation Designs and Settings

The target variable y_t is generated as:

$$y_t = \sum_{i=1}^{k} \beta_i x_{it} + \tau u_t, \quad u_t \sim IIDN(0,1), \text{ for } t = 1, 2, ..., T, \tag{26}$$

where $\beta_i = 1$, $i = 1, 2, ..., k$, and we set $k = 4$. The signals, $(x_{1t}, x_{2t}, \cdots, x_{kt})$, the pseudo-signals, and the noise variables in the active set are generated as

$$x_{it} = \kappa_i(\mu_i \varepsilon_{it} + (1 - \mu_i^2)^{1/2} \boldsymbol{\gamma}_i' \mathbf{f}_t), \tag{27}$$

where $\mathbf{f}_t = (f_{1t}, f_{2t}, ..., f_{m_0 t})'$ is an $m_0 \times 1$ vector of unobserved common factors, and $\boldsymbol{\gamma}_i = (\gamma_{1i}, \gamma_{2i}, ..., \gamma_{m_0 i})'$ is an $m_0 \times 1$ vector of factor loadings. To ensure the factors' share of variation

in x_{it} remains constant, as the number of factors increases, we set $\boldsymbol{\gamma}_i = \frac{\tilde{\boldsymbol{\gamma}}_i}{(\tilde{\boldsymbol{\gamma}}_i'\tilde{\boldsymbol{\gamma}}_i)^{1/2}}$ and generated $\tilde{\boldsymbol{\gamma}}_i$ as describe below. μ_i is defined as a positive constant between zero and one to control the average pair-wise correlation between the k signals and pseudo-signals in the active set. κ_i is the random positive number creating heterogeneity in the variance of x_{it}'s. For $i = 1, 2, \cdots, k$, we set $\kappa_i = 1$, and for the remaining ones, κ_i is independently drawn from a chi-square distribution with 3 degrees of freedom. The variables ε_{it}'s are also generated according to the spatial autoregressive (SAR) model.

$$\varepsilon_{it} = \lambda \sum_{j=1}^{N} \omega_{ij}\varepsilon_{jt} + \pi_i \nu_{it}, \tag{28}$$

where ω_{ij} are the spatial weights and λ is the SAR autocorrelation coefficient that measure the degree of spatial dependence across the idiosyncratic components. π_i is a positive constant, chosen to ensure that ε_{it} has a unit variance. We can write (28) in the following matrix format:

$$\boldsymbol{\varepsilon}_t = (\mathbf{I} - \lambda \mathbf{W})^{-1} \boldsymbol{\Pi} \mathbf{v}_t,$$

where $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \cdots, \varepsilon_{Nt})'$, $\mathbf{W}$ is an $N \times N$ matrix of weights, $\boldsymbol{\Pi} = \text{diag}(\pi_1, \cdots, \pi_N)$, and $\mathbf{v}_t = (\nu_{1t}, \cdots, \nu_{Nt})'$. So $\boldsymbol{\Sigma}_\varepsilon = \mathbf{Q}\boldsymbol{\Pi}^2\mathbf{Q}'$ where $\mathbf{Q} = (\mathbf{I} - \lambda\mathbf{W})^{-1}$, and π_i can be chosen so that the diagonal elements of $\boldsymbol{\Sigma}_\varepsilon$ are equal to one. We set $\lambda = 0.5$ and consider a particular rook type weight matrix in which for all i and j, $\omega_{ij} = 1/2$, if $|i - j| = 1$, and zero otherwise.

In addition, ν_{it}, and $f_{\ell t}$, for $\ell = 1, 2, \cdots, m_0$, are either independent draws from $\mathcal{N}(0, 1)$, or are generated as stationary AR(1) processes:

$$\nu_{it} = \rho_\nu \nu_{i,t-1} + \left(1 - \rho_\nu^2\right)^{1/2} \xi_{it,\varepsilon},\ \xi_{it,\varepsilon} \sim IID\mathcal{N}\left(0, 1\right),$$

$$f_{\ell t} = \rho_f f_{\ell t-1} + \left(1 - \rho_f^2\right)^{1/2} \xi_{\ell t,f},\ \xi_{\ell t,f} \sim IID\mathcal{N}\left(0, 1\right),$$

for all i, ℓ, and for $t = 1, 2, \cdots, T$ with starting values ν_{i0}, and $f_{\ell 0}$ drawn independently from $\mathcal{N}(0, 1)$. We set $\rho_\nu = \rho_f = 0.5$.

We consider the following two types of DGPs.

DGP I (Design with a single unobserved factor): We set $m_0 = 1$ and $\tilde{\gamma}_i = 1$ for $i = 1, 2, \cdots, [N_1^\alpha]$, and zero otherwise, where N_1 is a positive constant between k and N determining the total number of potential pseudo-signals.

DGP II (Design with multiple unobserved factors): We set $m_0 = 2$. $\tilde{\gamma}_{\ell i}$ for $i = 1, 2, \cdots, [N_1^{\alpha_\ell}]$, and $\ell = 1, 2$, are generated from $IIDU[0.5, 1.5]$ and zero otherwise.

Parameter τ in the DGP (26) is chosen so that the R^2 of the regression of y_t on the set of signals is 70%, 50%, 30% and 20%. We consider $\alpha_\ell = \alpha$ for all j, where α is equal to $0, 0.25, 0.45, 0.5, 0.55, 0.75, 0.9$, and 1. We also set $N \in \{100, 300, 1000\}$, $T \in \{100, 300, 500\}$ and $N_1 = 0.75N$. If $\tilde{\gamma}_{\ell i} = 0$, for all ℓ, then μ_i is set equal to one. Otherwise, $\mu_i = \mu$ is chosen to control the average pair-wise correlation between the k signals and pseudo-signals in the active

set. Using (27), the average pair-wise correlation is

$$\bar{\rho} = \frac{\sum_{i=1}^{k}\sum_{j=k+1}^{N_1}\rho_{ij}}{k(N_1-k)} \leq \frac{c+\sum_{i=1}^{k}\sum_{j=k+1}^{N_1}(1-\mu^2)\boldsymbol{\gamma}_i'\boldsymbol{\gamma}_j}{k(N_1-k)}$$
$$\leq \frac{c+(1-\mu^2)kN_1^{\alpha_{\max}}}{k(N_1-k)} = (1-\mu^2)O(N_1^{(\alpha_{\max}-1)}) + o(1),$$

where as before ρ_{ij} is the correlation between x_{it} and x_{jt}, c is a finite positive constant, and $\alpha_{\max} = \max_{\ell=\{1,\cdots,m^0\}}(\alpha_\ell)$. So when $\alpha_{\max}$ is equal to one, the average pair-wise correlation is approximately equal to $(1-\mu^2)$. We consider the values $\mu^2 = \{1/2, 2/3, 3/4\}$. Overall, we perform 384 experiments for all pairs of $(N,T) \in \{(100,100),(100,300),\cdots,(500,1000)\}$.

5.2 Variable Selection Procedures

We implement OCMT, LASSO, and A-LASSO as described in the online MC supplement of Chudik et al. (2018), and IPAD as outlined in Section 2.2 of Fan et al. (2019). In addition, we implement a couple of procedures based on GOCMT. The difference between the two procedures is in how the number of factors, m, is chosen. In the first procedure, GOCMT-m-max, we first use the CD test to determine whether the covariates are highly correlated or not. If the test indicates that the covariates are weakly correlated we set $m = 0$, otherwise, we set $m = m_{\max}$, where $m_{\max} > m_0$. In the second procedure, GOCMT-m-IC, we use Bai and Ng (2002) method to determine the number of factors between $m = 0$ and $m = m_{\max}$. We set $m_{\max} = 4$. The full description of the GOCMT procedures are provided in the online MC supplement.

The coefficients of the following post GOCMT selected model is estimated by LS:

$$y_t = a + \sum_{i\in\hat{S}} b_i x_{it} + e_t,$$

where $\hat{S}$ is the set of selected covariates. The coefficients of the covariates which are not selected by GOCMT are set to zero. We also compute the out of sample prediction of the target variable by

$$\hat{y}_{T+t} = \hat{a}_1 + \textstyle\sum_{i\in\hat{S}} \hat{b}_{i1} x_{it}, \quad \text{for } t = 1,2,\cdots,T_f,$$

or alternatively by

$$\hat{y}_{T+t} = \hat{a}_2 + \hat{\mathbf{f}}_{T+t}\hat{\boldsymbol{\delta}} + \textstyle\sum_{i\in\hat{S}} \hat{b}_{i2} x_{it}, \quad \text{for } t = 1,2,\cdots,T_f,$$

where $\hat{\mathbf{f}}_{T+t} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{V}_{NT}^{-1/2}\hat{\boldsymbol{\gamma}}_i x_{i,T+t}$, $\hat{\boldsymbol{\Gamma}} \equiv (\hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2, \cdots, \hat{\boldsymbol{\gamma}}_N)' = \sqrt{N}\mathbf{Q}$, with $\mathbf{Q}$ be the $N \times m$ matrix of orthonormal eigenvectors corresponding to the m largest eigenvalues of the $N \times N$ matrix, $\mathbf{X}'\mathbf{X}$, and $\mathbf{V}_{NT}$ is the $m \times m$ diagonal matrix of m largest eigenvalues of $(NT)^{-1}\mathbf{X}'\mathbf{X}$.

5.3 Performance Evaluation Criteria

We consider the following criteria to compare finite-sample performance of the aforementioned procedures:

1. Root Mean Square Forecast Error (RMSFE) $= \sqrt{\frac{1}{RT_f}\sum_{r=1}^{R}\sum_{t=1}^{T_f}(y_{T+t}^{(r)} - \hat{y}_{T+t}^{(r)})^2}$,
   where R is the total number of replications, and $y_{T+t}^{(r)}$ and $\hat{y}_{T+t}^{(r)}$ for $t = 1, 2, \cdots, T_f$ are the out of sample realized and predicted values of the target variable at replication r, respectively. We set $R = 1000$ and $T_f = 100$.

2. Root Mean Square Error of Coefficients (RMSE$_\beta$) $= \sqrt{\frac{1}{R}\sum_{r=1}^{R}\sum_{i=1}^{N}\left(\beta_i^0 - \hat{b}_i^{(r)}\right)^2}$,
   where $\beta_i^0 = \beta_i = 1$ for $i = 1, 2, \cdots, k$ and zero otherwise, and $\hat{b}_i^{(r)}$ is its' corresponding estimated value at replication r.

3. False Discovery Rate (FDR) $= \sum_{r=1}^{R}\left(\frac{|\hat{S}^{(r)} \cap S^c|}{|\hat{S}^{(r)}|+1}\right)/R$,
   where $\hat{S}^{(r)}$ is a set of selected covariates at replication r, and S^c is a set of covariates with true zero coefficients.[2]

4. False Positive Rate (FPR)$= \sum_{r=1}^{R}\left(\frac{|\hat{S}^{(r)} \cap S^c|}{N-k}\right)/R$.

5. True Positive Rate (TPR) $= \sum_{r=1}^{R}\left(\frac{|\hat{S}^{(r)} \cap S|}{k}\right)/R$,
   where S is a set of covariates with true non-zero coefficients.

5.4 Simulation Results

For a given value of the degree of cross correlations of the covariates, α, we summarize the MC results in Tables 1 and 2. Table 1 shows the summary statistics averaged across all the experiments. The reported RMSFE and RMSE$_\beta$ in Table 1 are relative to the Oracle procedure in which the signals are known and the model is estimated by LS. Table 2 reports the fraction of times that the variable selection procedures are beaten by LASSO, as the benchmark procedure, across all the experiments. The full set of MC results and the averaged summary statistics for different choices of DGPs, R^2s, μ^2s, Ns and Ts are provided in the online MC supplement.

In Table 1, when the covariates are weakly correlated, $\alpha < 0.5$, the performance of OCMT and GOCMT procedures is fairly similar. But, as the degree of correlation among the covariates becomes stronger, the OCMT procedure starts to select many pseudo-signals and hence its performance deteriorates relative to the GOCMT procedures. As the covariates becomes highly

[2]To ensure the denominator of FDR always has non-zero value even if $|\hat{S}^{(r)}| = 0$, we add plus one to the denominator.

correlated, $\alpha \geq 0.75$, the OCMT procedure fails to perform due to existence of too many pseudo-signals, whilst the suggested GOCMT procedures perform well. When the degree of correlation among the covariates is either very weak or relatively strong, $\alpha \in \{0, 0.25\}$ or $\alpha \in \{0.75, 0.9, 1\}$, GOCMT is the only procedure that on average beats LASSO in terms of RMSE_β, and RMSFE across all the experiments while controls the FDR at the rate lower than 10%. In the cases where $\alpha \in \{0.45, 0.50, 0.55\}$, and hence the unobserved common factors are weak, the GOCMT have FDR as high as 36% simply because the estimation of the unobserved factors is inaccurate. Note that even in these cases, the GOCMT methods still have FDR lower than OCMT, LASSO and A-LASSO, and also have the lowest RMSFE among the procedures.

The results from Table 2 are in line with the that of Table 1. Focusing on RMSFE, GOCMT is outperformed by LASSO in less than 15% of the experiments while A-LASSO and IPAD are beaten by LASSO most of the times. Moreover, when the degree of correlation among the covariates is either very weak or relatively strong, $\alpha \in \{0, 0.25\}$ or $\alpha \in \{0.75, 0.9, 1\}$, GOCMT is never outperformed by LASSO across all the experiments in terms of FDR and FPR, and it is barely beaten by LASSO in terms of RMSE_β, less than 8% of all the experiments.

Overall, the finite sample results show that GOCMT outperforms mainstream variable selection procedures in many cases, and is a valuable extension of OCMT to deal with a high degree of cross-sectional dependence among the variables in the active set.

6 Empirical Analysis

A current topic of research in the empirical asset pricing literature is how to evaluate the importance of hundreds of risk factors that are suggested for explaining risk premia in stock market. To address this issue, Feng et al. (2019) use the double-selection LASSO procedure of Belloni et al. (2014) to measure the contribution of 146 risk factors to asset pricing. In this procedure, to examine the importance of risk factor j in explaining the excess return of stock i, in the first step, a regression model with the excess return of stock i as a dependent variable and the remaining 145 risk factors as explanatory variables is estimated by LASSO. The covariates with corresponding non-zero estimated coefficients are selected. In the next step, the same exercise as in the previous step is repeated, but this time risk factor j is used as the dependent variable instead of the excess return of stock i. The union of selected risk factors from the both steps are considered as control variables when examining the importance of risk factor j in explaining excess return of stock i. As noted earlier, Zhao and Yu (2006) show that for LASSO to only select the signals - the covariates with non-zero marginal effects - the Irrepresentable Condition is required. Generally speaking, this condition requires the

degree of correlation between the signals and the rest of the covariates under consideration to be weak. However, as it is discussed below, the suggested risk factors in asset pricing literature are highly correlated and hence, the Irrepresentable Condition most likely is violated. In this case, the LASSO procedure is prone to the selection of too many noise variables in the second step of the double-selection procedure as control variables. Therefore, the power of the t-test for evaluating the importance of the risk factor decreases and the t-test becomes less efficient. In the rest of this section, we first introduce the return regression equations, and describe the data used. Next, we examine the degree of correlations among the 146 suggested risk factors in the literature. Finally, we apply the GOCMT and LASSO procedures to evaluate the importance of these factors in explaining risk premia in stock market, and compare the results.

Following the literature, we assume that returns on security i, at time t, r_{it}, is generated according to the linear multi-factor model

$$r_{it} - r_{t-1}^f = a_i + \sum_{j=1}^{k} \beta_{ij} g_{jt} + u_{it}, \tag{29}$$

where r_{t-1}^f is the risk free rate; a_i is the intercept in the factor model; g_{jt}, $j = 1, 2, ..., k$ are the relevant risk factors with non-zero associated factor loadings, β_{ij}; and u_{it} is the idiosyncratic component of asset return. Our sample period is from Jan. 1980 to Dec. 2017. We compile 146 monthly risk factors with no missing observations available at Feng et al. (2019). Monthly risk free rate, r_{t-1}^f, is obtained from Fama and French database[3]. We also compile daily close price and monthly dividend, D_{it}, for all the stocks that have been part of Standard and Poor's 500 (S&P500) index between Jan. 1990 and Dec. 2017 from Data Stream. For stock i, the price at the last trading day of each month is used to construct the corresponding monthly stock prices, P_{it}. Finally, monthly stock return is computed by $r_{it} = \frac{P_{it} - P_{i,t-1}}{P_{i,t-1}} + \frac{D_{it}}{P_{i,t-1}}$.

To evaluation the level of correlation among the risk factors over the full sample period, we first compute the proportion of statistically non-zero correlations among them. Note that for $i, j = 1, 2, \cdots, N$, where $N = 146$ is the number of risk factors, under the null hypothesis that $\rho_{ij} \equiv \text{corr}(g_{it}, g_{jt}) = 0$, $\sqrt{T}\hat{\rho}_{ij,T} \sim \mathcal{N}(0, 1)$, where $\hat{\rho}_{ij,T}$ is the estimated correlation coefficient. Therefore, we can compute the proportion of statistically non-zero correlations by $\hat{\pi} = \frac{\sum_{i=1}^{N} \sum_{j>i} \mathbf{I}[\sqrt{T}|\hat{\rho}_{ij,T}| > c_p(n)]}{n}$, where $n = N(N-1)/2$ and $c_p(n)$ is the critical value of the test. To deal with multiple testing problem, we use the Bonferroni correction idea and set $c_p(n) = \Phi^{-1}\left(1 - \frac{p}{2n}\right)$ with $\Phi^{-1}(.)$ being the inverse of a standard normal distribution function and p $(0 < p < 1)$ is the nominal size of the individual tests. The computed proportion is $\hat{\pi} = 0.533$, which means more than half of the elements in the estimated correlation matrix are

[3]Fama and French database is available at `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

significantly different from zero. Next, we use Bailey et al. (2018) procedure to measure the degree of linear dependence among the risk factors by $\hat{\alpha} = 1 + \frac{\log(\hat{\pi})}{\log(n)}$, where $\hat{\alpha}$ lies in the range $0 < \hat{\alpha} \leq 1$. $\hat{\alpha}$ is estimated as 0.933, which suggests a high degree of correlation among the factors. Finally as discussed in Section 3.3, the CD test suggested by Pesaran (2015) can be used to test for the degree of cross correlation among the risk factors. Roughly speaking the CD test which has standard normal distribution under the null hypothesis, allows us to test whether covariates under consideration are highly correlated or not. The computed CD test for weak dependence is 65.70, which differs substantially from the critical value of 1.96. Thus, we can conclude that the risk factors are strongly correlated.

The importance of risk factors in explaining risk premia in stock market can change through time. In particular, some factors might be able to explain risk premia at some points in time but not at others. To deal with this issue, we use 10-year rolling windows of 120 monthly observations to select the relevant risk factors at each point in time. Therefore, the observations for the first 10 years only use for the selection and estimation purposes, and our evaluation period starts from Dec. 1989 (337 rolling windows). For each rolling window, τ, we consider securities $i = 1, 2, \cdots, M_\tau$ that are part of S&P500 index at the end of the rolling window and have the required 120 monthly observations for return. The average, maximum and minimum number of securities across the rolling windows are 443.53, 461 and 412, respectively. Additionally note that to apply the GOMCT procedure, described in Section 3.3, we first need to select the number of common factors among the risk factors. Figure 1 shows the number of selected common factors for each rolling window using the procedure proposed by Onatski (2010) with the maximum possible number of common factors being set equal to five. As can be seen in Figure 1, the number of common factors chosen by the Onatski procedure is mostly equal to one during our sample period while at the end of the sample (between 2013 and 2017) there exists erratic switches between zero and two. To deal with these unusual switches, we set the number of common factors for the GOCMT procedure always equal to one.

Figure 2 shows the average number of risk factors selected by GOCMT and LASSO for each rolling window between Dec. 1989 and Dec. 2017. The figure indicates that on average LASSO selects considerably many more risk factors as compared to GOCMT. Additionally, Figure 3 shows that, most of the time, the number of selected risk factors by LASSO is greater than that of GOCMT in the case of 80 percent of securities considered. Since the 146 risk factors being considered are highly correlated, it is reasonable to expect that many of the risk factors selected by LASSO to be false discoveries.

In Figure 2, we see that among the 146 risk factors under consideration, GOCMT and

LASSO on average select between 3 and 18 risk factors. However, it is not clear that, at each rolling window, whether we have some particular risk factors which are selected for all the securities or whether the selected risk factors varies from one security to another. Note that if a risk factor is strong, we should expect it to explain excess returns for almost all securities. In other words, a risk factor g_{it} is considered as strong risk factor at the rolling window, τ, if its' corresponding factor loadings, β_{ij}'s, take non-zero values for all of the securities $i = 1, 2, \cdots, M_\tau$. We can now go one step further and evaluate the strength of the selected risk factors. Bailey et al. (2019) shows that we can estimate the strength of risk factor j at rolling window τ by $\hat{\delta}_{j\tau} = \frac{\log(\sum_{i=1}^{M_\tau} \hat{\mathcal{J}}_{j,i\tau})}{\log(M_\tau)}$, where $\hat{\mathcal{J}}_{j,i\tau}$ takes the value of one if risk factor j is chosen by a selection procedure for stock i at rolling window τ and zero otherwise. We can further define the average strength of factor j by $\hat{\delta}_j = T^{-1} \sum_{\tau=1}^{T} \hat{\delta}_{j\tau}$, where $T = 337$ is the total number of rolling windows, and divide the factors in four groups of weak ($\delta_j \leq 0.5$), semi-weak ($0.5 < \delta_j \leq 0.75$), semi-strong ($0.75 < \delta_j < 1$), and strong ($\delta_j = 1$). Pesaran and Smith (2021) argue that as the strength of risk factor j; δ_j; decreases, the convergence rate for the corresponding estimated risk premia drops, i.e. the convergence rate for factor j is $M^{\delta_j/2}$, where M is the number of securities used in estimation of risk premia. Note that if a factor is strong ($\delta = 1$), then the estimated risk premia has the conventional convergence rate; $\sqrt{M}$; while if a factor is weak ($\delta \leq 1/2$), the estimated risk premia converge very slowly; less than equal to $M^{1/4}$. To see the intuition, consider the case where we have 500 securities, and a risk factor with strength equal to $\delta = 0.5$. In this case, the risk factor only explains excess returns for less than 25 securities and therefore a large number of securities are needed to have accurate estimation of its' corresponding risk premia. As discussed in Pesaran and Smith (2021), weak factors can be related to pricing error rather than risk premia.

Figure 4 compares the frequency of average estimated strength of the 146 risk factors between GOCMT and LASSO over the sample period. The GOCMT procedure suggests that only less than 10% of the 146 suggested risk factors have the average estimated strength above 0.5. On the other hand, the LASSO procedure, which prone to have high false discoveries, indicates that as many as 65% of the risk factors have the average strength above 0.5. Table 3 shows the top ten risk factors based on GOCMT's estimated factor strength over different sub-periods of 1980s, 1990s, 2000s, 2010s. It also shows the top ten risk factors based on the average estimated factor strength over the full sample period. Notice that Excess Market Return always has the highest estimated strength and its' strength is fairly stable across time, while the estimated strength of other risk factors can change dramatically across time.

In the next step, we assume non-zero coefficient for Excess Market Return in equation (29),

as it always has the highest strength, and implement GOCMT and LASSO conditional on Excess Market Return to select among the remaining risk factors. Figure 5 compares the estimated risk factors' strength between Conditional GOCMT and Conditional LASSO. As can be seen in panel (a) of Figure 5, after controlling for Excess Market Return, the remaining risk factors which previously had estimated strength above 0.5, lose their strength such that there is only one risk factor (Market Beta) which still has the average strength above 0.5 (0.53). In section 4.2, it is shown that the GOCMT procedure will select all signals and strong-pseudo-signals with probability approaching to one. Therefore, the decrease in average strength of these factors after conditioning on Excess Market Return can be interpreted as if these risk factors were playing the role of strong-pseudo-signals. One the other hand, the result from Conditional LASSO still suggests that there exists many risk factors with estimated strength above 0.5. Again, this can be related to the fact that LASSO cannot handle the strong linear dependence among the remaining risk factors and prone to select too many irrelevant risk factors.

Finally, Table 4 reports the name and strength of the ten risk factors with the highest estimated strength by Conditional GOCMT across different sub-periods. As can be seen, relative to the results in Table 3, the estimated strength of the risk factors drops significantly. For instant, when we do not condition on Excess Market Return, during 2000s all the top ten risk factors have estimated strength above 0.8, but once we condition on the Market factor, there exists no risk factors with estimated strength above 0.75, see Tables 3 and 4. These results suggest that Excess Market Return is the only risk factors that can explain risk premia in the stock market and the rest of the factors which happens to be weak or semi-weak, are more related to pricing error rather than risk premia.

7 Conclusion

Current tools for variable selection in high-dimensional settings, like the OCMT and penalized regression methods, require a degree of cross-sectional correlation among the covariates in the active set to be weak. In practice, however, there often exists a strong degree of cross-sectional correlation among the covariates in the selection set. In this paper, we generalize the OCMT procedure proposed by Chudik et al. (2018) to allow for a strong degree of cross-sectional correlation among the covariates in the active set. We refer to the proposed method as GOCMT.

The GOCMT procedure exploits ideas from the latent factors and multiple testing literature to control the probability of selecting the approximating model when the collinearity among the set of covariates under consideration are high. Our theoretical results are valid under general assumptions. Monte Carlo experiments indicate that the newly suggested method have

appealing finite-sample performance relative to competing methods, like LASSO, under many different settings. The benefits of the procedure are also illustrated by an empirical application to selection of risk factors in asset pricing literature. Our analysis indicate that among 146 risk factors available at Feng et al. (2019), only Excess Market Return is strong and can be used to estimate the risk premia. The other risk factors are mostly found to be weak and hence can be related to pricing error rather than risk premia.

In this paper, it is assumed that there exist no dominant units (observable common factor) in the active set. For future research, it would be interesting to relax this assumption and allow for the existence of dominant units in the active set. Moreover, in this work, we assume that the unobserved common factors are strong. For empirical economic application it is also important to allow for weak unobserved common factors affecting both signals and pseudo-signals.

Tables and Figures

Table 1: Average Value of the Selected Statistics across all Experiments for a given Value of α

		OCMT	GOCMT-m-max	GOCMT-m-IC	Lasso	Adaptive Lasso	IPAD	IPAD$_+$
$\alpha = 0$	TPR	0.856	0.854	0.855	0.926	0.918	0.858	0.362
	FPR	0.002	0.002	0.002	0.046	0.04	0.008	0.004
	FDR	0.054	0.051	0.052	0.582	0.555	0.206	0.105
	RMSE$_\beta$	1.902	1.914	1.91	3.756	11.638	5.194	5.369
	RMSFE	1.017	1.018	1.017	1.039	1.09	1.04	1.246
	alt. RMSFE	-	1.018	1.018	-	-	-	-
$\alpha = 0.25$	TPR	0.893	0.891	0.891	0.911	0.896	0.843	0.342
	FPR	0.002	0.002	0.002	0.043	0.038	0.008	0.004
	FDR	0.076	0.074	0.074	0.55	0.524	0.202	0.1
	RMSE$_\beta$	1.655	1.654	1.654	2.984	9.295	4.438	4.396
	RMSFE	1.011	1.011	1.011	1.033	1.082	1.035	1.249
	alt. RMSFE	-	1.012	1.011	-	-	-	-
$\alpha = 0.45$	TPR	0.927	0.92	0.922	0.9	0.877	0.836	0.375
	FPR	0.013	0.011	0.012	0.038	0.033	0.008	0.004
	FDR	0.362	0.331	0.337	0.532	0.505	0.222	0.116
	RMSE$_\beta$	5.337	5.129	5.292	2.795	8.785	4.022	3.589
	RMSFE	1.014	1.014	1.014	1.03	1.076	1.034	1.232
	alt. RMSFE	-	1.014	1.014	-	-	-	-
$\alpha = 0.5$	TPR	0.927	0.912	0.916	0.898	0.874	0.834	0.388
	FPR	0.02	0.016	0.017	0.039	0.034	0.008	0.004
	FDR	0.439	0.362	0.374	0.539	0.511	0.232	0.123
	RMSE$_\beta$	6.608	5.486	5.723	2.851	8.696	4.148	3.606
	RMSFE	1.019	1.016	1.017	1.03	1.076	1.034	1.226
	alt. RMSFE	-	1.017	1.018	-	-	-	-
$\alpha = 0.55$	TPR	0.927	0.898	0.905	0.896	0.871	0.832	0.401
	FPR	0.029	0.019	0.019	0.039	0.034	0.009	0.005
	FDR	0.494	0.335	0.355	0.546	0.515	0.242	0.131
	RMSE$_\beta$	17.307	5.103	5.642	2.86	9.016	4.269	3.625
	RMSFE	1.025	1.018	1.019	1.03	1.076	1.035	1.22
	alt. RMSFE	-	1.019	1.02	-	-	-	-
$\alpha = 0.75$	TPR	-	0.793	0.794	0.884	0.853	0.819	0.453
	FPR	-	0.007	0.007	0.042	0.034	0.011	0.006
	FDR	-	0.095	0.096	0.584	0.541	0.294	0.171
	RMSE$_\beta$	-	2.271	2.289	2.853	7.677	4.309	3.462
	RMSFE	-	1.023	1.023	1.031	1.075	1.038	1.196
	alt. RMSFE	-	1.019	1.018	-	-	-	-
$\alpha = 0.9$	TPR	-	0.75	0.751	0.872	0.834	0.803	0.474
	FPR	-	0.002	0.002	0.045	0.034	0.012	0.007
	FDR	-	0.05	0.05	0.619	0.554	0.329	0.201
	RMSE$_\beta$	-	1.761	1.76	3.079	7.756	4.4	3.593
	RMSFE	-	1.028	1.028	1.032	1.072	1.039	1.185
	alt. RMSFE	-	1.02	1.02	-	-	-	-
$\alpha = 1$	TPR	-	0.736	0.736	0.865	0.818	0.787	0.463
	FPR	-	0.002	0.002	0.047	0.032	0.013	0.008
	FDR	-	0.041	0.041	0.645	0.55	0.335	0.208
	RMSE$_\beta$	-	1.608	1.607	3.24	7.992	4.093	3.589
	RMSFE	-	1.03	1.03	1.032	1.067	1.038	1.187
	alt. RMSFE	-	1.02	1.02	-	-	-	-

Note that the reported RMSFE and RMSE$_\beta$ are relative to the Oracle procedure in which Signals are known and the model is estimated by Least Square (LS).

Table 2: Percentage of All Experiments that the Selection Procedures are beaten by LASSO for a given Value of α

		OCMT	GOCMT-m-max	GOCMT-m-IC	Adaptive LASSO	IPAD	IPAD$_+$
$\alpha = 0$	TPR	59.7	61.3	61.3	65.3	75.7	100
	FPR	0	0	0	0	0	0
	FDR	0	0	0	0	0	0
	RMSE$_\beta$	6.9	6.9	6.9	100	81.5	70.8
	RMSFE	6.9	6.9	6.9	100	54.6	100
	alt. RMSFE	-	7.2	6.9	-	-	-
$\alpha = 0.25$	TPR	30.3	31.7	31.5	76.6	80.1	100
	FPR	0	0	0	0	0	0
	FDR	0	0	0	0	0	0
	RMSE$_\beta$	2.5	2.3	2.3	100	88.4	75.2
	RMSFE	2.8	2.8	2.8	100	74.1	100
	alt. RMSFE	-	3	2.8	-	-	-
$\alpha = 0.45$	TPR	10.4	12.5	12.3	82.9	85	100
	FPR	9.7	5.8	7.4	0	0	0
	FDR	28.7	22.5	22.9	0	0	0
	RMSE$_\beta$	58.6	59	59	100	83.6	69.2
	RMSFE	7.9	3.7	4.6	100	91	100
	alt. RMSFE	-	5.8	6	-	-	-
$\alpha = 0.5$	TPR	9.7	13.2	12	82.4	84.7	100
	FPR	26.2	11.6	14.1	0	0	0
	FDR	33.1	22	23.6	0	0	0
	RMSE$_\beta$	66.2	63.2	64.1	100	84.7	68.5
	RMSFE	19.9	11.1	11.6	100	92.6	100
	alt. RMSFE	-	14.1	15	-	-	-
$\alpha = 0.55$	TPR	9	15.5	14.1	81.9	84.3	100
	FPR	33.8	11.6	14.8	0	0	0
	FDR	54.2	16.4	19.2	0	0	0
	RMSE$_\beta$	70.4	62	63.7	100	88.9	69
	RMSFE	30.6	13.2	15	100	94.2	100
	alt. RMSFE	-	16.9	19.4	-	-	-
$\alpha = 0.75$	TPR	-	51.9	51.9	83.3	84.7	100
	FPR	-	0	0	0	0	0
	FDR	-	0	0	0	0	0
	RMSE$_\beta$	-	22.5	22.7	100	88.7	68.5
	RMSFE	-	21.3	21.3	100	96.3	100
	alt. RMSFE	-	11.8	11.3	-	-	-
$\alpha = 0.9$	TPR	-	61.3	61.3	82.6	86.6	100
	FPR	-	0	0	0	0	0
	FDR	-	0	0	0	0	0
	RMSE$_\beta$	-	5.8	5.8	100	87.7	68.5
	RMSFE	-	26.4	26.4	100	95.6	100
	alt. RMSFE	-	14.6	14.6	-	-	-
$\alpha = 1$	TPR	-	64.6	64.4	84	88.2	100
	FPR	-	0	0	0	0	0
	FDR	-	0	0	0	0	0
	RMSE$_\beta$	-	2.1	2.1	100	80.3	62.3
	RMSFE	-	28.7	28.5	100	79.2	100
	alt. RMSFE	-	17.4	16.9	-	-	-

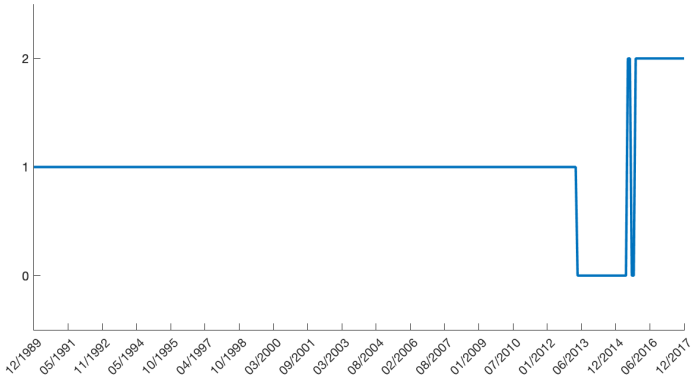


Figure 1: Number of Common Factors Selected by the Onatski(2010) criterion between Dec. 1989 and Dec. 2017

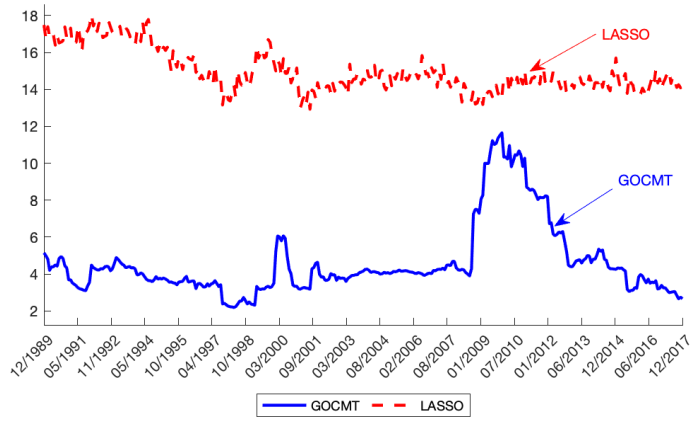


Figure 2: Number of Selected Risk Factors by GOCMT and LASSO between Dec. 1989 and Dec. 2017, Averaged across securities

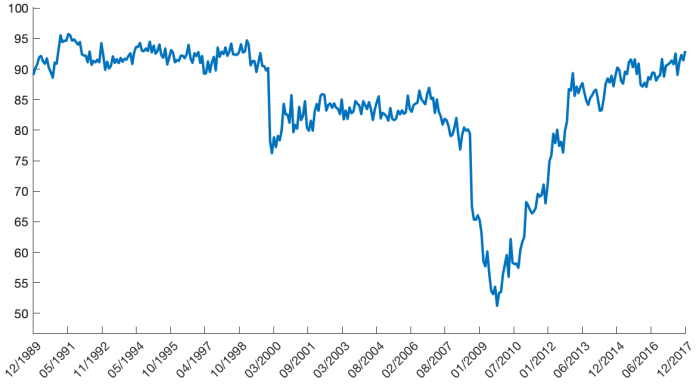
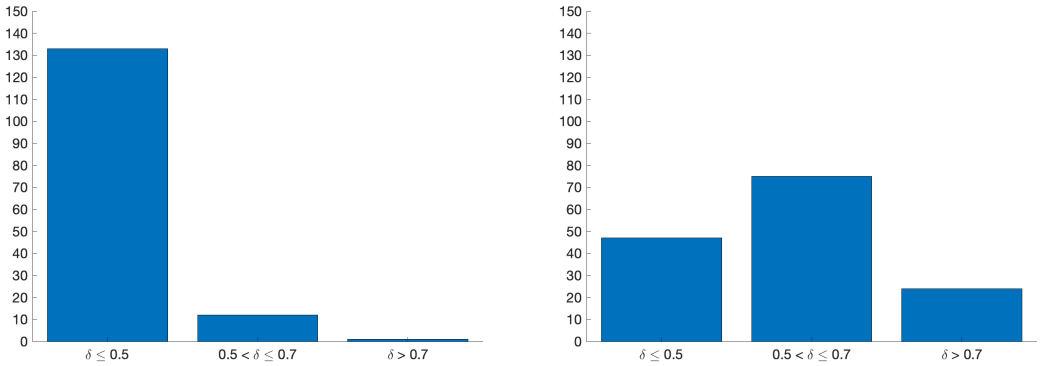


Figure 3: Percentage of securities for which LASSO selects more risk factors compared to GOCMT between Dec. 1989 and Dec. 2017

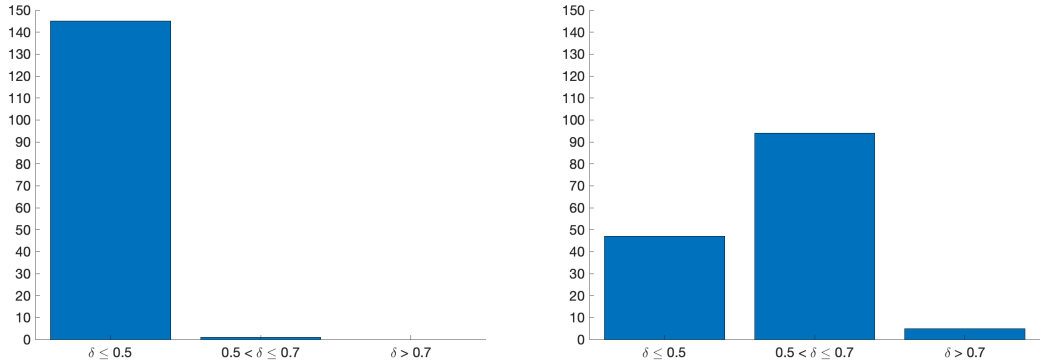


(a) GOCMT (b) LASSO

Figure 4: Frequency of average estimated strength of risk factors

Table 3: Top 10 risk factors based on estimated factor strength by GOCMT over different sub-periods

1980s		1990s	
Excess Market Return	0.96	Excess Market Return	0.87
Betting Against Beta	0.88	Sales to price	0.71
Profit margin	0.76	Market Beta	0.68
Kaplan-Zingales Index	0.74	Sales to inventory	0.67
Leverage	0.74	Change in Net Financial Assets	0.67
Enterprise book-to-price	0.71	Zero trading days	0.61
Sales to price	0.69	Share turnover	0.60
Altman's Z-score	0.69	HML Devil	0.60
Quality Minus Junk	0.69	Quality Minus Junk	0.58
Cash flow to debt	0.67	Industry Concentration	0.58
2000s		**2010s**	
Excess Market Return	0.94	Excess Market Return	0.90
Altman's Z-score	0.89	Sales to price	0.73
Net debt-to-price	0.89	Industry Concentration	0.66
Leverage	0.89	Market Beta	0.63
Market Beta	0.88	years since first Compustat coverage	0.63
Enterprise book-to-price	0.88	HML Devil	0.60
Kaplan-Zingales Index	0.87	High Minus Low	0.58
HML Devil	0.83	Leverage	0.57
Zero trading days	0.82	Altman's Z-score	0.56
Bid-ask spread	0.81	Change in shares outstanding	0.56
Average Estimated Strength Between Jan. 1980 and Dec. 2017			
Excess Market Return	0.90	Kaplan-Zingales Index	0.62
Sales to price	0.67	Market Beta	0.60
Leverage	0.66	HML Devil	0.59
Altman's Z-score	0.66	Quality Minus Junk	0.57
Enterprise book-to-price	0.62	Net debt-to-price	0.55



(a) Conditional GOCMT

(b) Conditional LASSO

Figure 5: Frequency of average estimated strength of risk factors conditional on Excess Market Return

Table 4: Top 10 risk factors based on estimated factor strength by Conditional GOCMT over different sub-periods

1980s		1990s	
Organizational Capital	0.64	Sales to price	0.66
Current ratio	0.62	Market Beta	0.59
Small Minus Big	0.60	Return volatility	0.59
Gross profitability	0.59	Volatility of liquidity (share turnover)	0.57
Profit margin	0.59	HML Devil	0.57
Capital turnover	0.58	Industry Concentration	0.56
Percent Operating Accruals	0.58	Operating Leverage	0.56
Industry-adjusted book to market	0.58	Sales to inventory	0.55
Sales to inventory	0.57	Share turnover	0.54
Operating Leverage	0.57	Small Minus Big	0.53
2000s		**2010s**	
Organizational Capital	0.62	High Minus Low	0.65
Order backlog	0.58	HML Devil	0.62
Book Asset Liquidity	0.57	Market Beta	0.60
Cash flow-to-price	0.56	Bid-ask spread	0.50
Market Beta	0.56	Volatility of liquidity (share turnover)	0.50
Depreciation / PP&E	0.56	Return volatility	0.47
Capital turnover	0.56	Organizational Capital	0.44
Cash flow to price ratio	0.53	Share turnover	0.43
Earnings to price	0.51	1-month momentum	0.42
Operating Leverage	0.51	Zero trading days	0.42
Average Estimated Strength Between Jan. 1980 and Dec. 2017			
Market Beta	0.53	Share turnover	0.46
HML Devil	0.48	Organizational Capital	0.45
Enterprise book-to-price	0.47	Book Asset Liquidity	0.45
Cash flow-to-price	0.47	Operating Leverage	0.45
Sales to price	0.47	Kaplan-Zingales Index	0.43

Appendix

A Mathematical Derivations

The proof of Lemma 1, Proposition 1, and Theorems 1 and 2 are provided here. The proofs are based on lemmas presented in the online theory supplement. Among these, Lemmas S1.1 and S1.14 are key. Lemma S1.1 establishes the exponential probability inequality for the deviation of the rescaled estimated factors from their underlying rotated unobserved factors. The proof of Lemma 1 is built up on this result. For covariates $i = 1, 2, \cdots, N$, Lemma S1.14 establishes the exponential probability inequalities for the t-ratio multiple tests conditional on the underling net effect, θ_i, being either zero or non-zero.

Additional Notations and Definitions

Throughout this section we consider the following events:

$$\mathcal{A}_0 = \mathcal{H} \cap \mathcal{G}, \text{ where } \mathcal{H} = \left\{ \sum_{i=1}^{k} \hat{\mathcal{J}}_i = k \right\} \text{ and } \mathcal{G} = \left\{ \sum_{i=k+k^*+1}^{N} \hat{\mathcal{J}}_i = 0 \right\}. \tag{A.1}$$

$\mathcal{A}_0$ is the event of selecting the approximating model, $\mathcal{H}$ is the event that all signals are selected, and $\mathcal{G}$ is the event that no noise variable is selected. We also denote the event that exactly j noise variables are selected by $\mathcal{G}_j = \left\{ \sum_{i=k+k^*+1}^{N} \hat{\mathcal{J}}_i = j \right\}$, for $j = 0, 1, \cdots, N - k - k^*$, with $\mathcal{G} \equiv \mathcal{G}_0$. For the analysis of different stages of GOCMT, we also introduce the events $\mathcal{B}_{i,s}$ which is the event that variable i is selected up to and including stage s, namely in any of the stages $j = 1, 2, \cdots, s$ of GOCMT procedure, and $\mathcal{L}_s = \cap_{i=1}^{k} \mathcal{L}_{i,s}$ is the event that all the signals are selected up to and including stage s of the GOCMT procedure. $\mathcal{T}_s$ is the event that GOCMT stops after s stages or less. $\mathcal{D}_{s,h}$ is the event that the number of variables selected in the first s stages of GOCMT is smaller than or equal to h, where h is a finite positive constant greater than $k(k + k^*)$.

Proof of Lemma 1

As it shown in Bai and Ng (2002), $\tilde{\mathbf{F}} = \hat{\mathbf{F}}\mathbf{V}_{NT}^{-1}$, *where* $\hat{\mathbf{F}} = (\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, \cdots, \hat{\mathbf{f}}_T)'$ *is the* $T \times m_0$ *matrix of rescale estimator of the factors given by* $\hat{\mathbf{F}} = \bar{\mathbf{F}} \left(\bar{\mathbf{F}}'\bar{\mathbf{F}}/T \right)^{1/2}$. $\bar{\mathbf{F}}$ *is the* $T \times m_0$ *matrix of estimated factors given by*

$$\bar{\mathbf{F}} = \mathbf{X}\bar{\mathbf{\Gamma}}/N,$$

where $\bar{\mathbf{\Gamma}} = \sqrt{N}\bar{\mathbf{Q}}$ *with* $\bar{\mathbf{Q}}$ *is an* $N \times m_0$ *matrix of orthonormal eigenvectors corresponding to the* m_0 *largest eigenvalues of the* $N \times N$ *matrix,* $\mathbf{X}'\mathbf{X}$*. Hence,*

$$\tilde{\mathbf{F}} - \mathbf{F}^0\mathbf{G} = (\hat{\mathbf{F}} - \mathbf{F}^0\mathbf{H})\mathbf{V}_{NT}^{-1},$$

where $\mathbf{H} = \left(\mathbf{\Gamma}^{0\prime}\mathbf{\Gamma}^0/N\right)\left(\mathbf{F}^{0\prime}\tilde{\mathbf{F}}/T\right)$*. Therefore, by Lemma S2.10*

$$\|\tilde{\mathbf{F}} - \mathbf{F}^0\mathbf{G}\|_F^2 \leq \|\hat{\mathbf{F}} - \mathbf{F}^0\mathbf{H}\|_F^2 \|\mathbf{V}_{NT}^{-1}\|_2^2 = \left(\sum_{t=1}^{T} \|\tilde{\mathbf{f}}_t - \mathbf{H}'\mathbf{f}_t^0\|_F^2\right) \|\mathbf{V}_{NT}^{-1}\|_2^2.$$

Let $v_{\min}$ *denote the smallest eigenvalue of* $\mathbf{V}_{NT}$*. Since* m_0 *largest eigenvalues of* $\mathbf{X}'\mathbf{X}/(NT)$ *are bounded away from zero, there exists a positive constant* $\underline{v}$ *such that* $0 < \underline{v} < v_{\min}$*. Therefore,* $\|\mathbf{V}_{NT}^{-1}\|_2^2 = 1/v_{\min} > 1/\underline{v}$ *and we have*

$$\|\tilde{\mathbf{F}} - \mathbf{F}^0\mathbf{G}\|_F^2 \leq \left(\sum_{t=1}^{T} \|\tilde{\mathbf{f}}_t - \mathbf{H}'\mathbf{f}_t^0\|_F^2\right) \|\bar{\mathbf{V}}_{NT}^{-1}\|_2^2 \leq \frac{1}{\underline{v}} \left(\sum_{t=1}^{T} \|\tilde{\mathbf{f}}_t - \mathbf{H}'\mathbf{f}_t^0\|_F^2\right).$$

Hence,

$$\begin{aligned}\Pr\left(\|\tilde{\mathbf{F}} - \mathbf{F}^0\mathbf{G}\|_F^2 > d_T\right) &\leq \Pr\left(\frac{1}{\underline{v}} \sum_{t=1}^{T} \|\tilde{\mathbf{f}}_t - \mathbf{H}'\mathbf{f}_t^0\|_F^2 > d_T\right) \\ &= \Pr\left(\sum_{t=1}^{T} \|\tilde{\mathbf{f}}_t - \mathbf{H}'\mathbf{f}_t^0\|_F^2 > \underline{v} d_T\right),\end{aligned}$$

and by Lemma S1.1, there exists a finite positive constant C_0*, such that if* $\lambda \leq (s+2)/(s+4)$*,*

$$\Pr\left(\|\tilde{\mathbf{F}} - \mathbf{F}^0\mathbf{G}\|_F^2 > d_T\right) \leq N^2 \exp\left[-C_0 T^{-1} d_T^2\right],$$

if $(s+2)/(s+4) < \lambda < 1$*,*

$$\Pr\left(\|\tilde{\mathbf{F}} - \mathbf{F}^0\mathbf{G}\|_F^2 > d_T\right) \leq N^2 \exp\left[-C_0 d_T^{s/(s+2)}\right],$$

and if $\lambda \geq 1$*,*

$$\Pr\left(\|\tilde{\mathbf{F}} - \mathbf{F}^0\mathbf{G}\|_F^2 > d_T\right) \leq N^2 \exp\left[-C_0 (T d_T)^{s/2(s+2)}\right].$$

Proof of Proposition 1

Note that $\mathbb{S}_0$ *is the population value of number of stages required for selecting all the signals in which it is assumed that* $\Pr[t_{i,T,(j)} > c_p(N,\delta) | \theta_{i,(j)} \neq 0] = 1$ *and* $\Pr[t_{i,T,(j)} > c_p(N,\delta) | \theta_{i,(j)} = 0] = 0$ *for all* i *and* j*. So, if* $\theta_{i,(1)} \neq 0$ *for all* i *with* $\boldsymbol{\beta}_i \neq 0$*, it obviously follows that* $\mathbb{S}_0 = 1$*. Next, assume that* $\theta_{i,(1)} = 0$ *for a non-empty subset of signals. Then these signals will not be selected in the first stage. By Lemma S1.19, it follows that at least for one signal* $\theta_{i,(1)} \neq 0$ *and therefore this signal will be picked up in the first stage. Similarly, By Lemma S1.19, in the next stage at least one hidden signal, for which* $\theta_{i,(1)} = 0$*, will have* $\theta_{i,(2)} \neq 0$ *and hence will picked up in this stage. Proceeding recursively using Lemma S1.19, it then follows that all hidden signals, for which* $\theta_{i,(1)} = 0$*, will satisfy* $\theta_{i,(j)} \neq 0$ *for some* $j \leq k$*, proving the proposition.*

Proof of Theorem 1

To establish result (17), note that $\mathcal{T}_k$ is the event that the GOCMT procedure stops after k stages or less. Therefore $\Pr(\hat{\mathbb{S}} > k) = \Pr(\mathcal{T}_k^c) = 1 - \Pr(\mathcal{T}_k^c)$*, where* $\hat{\mathbb{S}}$ *is defined by (14). By Lemma S1.17, there exist positive constants* C_0*,* C_1 *and* C_2 *such that*

$$\Pr(\mathcal{T}_k) = 1 - O(N^{4-2C_0\delta}) - O(N^{4-2C_0\delta^*}) - O[N\exp(-C_1N^{C_2\kappa_1})],$$

and hence

$$\Pr\left(\hat{\mathbb{S}} > k\right) = O(N^{4-2C_0\delta}) + O(N^{4-2C_0\delta^*}) + O[N\exp(-C_1N^{C_2\kappa_1})],$$

To establish result (18), first note that

$$\Pr(\mathcal{A}_0^c) = \Pr(\mathcal{A}_0^c|\mathcal{D}_{k,h})\Pr(\mathcal{D}_{k,h}) + \Pr(\mathcal{A}_0^c|\mathcal{D}_{k,h}^c)\Pr(\mathcal{D}_{k,h}^c) \leq \Pr(\mathcal{A}_0^c|\mathcal{D}_{k,h}) + \Pr(\mathcal{D}_{k,h}^c).$$

By Lemma S1.16, for some finite positive constants C_0*,* C_1 *and* C_2

$$\Pr(\mathcal{D}_{k,h}^c) = O(N^{4-2C_0\delta}) + O(N^{4-2C_0\delta^*}) + O[N\exp(-C_1N^{C_2\kappa_1})].$$

Moreover,

$$\Pr(\mathcal{A}_0^c|\mathcal{D}_{k,h}) \leq \Pr(\mathcal{H}^c|\mathcal{D}_{k,h}) + \Pr(\mathcal{G}^c|\mathcal{D}_{k,h}),$$

where $\mathcal{H}$ *and* $\mathcal{G}$ *is given by (A.1). Therefore,* $\mathcal{H}^c = \{\sum_{i=1}^k \hat{\mathcal{J}}_i < k\}$ *and* $\mathcal{G}^c = \{\sum_{i=k+k^*+1}^N \hat{\mathcal{J}}_i > 0\}$*. Let's consider* $\Pr(\mathcal{H}^c|\mathcal{D}_{k,h})$ *and* $\Pr(\mathcal{G}^c|\mathcal{D}_{k,h})$ *in turn:*

$$\Pr(\mathcal{H}^c|\mathcal{D}_{k,h}) \leq \sum_{i=1}^k \Pr(\hat{\mathcal{J}}_i = 0|\mathcal{D}_{k,h}).$$

but the event $\{\hat{\mathcal{J}}_i = 0|\mathcal{D}_{k,h}\}$ *can only occur only if* $\{\cap_{j=1}^k \mathcal{B}_{i,j}^c|\mathcal{D}_{k,h}\}$ *occurs, while* $\{\cap_{j=1}^k \mathcal{B}_{i,j}^c|\mathcal{D}_{k,h}\}$ *can occur without* $\{\hat{\mathcal{J}}_i = 0|\mathcal{D}_{k,h}\}$ *occurring. Therefore,* $\Pr(\hat{\mathcal{J}}_i = 0|\mathcal{D}_{k,h}) \leq \Pr(\cap_{j=1}^k \mathcal{B}_{i,j}^c|\mathcal{D}_{k,h})$*. Note that*

$$\Pr(\cap_{j=1}^k \mathcal{B}_{i,j}^c|\mathcal{D}_{k,h}) = \Pr(\mathcal{B}_{i,1}^c|\mathcal{D}_{k,h}) \times \Pr(\mathcal{B}_{i,2}^c|\mathcal{B}_{i,1}^c, \mathcal{D}_{k,h}) \times \Pr(\mathcal{B}_{i,3}^c|\mathcal{B}_{i,2}^c \cap \mathcal{B}_{i,1}^c, \mathcal{D}_{k,h}) \times \cdots \times$$
$$\Pr(\mathcal{B}_{i,k}^c|\mathcal{B}_{i,k-1}^c \cap \cdots \cap \mathcal{B}_{i,1}^c, \mathcal{D}_{k,h}).$$

By Proposition 1, we know that for each $i = 1, \cdots, k$ *there exist some steps* $1 \leq j \leq k$ *such that* $\theta_{i,(j)} \neq 0$*. Therefore, for such the j,*

$$\Pr(\mathcal{B}_{i,j}^c|\mathcal{B}_{i,j-1}^c \cap \cdots \cap \mathcal{B}_{i,1}^c, \mathcal{D}_{k,h}) = \Pr(\mathcal{B}_{i,j}^c|\mathcal{B}_{i,j-1}^c \cap \cdots \cap \mathcal{B}_{i,1}^c, \theta_{i,(j)} \neq 0, \mathcal{D}_{k,h}),$$

and by Lemma S1.14,

$$\Pr(\mathcal{B}_{i,j}^c|\mathcal{B}_{i,j-1}^c \cap \cdots \cap \mathcal{B}_{i,1}^c, \theta_{i,(j)} \neq 0, \mathcal{D}_{k,h}) = O[\exp(-C_0N^{C_2\kappa_1})].$$

Hence,

$$\Pr(\cap_{j=1}^k \mathcal{B}_{i,j}^c|\mathcal{D}_{k,h}) \leq \Pr(\mathcal{B}_{i,j}^c|\mathcal{B}_{i,j-1}^c \cap \cdots \cap \mathcal{B}_{i,1}^c, \theta_{i,(j)} \neq 0, \mathcal{D}_{k,h}) = O[\exp(-C_0N^{C_2\kappa_1})].$$

Therefore,

$$\Pr(\hat{\mathcal{J}}_i = 0|\mathcal{D}_{k,h}) = O[\exp(-C_0N^{C_2\kappa_1})],$$

and since k is finite, we have,

$$\Pr(\mathcal{H}^c|\mathcal{D}_{k,h}) = O[\exp(-C_0 N^{C_2\kappa_1}].$$

For $\Pr(\mathcal{G}^c|\mathcal{D}_{k,h})$ *we first note that,*

$$\Pr(\mathcal{G}^c|\mathcal{D}_{k,h}) = \Pr\left(\sum_{i=k+k^*+1}^{N} \hat{\mathcal{J}}_i > 0|\mathcal{D}_{k,h}\right) \leq \sum_{i=k+k^*+1}^{N} \Pr\left(\hat{\mathcal{J}}_i = 1|\mathcal{D}_{k,h}\right).$$

Also,

$$\begin{aligned}\Pr\left(\hat{\mathcal{J}}_i = 1|\mathcal{D}_{k,h}\right) &= \Pr\left(\hat{\mathcal{J}}_i = 1|\mathcal{D}_{k,h}, \mathcal{T}_k\right) Pr(\mathcal{T}_k|\mathcal{D}_{k,h}) + \Pr\left(\hat{\mathcal{J}}_i = 1|\mathcal{D}_{k,T}, \mathcal{T}_k^c\right) Pr(\mathcal{T}_k^c|\mathcal{D}_{k,h}) \\ &\leq \Pr\left(\hat{\mathcal{J}}_i = 1|\mathcal{D}_{k,h}, \mathcal{T}_k\right) + Pr(\mathcal{T}_k^c|\mathcal{D}_{k,h}).\end{aligned}$$

Therefore,

$$\Pr(\mathcal{G}^c|\mathcal{D}_{k,h}) \leq \sum_{i=k+k^*+1}^{N} \Pr\left(\hat{\mathcal{J}}_i = 1|\mathcal{D}_{k,h}, \mathcal{T}_k\right) + (N-k-k^*)Pr(\mathcal{T}_k^c|\mathcal{D}_{k,h}).$$

Consider now the first term of the above and note that, since the net effect coefficient, $\theta_{i,(j)}$*, of noise variables are zero for* $i = k+k^*+1, k+k^*+2, \cdots, N$ *and all* j*, we have*

$$\begin{aligned}&\sum_{i=k+k^*+1}^{N} \Pr\left(\hat{\mathcal{J}}_i = 1|\mathcal{D}_{k,h}, \mathcal{T}_k\right) \\ &\quad\leq \sum_{i=k+k^*+1}^{N} \Pr(t_{i,T,(1)} > c_p(N,\delta)|\theta_{i,(1)} = 0, \mathcal{D}_{k,h}, \mathcal{T}_k) + \\ &\qquad \sum_{i=k+k^*+1}^{N} \sum_{s=2}^{k} \Pr(t_{i,T,(s)} > c_p(N,\delta^*)|\theta_{i,(s)} = 0, \mathcal{D}_{k,h}, \mathcal{T}_k).\end{aligned}$$

Therefore, by Lemma S1.14, and result (II) of Lemma S2.2, we have

$$\begin{aligned}&\sum_{i=k+k^*+1}^{N} \Pr\left(\hat{\mathcal{J}}_i = 1|\mathcal{D}_{k,h}, \mathcal{T}_k\right) \\ &\quad\leq (N-k-k^*)N^3 \exp(-C_0 c_p^2(N,\delta)) + \\ &\qquad (k-1)(N-k-k^*)N^3 \exp(-C_0 c_p^2(N,\delta^*)) + O[N\exp(-C_0 N^{C_2\kappa_1})] \\ &\quad= O(N^{4-2C_0\delta}) + O(N^{4-2C_0\delta^*}) + O[N\exp(-C_0 N^{C_2\kappa_1})].\end{aligned}$$

Furthermore, by (S.1),

$$(N-k-k^*)Pr(\mathcal{T}_k^c|\mathcal{D}_{k,h}) = O(N^{5-2C_0\delta^*}) + O[N^2\exp(-C_1 N^{C_2\kappa_1})].$$

So, overall,

$$\Pr(\mathcal{G}^c|\mathcal{D}_{k,h}) = O(N^{4-2C_0\delta}) + O(N^{5-2C_0\delta^*}) + O[N^2\exp(-C_1 N^{C_2\kappa_1})].$$

where we used that $O[N\exp(-C_1 T^{C_2})]$ *is dominated by* $O[N^2\exp(-C_1 T^{C_2})]$*, and* $O(N^{4-2C_0\delta^*})$ *is dominated by* $O(N^{5-2C_0\delta^*})$*. Substituting* $\Pr(\mathcal{H}^c|\mathcal{D}_{k,h})$ *and* $\Pr(\mathcal{G}^c|\mathcal{D}_{k,h})$ *in* $\Pr(\mathcal{A}_0^c|\mathcal{D}_{k,h})$*, and using* $\Pr(\mathcal{D}_{k,h}^c)$ *we obtain*

$$\Pr(\mathcal{A}_0^c) = O(N^{4-2C_0\delta}) + O(N^{5-2C_0\delta^*}) + O[N^2\exp(-C_1 N^{C_2\kappa_1})],$$

and therefore,

$$\Pr(\mathcal{A}_0) = 1 - O(N^{4-2C_0\delta}) - O(N^{5-2C_0\delta^*}) - O[\exp(-N^{C_1\kappa_1})].$$

Proof of Theorem 2

For any $B > 0$,

$$\Pr\left(T^{\frac{1}{2}}\left\|\hat{\mathbf{b}}_T - \mathbf{b}_T^*\right\| > B\right) = \Pr\left(T^{\frac{1}{2}}\left\|\hat{\mathbf{b}}_T - \mathbf{b}_T^*\right\| > B|\mathcal{A}_0\right)\Pr\left(\mathcal{A}_0\right) + \Pr\left(T^{\frac{1}{2}}\left\|\hat{\mathbf{b}}_T - \mathbf{b}_T^*\right\| > B|\mathcal{A}_0^c\right)\Pr\left(\mathcal{A}_0^c\right).$$

Since $\Pr\left(T^{\frac{1}{2}}\left\|\hat{\mathbf{b}}_T - \mathbf{b}_T^*\right\| > B|\mathcal{A}_0^c\right)$ *and* $\Pr\left(\mathcal{A}_0\right)$ *are less than or equal to one, we can further write,*

$$\Pr\left(T^{\frac{1}{2}}\left\|\hat{\mathbf{b}}_T - \mathbf{b}_T^*\right\| > B\right) \leq \Pr\left(T^{\frac{1}{2}}\left\|\hat{\mathbf{b}}_T - \mathbf{b}_T^*\right\| > B|\mathcal{A}_0\right) + \Pr\left(\mathcal{A}_0^c\right).$$

By conditioning on $\mathcal{A}_0$ *the dimension of vector* $\hat{\mathbf{b}}_T$ *is at most equal to* $k + k^*$ *and hence it is finite. Therefore, by Lemma S1.18 in online theory supplement, conditional on* $\mathcal{A}_0$, $\left\|\hat{\mathbf{b}}_T - \mathbf{b}_T^*\right\|$ *is* $O_p\left(T^{-\frac{1}{2}}\right)$. *By Theorem 1, we also have* $\lim_{T\to\infty}\Pr\left(\mathcal{A}_0^c\right) = 0$. *Hence, for any* $\varepsilon > 0$, *there exist* $B_\varepsilon > 0$ *and* $T_\varepsilon > 0$ *such that*

$$\Pr\left(T^{\frac{1}{2}}\left\|\hat{\mathbf{b}}_T - \mathbf{b}_T^*\right\| > B_\varepsilon|\mathcal{A}_0\right) + \Pr\left(\mathcal{A}_0^c\right) < \varepsilon \text{ for all } T > T_\varepsilon,$$

Therefore, $\Pr\left(T^{\frac{1}{2}}\left\|\hat{\mathbf{b}}_T - \mathbf{b}_T^*\right\| > B_\varepsilon\right) < \varepsilon$ *for all* $T > T_\varepsilon$, *and we conclude that*

$$\left\|\hat{\mathbf{b}}_T - \mathbf{b}_T^*\right\| = O_P\left(T^{-\frac{1}{2}}\right).$$

Since $T = \ominus(N^{\kappa_1})$ *for* $0 < \kappa_1 \leq 2$, *we can further write*

$$\left\|\hat{\mathbf{b}}_T - \mathbf{b}_T^*\right\| = O_P\left(N^{-\frac{\kappa_1}{2}}\right).$$

as required. Following a similar line of argument, we can also show that

$$T^{-1}\sum_{t=1}^{T}\hat{\xi}_t^2 - \bar{\sigma}_{u,T}^2 = O_p\left(N^{-\frac{\kappa_1}{2}}\right),$$

which completes the proof.

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2008). Large dimensional factor analysis. *Foundations and Trends® in Econometrics*, 3(2):89–163.

Bailey, N., Kapetanios, G., and Pesaran, H. (2019). Measurement of factor strength: Theory and practice. *unpublished manuscript*.

Bailey, N., Kapetanios, G., and Pesaran, M. H. (2016). Exponent of cross-sectional dependence: estimation and inference. *Journal of Applied Econometrics*, 31(6):929–960.

Bailey, N., Kapetanios, G., and Pesaran, M. H. (2018). Exponent of cross-sectional dependence for residuals. *Sankhya B*, pages 1–57.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Buhlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics*, 34(2):559–583.

Chudik, A., Kapetanios, G., and Pesaran, M. H. (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica*, 86(4):1479–1512.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, Y. and Lv, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association*, 108(503):1044–1061.

Fan, Y., Lv, J., Sharifvaghefi, M., and Uematsu, Y. (2019). Ipad: stable interpretable forecasting with knockoffs inference. *Journal of American Statistical Association to appear*.

Feng, G., Giglio, S., and Xiu, D. (2019). Taming the factor zoo: A test of new factors. *Journal of Finance, forthcoming*.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Annals of statistics*, 28(2):337–374.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 29:1189–1232.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.

Pesaran, H. and Smith, R. (2021). Factor strengths, pricing errors, and estimation of risk premia. *working paper*.

Pesaran, M. H. (2015). Testing weak cross-sectional dependence in large panels. *Econometric Reviews*, 34(6-10):1089–1117.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

Zheng, Z., Fan, Y., and Lv, J. (2014). High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):627–649.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.