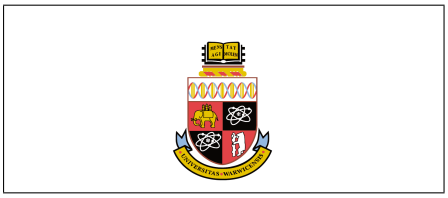
Weak factor model in large dimension

by

Quang Phan

March 21, 2015



Thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Warwick

Department of Economics

Contents

List of Figures

List of Tables

Acknowledgment

Completing the PhD is indeed a fascinating journey. Looking back at the whole period, I have realised many parts of me have changed in a positive way from the date I started the journey. Days and nights of thinking about my research questions and sitting in front of a screen writing up my solutions for these open-ended questions surely will have big impacts for my future career as a researcher.

This completion will start a new chapter of my life and I am obliged to express my special thank to all the people who have been by my side and supported me in this fascinating journey.

First of all, I can not express how grateful I am to my supervisor, Prof. Corradi, for her great support throughout my whole PhD period. She is the best mentor that I could ever ask for, both academically and personally. None of the results in this thesis would happen without her great comments and feedback.

Also, I really appreciate the Department of Economics at Warwick for letting me into the PhD program and provides very generous supports in my 4 years completing this thesis.

Some results in this thesis are obtained based on some advices and comments from Yuan Liao, Chris Heaton, Mike Pitt, and my brother-in-law Long Tran Thanh. Thank you for bringing such insightful and valuable ideas that contribute toward the results in here.

To my parents, grandmother, sister and the rest of my family, I owe them much for their unconditional support, love and encouragement. Last but not least, to my

girlfriend, Chi, who inspired and gave me strengths to keep going when I almost lost hope for my work. Thank you all for having faith in me throughout these challenging times of my life. This thesis would never have been written without their love and support. To them I dedicate this thesis.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. All of the materials have not been published.

Abstract

This thesis presents some extensions to the current literature in high-dimensional static factor models. When the cross-section dimension (represented by N henceforth) is very large, the standard assumption for each common factor is to have the number of non-zero loadings grow linearly with N. On the other hand, an idiosyncratic error for each component can only be correlated with a finite number of other components in the cross-section. These two assumptions are crucial in standard high-dimensional factor analysis, as they allow us to obtain consistent estimators for the factors, the loadings and the number of factors. However, together they rule out the possibility that we may have some factors that have strictly less than N but still non-negligible number of non-zero loadings, e.g. N^α for some $0 < \alpha < 1$. The existence of these weak factors will decrease the signal-to-noise ratio as now the gap between the systematic and idiosyncratic eigenvalues is more narrow. As the consequence, in such model it is harder to establish the consistency of the factors estimated by sample principle components. Furthermore, the number of factors is even more challenging to identify because most existing methods rely on the large signal-to-noise ratio. In this thesis, I consider a factor model that allows general strength for each factor, i.e. both strong and weak factors can exist. Chapter 1 gives more discussions about the current literature on this and the motivation for my contribution.

In Chapter 2, I show that the sample principle components are still the consistent estimators for the factors (up to the spanning space), provided that the factors are

not too weak. In addition, I derive the lower bound that the strength of the weakest factor needs to achieve for being consistently estimated. More precisely, what I mean by strength is the order of the number of non-zero loadings of the factor.

Chapter 3 presents a novel method to determine the number of factors, which is asymptotically consistent even when the factors are weak. I run extensive Monte Carlo simulations to compare the performance of this method to the two well-known ones, i.e. the class of criteria proposed in Bai and Ng (2002) and the eigenvalue ratio method in Ahn and Horenstein (2013).

In Chapter 4 and 5, I show some applications that are based on the work of this thesis. I mainly focus on two issues: selecting the factor models in practice and using factor analysis to compute the large static covariance matrix.

1 Introduction

Factor analysis first arises in the field of Psychometrics, when Spearman (1904) obtained results of several tests taken by schoolchildren and proposed that the correlations between those tests were due to a single factor, which he referred to as intelligence. Since then, there has been a rapid growth in applications of factor analysis in social science, particularly in Finance and Economics. It is very useful and interesting to find a small number of factors (either observed or unobserved) that capture the movements of a much larger number of variables. For examples, Boivin et al. (2013) address a strong factor, can be interpreted as credit shock, which has big impacts on several other financial and economic variables such as credit spreads, interest rates, etc. Additionally, from the statistical angle identifying the common factors brings a great advantage of dimension reduction in the large-dimensional setting.

In this thesis, I focus on the case where factor model is used as a dimension reduction technique. For example, in some applications such as estimating large covariance matrix or forecasting with many explanatory variable, the factors after extracted are used in place of the original components. Therefore, this gains benefit of reducing the dimension significantly.

In brief, this thesis presents some theoretical extensions to the current literature in factor analysis. Particularly, I replace the strongly pervasive factors condition with a less restrictive one that allows the factors to affect a relatively small but non-negligible number of components. This replacement eases the requirement for the

consistency of the factors estimated by the standard principle components technique. In addition, changing this assumption also has some impacts for other areas of research regarding factor analysis, such as determining the number of factors and the estimation of covariance matrix using factor analysis. Therefore, other contributions in this thesis are about determining the number of factors and applications of factor analysis in computing the large covariance matrix.

The main contributions here belong to the theory of Econometrics, rather than Economics empirical findings. Therefore, discussions and application of the common factors identification are mainly approached from a statistical point of view. I do not focus on the case where there is a need to interpret the meaning of the underlying factor processes.

In contrast, there are many other empirical works exploiting factor analysis and interpret the factors as some meaningful variables for insights. Examples within this line of research including studies regarding identifying the factors (or shocks) in yield curve (Diebold et al. (2006)), stock returns (Fama and French (1993)), credit market (Boivin et al. (2013), Gilchrist et al. (2009), etc.), credit default swaps (Chen and Härdle (2012)), corporate bond spreads (Elton et al. (2001)), etc. Nevertheless, the centre of discussion in this thesis regarding general factor identification issues in large-dimensional setting, rather than these financial and economic applications.

Over the next few sections in this chapter, I will gradually discuss some recent relevant advances in factor analysis. Also, some applications are mentioned to illustrate how this can be used in practice.

1.1 Literature review

Parallel to the practical aspects, theoretical research regarding factor model is comparably active, and will be reviewed in section 1.1.1. On the other hand, some well-known applications are reviewed in section 1.1.2.

1.1.1 Developments in factor analysis

Since the literature is extremely large, it is impossible to present all the important related works in the review, hence there are many significant results missing in these subsequent sections. For example, I will not discuss dynamic factor model in details despite its importance, because static factor is the main focus of this thesis. In contrast, some results in the large covariance matrix estimation will be mentioned, due to its link with the main contribution.

1.1.1.1 Overview of factor analysis

In particular, a static factor model for y_{it}, $i = 1, ..., N$ is given by:

$$y_{it} = \lambda_i^{(1)} f_t^{(1)} + ...\lambda_i^{(r)} f_t^{(r)} + u_{it}. \tag{1.1}$$

or

$$y_{it} = \alpha + \lambda_i' f_t + u_{it}. \tag{1.2}$$

where $\lambda_i = [\lambda_i^{(1)}, ..., \lambda_i^{(r)}]'$ is the factor loadings vector for component y_{it}, $f_t = [f_t^{(1)}, ..., f_t^{(r)}]'$ is the common factors vector, u_{it} is the idiosyncratic error (shock) which is not explained by the common factors. The λ_i term corresponds to the exposure of y_{it} to the common factors f_t. In vector form, we can write:

$$\begin{gathered} Y_t = \Lambda f_t + u_t. \\ (N \times 1) = (N \times r)(r \times 1) + (N \times 1) \end{gathered} \tag{1.3}$$

In here, $Y_t = [y_{1t}, ..., y_{Nt}]$, $\Lambda = [\lambda_1, ..., \lambda_N]$ is the matrix of factor loadings and u_t is the vector of idiosyncratic errors. W.l.o.g we assume that Y_t, f_t and u_t all have means 0. In matrix form, given that the length of the time dimension is T we will denote $Y = [Y_1', ..., Y_T']$, $F = [f_1', ..., f_T']$, and $U = [u_1', ..., u_T']$. Hence, (1.3) can also

be written as:

$$Y = F\Lambda' + U.$$

$$(T \times N) = (T \times r)(r \times N) + (T \times N)$$

Recently there has been a rapid growth in applications of factor analysis for social science, particularly in Economics and Finance. This is due to the need to seek for a small set of factors that can contain a large proportion of information from the vast original multivariate series. Some well-known examples of factor model in economic theory are the capital asset pricing model (CAPM, Sharp (1964)) and the arbitrage pricing theory (APT, Ross (1976)).

The factor model in Ross (1976) is referred to as strict factor model because it assumes the common factors capture all the correlations between all variables, which means $\Sigma_u \equiv \text{cov}(u_t)$ is a diagonal matrix. However, this assumption may be too stringent in practice and we normally need to allow for some level of cross-section correlations between the idiosyncratic errors. Therefore the approximate factor model of Chamberlain and Rothschild (1983) seems more appropriate. In this model, the key assumption is that the idiosyncratic covariance matrix is not diagonal, but its eigenvalues must be bounded as $N \to \infty$. I will come back to this in more details in some later paragraphs.

Clearly, the main objective in factor analysis is to identify the set of common factors, assuming that they exist. Generally, estimating the factors can be done in many ways. We can even specify some observed variables as common factors, based on a theoretical framework or from many experiments. Some examples of this approach are the CAPM or the 3-factor model of Fama and French (1993). On the other hand, factors can be considered as latent variables and require statistical techniques to estimate. This is a more popular direction in current literature, as we usually have no prior knowledge about the common factors. Based on this approach, a large literature now in factor models are contributed by extending the factor structure (e.g. dynamic factor model, multi-level factors, etc,) and identification techniques

(e.g. principle component (PC) analysis, maximum likelihood estimator, etc.).

Another aspect that plays an important role in theoretical and empirical work is to determine the number of factors in the model. A few methods have already been proposed and used in applications. The simplest method is to select the number of factors from the scree plot of the descending sample eigenvalues of $\Sigma \equiv \text{cov}(Y_t)$ (i.e. eigenvalues of the sample covariance matrix of Y_t) as in Cattell (1966). Related procedures are suggested by Onatski (2009, 2010) using the slope of the scree plot and the difference of ordered sample eigenvalues, respectively. In addition to these, Ahn and Horenstein (2013) consider maximising the ratio of successive eigenvalues or their growth ratio.

Information criteria have also been used to select the number of factors. Choi and Jeong (2013) study the consistency of using AIC, BIC or HQIC in choosing the true factor model. In addition, Bai and Ng (2002) propose several criteria for the number of factors in approximate factor models and show them to be consistent. The relationship between the information criteria and those based on eigenvalues is discussed in Onatski (2010) and Ahn and Horenstein (2013). Once the number of static factors is determined, the number of dynamic (or primitive) factors can be determined using methods proposed by Amengual and Watson (2007), Bai and Ng (2007), and Breitung and Pigorsch (2012).

This is also worth mentioning at this stage that there can be two different ways when specifying the model in (1.2). In the first one, the common factors are assumed to affect most components in the cross-section, which is called pervasiveness. This formally means the number of non-zero loadings for each factor needs to grow with N. Literature for this model can be founded in Bai (2003), Stock and Watson (2002), Bai and Ng (2002) and some references within. A second type of factor model is less common, but starting to attract some attentions recently. In contrast, the factors in the second type are not defined to capture the cross-section correlation but rather drive the serial dependence of the original time series. Following this direction, we

assume there is a set of common factors that account for all the serial correlations and hence the idiosyncratic components are just white noise. Some attempts in this direction include Anderson (1963), Priestley et al.(1974), Brillinger (1981), Peña and Box (1987), and Pan and Yao (2008). More recent efforts focus on the inference when the dimension of time series is as large as or even greater than the sample size; see, for example, Lam, Yao and Bathia (2011), Lam and Yao (2013) and the references within. In summary, the first class of model assumes the common factors leave very little cross-section correlation in the idiosyncratic components but allow for serial correlation, whereas the second class assumes u_t is serially uncorrelated but the factors can be less pervasive.

In this thesis, I mainly focus on the static factor model as shown in (1.3) and I adopt the model setup similar to the one discussed in Bai and Ng (2002), Bai (2003) or Stock and Watson (2002), in which u_t is allowed for serial correlation. However, as we shall see, I relax the pervasiveness condition that usually comes with the model.

1.1.1.2 Factors identification

It is very important to notice that the latent factors can not be uniquely identified without further restriction. For example, we can always linearly transform f_t and Λ by an $r \times r$ invertible matrix and its inverse and they still generate exactly Y_t. Therefore, we can only estimate the loadings and the factors up to their spanning spaces without any restrictions.

If f_t is a stationary process, a well-known restriction is that $\Sigma_f \equiv \text{cov}(f_t) = I_r$, where I_r is the $r \times r$ identity matrix. This is simply done by replacing f_t by $\Sigma_f^{-1/2} f_t$. However, this is still not sufficient for unique identification, because for now we can still rotate the factors by an orthonormal matrix and still having $\text{cov}(f_t) = I_r$. Therefore, together with this we usually impose an extra restriction that $\Lambda'\Lambda$ is a diagonal matrix (with distinct diagonal elements in descending order). This extra assumption helps us exactly identify f_t and Λ (up to the signs) instead of the rotations

of them, without loss of generality because we know that any other restrictions can be retrieved by a linear transformation. These restrictions are often found in the maximum likelihood estimation, e.g. see Lawley and Maxwell (1971). Furthermore, as discussed in Bai and Ng (2013), it is not as stringent as it seems, and can be useful for economics applications. An example shown in Bai and Ng (2013) is the case where $r = 3$ and

$$\Lambda = \begin{bmatrix} \Lambda_1 & 0 & 0 \\ 0 & \Lambda_2 & 0 \\ 0 & 0 & \Lambda_3 \end{bmatrix} \tag{1.4}$$

where Λ_i is the $N_i \times 1$ vectors of loadings, and $N_1 + N_2 + N_3 = N$. This model implies that the first factor generates the first N_1 group of cross-section components, and so on. This can be applied in models for regional panel data. Even when the order of the cross-section components is shuffled, the loadings matrix restriction still holds, which makes it useful because we do not require the knowledge of the grouped structure. In this thesis, these restrictions regarding the factors and loadings are not needed for the main results, although I shall often refer to this restriction in some discussions for convenience.

Having discussed about estimators for the factors and the loadings matrices, it is also important to point out that in large dimensional setting (N is as large as T), principle components (PCs) analysis is considered as the most efficient methods to achieve this task. The first r (population) PCs of Y_t, denoted as g_t, are defined as follows:

$$g_t = B'Y_t$$

where B is the $N \times r$ matrix whose columns consisting of r eigenvectors corresponding to the r largest eigenvalues of Σ, normalised so that $B'B = I_r$. We can also write:

$$Y_t = Bg_t + w_t \tag{1.5}$$

with $E(g_t w_t') = 0$. Intuitively, if the first r principle components already capture the large proportion of variation in Y_t, the term w_t can be interpreted as the disturbance. Therefore, as in (1.3) and (1.5) there is a similarity between the PCs and factors. In fact, Schneeweiss (1997) develops a result which shows the convergence of PCs to the factors. The key requirement for this convergence is:

$$\frac{\mu_r(\Lambda'\Lambda)}{\mu_1(\Sigma_u)} \to \infty. \tag{1.6}$$

where $\mu_k(A)$ is the kth-largest eigenvalues of a square matrix A. The ratio in (1.6) can be interpreted as the signal-to-noise, and is a key parameter that determines how well one can identify the factors.

The results in Schneeweiss (1997) are developed for population PCs, where Σ is assumed to be known. However, replacing Σ by the sample covariance matrix introduces further sampling errors, especially when N is large. The convergence of sample PCs to factors space is one of the crucial developments recently, and can be found in Bai (2003), Bai and Ng (2002) or Stock and Watson (2002). The authors show that when $(N, T) \to \infty$, the sample PCs converge to the factors space under some conditions, in which some among them imply (1.6).

In order to get to our main contribution, it is worth explaining the intuitive interpretation behind the seemingly technical condition (1.6). What $\mu_1(\Sigma_u)$ and $\mu_r(\Lambda'\Lambda)$ represent are really the amount of cross-section correlations in the idiosyncratic components and the pervasiveness of the factors. First we discuss about Σ_u, which was originally assumed to be diagonal in Ross (1976). However, since the introduction in Chamberlain and Rothschild (1983), the idiosyncratic errors u_{it} are allowed to be cross-sectionally correlated, i.e. we can have a pair (i, j) such that $\text{cov}(u_{it}, u_{jt}) \neq 0$. This is called "approximate factor model", as opposed to the "strict factor model" where $\text{cov}(u_{it}, u_{jt}) = 0$ for all $i \neq j$. Although allowing Σ_u to be different than a diagonal matrix, Chamberlain and Rothschild (1983) require $\mu_1(\Sigma_u)$ to be bounded

as $N \to \infty$.

In this case, if the factors are pervasive enough, then (1.6) is satisfied. To see why, the pervasive condition is usually stated as: $\sum_{i=1}^{N} \left(\lambda_i^{(k)}\right)^2$ grows linearly with N for any $k \in (1, ..., r)$. Equivalently, for any factors, the number of non-zero loadings must grow strictly with order N. If $\Lambda'\Lambda$ is a diagonal matrix as usually assumed for unique identification, then the eigenvalues lie on the diagonal, and the kth eigenvalue is $\sum_{i=1}^{N} \left(\lambda_i^{(k)}\right)^2$. So condition (1.6) is satisfied if the factors are strongly pervasive and $\mu_1(\Sigma_u)$ is bounded.

These two conditions regarding low cross-section correlations of idiosyncratic errors and pervasiveness of factors can be founded in most recent works of factor models such as in Bai and Ng (2002), Bai (2003), Stock and Watson (2002) and the references therein. For example, I recall the two assumption B and E2 in Bai (2003) and denote them as Assumption 0 in this paper:

Assumption 0. *(i) $\Lambda'\Lambda/N$ converges to a positive definite matrix D whose eigenvalues are bounded away from both 0 and infinity*

(ii) Σ_u has bounded row sum of absolute entries, i.e. $\max_i \sum_j |\sigma_{ij}| = O(1)$ where $\sigma_{ij} = cov(u_{it}, u_{jt})$.

Assumption 0 (i) makes sure that each factor has impacts on the majority of the components in the cross-section. Assumption 0 (ii) describes the level of cross-section correlations between the idiosyncratic errors. Slightly weaker one is used in Bai and Ng (2002): $\frac{1}{N}\sum_i \sum_j |\sigma_{ij}| = O(1)$. The main idea behind these restriction on Σ_u is that although the model allows for approximate factor, the level of correlations across the idiosyncratic errors can not exceed a certain level. Notice that if $|\sigma_{ij}|$ is bounded above and below for all i, j, Assumption 0 (ii) leads to $\max_i \sum_j \mathbf{I}\{|\sigma_{ij}| > 0\} = O(1)$.

To see why, notice that:

$$\begin{aligned}\max_i \sum_j \mathbf{I}\{|\sigma_{ij}| > 0\} &= \max_i \sum_j |\sigma_{ij}|^0 \mathbf{I}\{|\sigma_{ij}| > 0\} \\ &= \max_i \sum_j |\sigma_{ij}|(|\sigma_{ij}|)^{-1}\mathbf{I}\{|\sigma_{ij}| > 0\}. \\ &\leq \max_{i,j} \left[(|\sigma_{ij}|)^{-1}\mathbf{I}\{|\sigma_{ij}| > 0\}\right] \max_i \sum_j |\sigma_{ij}| = O(1).\end{aligned}$$

Intuitively, $\max_i \sum_j \mathbf{I}\{|\sigma_{ij}| > 0\} = O(1)$ means that the number of non-zero entries in each row of Σ_u must be bounded while its dimension N grows to infinity. Therefore, later on we will use the fact that $\max_i \sum_j \mathbf{I}\{|\sigma_{ij}| > 0\} = O(1)$ can be derived from of Assumption 0 (ii)[1]. The reason for looking at $\max_i \sum_j \mathbf{I}\{|\sigma_{ij}| > 0\}$ is that we want to use some important results in the sparse matrix[2] literature. This is useful for us later to construct a method to estimate the number of factors (see Chapter 3).

In addition, Assumption 0 (ii) implies that $\mu_1(\Sigma_u)$ is bounded[3]. Therefore, together conditions (i) and (ii) of Assumption 0 imply (1.6), which contributes to the sufficient conditions required for the population PCs to converge to the factors. However, it may be stronger than necessary because we only need (1.6) to hold. It is interesting to consider the cases where Assumption 0 does not hold and examine whether the population and sample PCs still converge to the factors space. Clearly, the sample PCs case will be the ultimate goal, so most of the current studies directly establish the consistency result for this.

One such interesting case is discussed in the PhD thesis of Heaton (2008). This is the case where $\sum_j |\sigma_{ij}|$ grows with rate $N^{1-\alpha}$ for some i and $0 < \alpha \leq 1$. In this case, Heaton (2008) shows that the sample PCs still converge to the factors, but with a

[1] The condition $\max_{i,j} \left[(|\sigma_{ij}|)^{-1}\mathbf{I}\{|\sigma_{ij}| > 0\}\right] = O(1)$ is reasonable, as it just simply states all the non-zero entries in Σ_u must be bounded away from 0, which is true when Σ_u is non-stochastic.

[2] A large matrix with many zero entries is called sparse matrix, and this has attracted a large number of studies recently

[3] As from standard Linear Algebra result, we have $\mu_1(\Sigma_u) < \max_i \sum_j \sigma_{ij} \leq \max_i \sum_j |\sigma_{ij}|$.

slower rate. Particularly, he proves that (using our notations):

$$\frac{1}{T}\left\|\tilde{F}^r - FA\right\|^2 = O_p(\frac{1}{N^\alpha} + \frac{1}{T})$$

where $\tilde{F}^r$ is the sample PCs and A is a rotational matrix. In Heaton's thesis, where he assumes the column of F is orthonormal, A becomes only a signs matrix. Under Assumption 0, the rate of convergence for similar quantity in the left hand side above is established in Bai and Ng (2002) and is $O_p(N^{-1} + T^{-1})$. Therefore we can see how weakening Assumption 0 (ii) affects the rate of convergence.

In this thesis, I mainly focus on the case where Assumption 0 (ii) is not violated, but we relax Assumption 0 (i). There are some reasons for this to be too stringent, because under this condition a factor that only affects a relatively small number of cross-section components (say $N^{2/3}$ out of N) will not be assumed to exist. However, recently some empirical works suggest the potential loss of information when ruling out these not-so-strong factors. For example, Boivin and Ng (2001) provide an empirical study illustrating that a smaller but carefully chosen set of cross-section variables yields better factors than the whole original set. One potential reason for this is that the amount of correlations from the large number of idiosyncratic errors will reduce the sharpness of the estimators. In addition, some factors that are extracted from the subset can be identified as idiosyncratic errors when applying factor analysis to the full set, due to their small explanatory powers for the majority of cross-section variables.

At what level can we relax Assumption 0 (i) is also an interesting issue. If we relax it so that $\Lambda'\Lambda$ converges to a positive definite matrix D, then clearly (1.6) is not satisfied, and so even population PCs are not consistent for the factors. To strengthen our argument, Onatski (2012) shows that under this case, the factors are so weak that sample PCs will be poor estimators for the factors, and can even be orthogonal to the factors space. A studying case that we present in this paper lies

in the middle, that is we assume the number of non-zero loadings for each factor is at order $d(N)$, which can be dominated by N but has to grow to infinity with N. We even consider the general case where $d(N)$ varies from factors to factors. This is found useful in the case where some factors are global while some of them are regional, but both the number of regions and the sizes of them can be large.

Another related area of research with weak factors is the multi-level factor model, which usually includes global and regional levels. In such model, the factors are separated into levels, where the top level factors (global factors) are pervasive and affect almost every cross-section component. The second level factors are not as pervasive because they only affect components in each region. In a special case where the number of components in each sector grow at a slower rate then N, this 2-level factor is a special case of a mixture model of strong and weak factor. Restriction for such model and effective identification method can be found in Wang (2008). The most important similarity between the multi-level factor model and the one presented in this thesis is that the loadings matrix can be allowed to have many zeros. However, there is a key difference, which is that in my proposed model we neither know how to separate the original cross-section into sectors nor if such separation is possible.

1.1.2 Applications of factor model

There are a wide range of applications of factor model that can be found in the Economics and Finance literature. They can be separated into two classes: the first one links the factors with some interpretations for meaningful insights about the observed variables (e.g. APT, CAPM, Fama-French 3 factor model, business cycle[4], yield curve modelling[5], etc.), whereas the second class uses the factor analysis as a tool for dimension reduction. In here, I focus more on the second one.

High-dimensional settings can be found in many applications recently, due to the growth in available data and the advance in computational techniques. Typically

[4]Gregory et al. (1997)
[5]Diebold et al. (2006)

the vast dimension can be hundreds or thousands, e.g. number of firms in the stock markets or macro variables in the global economy. An unarguable advantage with the growth in the size of data is to capture more information which can not be revealed from any smaller sets.

However, this clear advantage is not taken for granted, because the suitability of analysis tools used for these large dimensional data has to be examined before being applied. For example, in the case where the number of interested variables expands faster than their sample sizes, many traditional theoretical estimators in data analysis break down due to undesirable bounds required for convergence, such as the sample covariance matrix of these data. Therefore, a large class of innovative methods has arisen which either seek for dimension reducing techniques or extend the theoretical results for large dimension, including factor analysis.

1.1.2.1 Large covariance matrix estimation:

To begin we give one such technically challenging example that arises in high-dimensional setting: i.e. estimating the covariance matrix when the cross-section dimension (N) of the data is as large as the sample size (T). It is well known that in this situation, the sample covariance matrix is very ill-behaved, and it is not even invertible when $N > T$. Since the development of Markowitz portfolio theory, covariance matrix of returns has been an important concept in Finance and Econometrics, and we would definitely want to have a "good"[6] estimator for this, no matter how large N is.

For some backgrounds in this area: suppose we have a portfolio of N assets. Based on Markowitz portfolio theory, finding the optimal portfolio allocation requires us to estimate the $N \times N$ covariance matrix across the assets returns (assume constant in this period), denoted Σ. The diagonal of this matrix is the variance of each asset return in the portfolio, where the (i, j) off-diagonal entry is the covariance of returns

[6] By "good" I mean relatively low bias and variance.

between asset i and asset j. In order to allocate the weighs of investment for these assets within a portfolio, we may choose the one that reaches our required expected return with minimal variance. If we denote $Y_t = \{y_{it}\}_{i=1}^N$ the vector of N assets returns at time t, the variance of the portfolio with weighs $w = \{w_i\}_{i=1}^N$ is:

$$\text{var}(w'Y_t) = w'\Sigma w.$$

Therefore, it can be seen that the covariance matrix has a closed link with risk management in practice. Solutions for estimating high-dimensional Σ are normally obtained by proposing a structure for the covariance matrix (or in other words, for the data generation process of Y_t). This is usually called regularisation. Some popular regularised restrictions are banded and sparse, which restrict the number of non-zero entries in Σ. However, applying a sparse (or banded) structure directly to Σ is not realistic, for example it is possible that all assets returns are correlated with each other. Therefore a more rational restriction is that Σ can be decomposed into a sum of a low rank matrix and a sparse matrix, a property that can be resulted if Y_t has a factor structure representation. If (1.2) holds true and f_t and u_t are independent then:

$$\Sigma = \Lambda\Sigma_f\Lambda' + \Sigma_u. \tag{1.7}$$

where Σ_f and Σ_u are the covariance matrices of f_t and u_t. The decomposition in (1.7) provides an efficient estimator for Σ if r is small and Σ_u has many zero entries. In this case, the product matrix $\Lambda\Sigma_f\Lambda'$ has rank r and Σ_u is sparse, so we significantly reduces the number of parameters required to estimate.

Apart from this, there are many other applications involving extracting factors from the original data set with high cross-section dimension, because certainly it is desirable to find a smaller set of variables that can contain a large proportion of information from the vast original multivariate data set. In section 1.1.2, there will be a survey about popular applications of factor model in Economics and Finance.

1.1.2.2 Forecasting with diffusion indexes[7]:

Assuming we want to forecast a h-step ahead for a variable x_t and know that x_{t+h} can be predicted by the following forecasting model:

$$x_{t+h} = \alpha + F_t'\beta + \epsilon_{t+h}$$

In this case, although the factors F_t is not observed, it can be extracted from other observable variables in the market, i.e. Y_t if in fact we have a factor structure as in (1.3). Given that the dimension of Y_t (N) is very large, it is desirable to use the factors as the explanatory variables in the forecasting equation.

1.1.2.3 Large-dimensional vector autoregressive:

Consider a task where one may want to forecast the future values for $\{y_{it}\}$ for some $i \in (1, .., N)$. We can imply they follow a VAR structure with some added exogenous variables. i.e.

$$Y_t = \Theta(L)Y_t + \Gamma Z_t + \epsilon_t$$

where $Y_t = (y_{1t}, ..., y_{NT})$, Z_t represents the exogenous variables and ϵ_t consists of some noises that are spatially uncorrelated. When N is small we can estimate all the unknown parameters by maximum likelihood, as usual. Where N is large, or even extremely large this can yield further problem due to infeasibility to cope with large number of unknown parameters. One way to solve this problem is that we can assume Y_t has the factors structure as in (1.3) and replace the Y_t on the right hand-side with Λf_t. After obtaining f_t from Y_t then $\Theta(L)\Lambda$ can be estimated together as a single lag matrix and has much less parameters then $\Theta(L)$. For example, if the lag of our VAR model is 1 then $\Theta(L)\Lambda$ is a $N \times r$ matrix whereas $\Theta(L)$ is the $N \times N$ matrix.

[7] Stock and Watson (1998)

1.2 Contributions of this thesis

In this thesis, I attempt to replace the strong pervasiveness condition for the factors with a less stringent one, while assuming that the idiosyncratic covariance matrix is sparse. The sparsity of Σ_u is stated through Assumption 2 (iii), which is same as Assumption 0 (ii). Particularly, when Σ_u has bounded row sum uniformly and finite entries, the number of non-zero entries of Σ_u must be bounded. This has better interpretation in some applications, e.g. conditional on the common factors, most asset returns in the market are uncorrelated. I later on define the sparsity level of a matrix as its maximum number of non-zero entries in a row, and by this definition Σ_u must have bounded sparsity level. This fact is also found useful later when a novel criterion to choose the number of factors is proposed.

In summary, together with the sparsity assumption for Σ_u, here are some important questions that are studied in this thesis:

1. If we loosen the restriction that $\Lambda'\Lambda/N \to D$ to $\Lambda'\Lambda/d(N) \to D$ for a function $d(N)$ that grows to infinity at a slower rate than N can we still identify the factors, given that the number of factors is known. Recall from above that D is a positive definite matrix whose eigenvalues are bounded away from *both* 0 and infinity. Furthermore, replacing N with a single term $d(N)$ means that all the factors still have same strength order. I will also generalise to discuss the case where each factor can have different strength, which also generalises the multi-level factors model in Wang (2008). This is described later in Assumption 1.

2. If it turns out that the factors can be identified in the weak factor model but we do not know the number of factors, how do we consistently estimate it?

All of these questions are not trivial considering the current literature in high dimensional factor analysis. For example, in Bai and Ng (2002) or Bai (2003), it is known that with Assumption 0 among others, the factors are identified up to a ro-

tation matrix which equals $(\Lambda'\Lambda/N)(F'\tilde{F}^k/T)(V^k)^{-1}$ where $\tilde{F}^k$ is the matrix of k estimated factor by PCs and V^k is the $k \times k$ diagonal matrix of the first k largest eigenvalues of $YY'/(NT)$ in decreasing order. Therefore an important condition is that this rotation matrix has eigenvalues bounded away from both 0 and ∞. For the case when only $\Lambda'\Lambda/d(N) \to D$, this is not straightforward even when k is the true number of factors, so it requires further investigation and modification to the work of Bai and Ng.

In addition, most current methods in determining the number of factors such as in Bai and Ng (2002), Onatski (2010), Ahn and Horenstein (2013) exploit the sharp edge in the set of sample eigenvalues of Σ, which separates the factors and the errors. These are all based on Assumption 0 that intuitively states that Σ has exactly r spiked eigenvalues growing at rate N. In our setting, because of replacing this crucial condition, the number of factors is now much harder to determined. For example, using the ratio of eigenvectors method of Ahn and Horenstein (2013), there might easily be a case where the ratio of orders N^α and N eigenvalues for $\alpha < 1$ is less than the ratio of N^0 and N^α eigenvalues. Clearly, we want the number of factors to be at the point where the eigenvalues of Σ begin to be bounded but it will not always be possible to identify in this case, using the sample eigenvalues ratio based method. In support of our argument, Yu and Samworth (2013) provide some Monte Carlo results that show how the Bai and Ng (2002) criteria can underestimate the number of factors when the first r largest eigenvalues of the data are not as spiked as rate linear with N.

Apart from showing the consistency of the sample PCs for estimating the factors space (Chapter 2) and propose a novel criterion for choosing the number of factors (Chapter 3), I will look at two applications of the weak factor model in practice in Chapter 4 and 5. One of them regards estimating large covariance matrix, particularly the POET estimator proposed in Fan et al. (2013). I discuss the impact of weak factor model to the POET estimator and show that the number of factor is not

crucial for a consistent estimator. In addition, another application of the proposed criterion for the number of factors can be used for factor model selection, where the candidates include both observed and unobserved factors.

2 Factor identification under the weaker assumption

In this chapter, I discuss the identification of factors under the weaker assumption. It will later be seen that a factor and its loading space are still consistently estimated if the factor has the corresponding number of non-zero loadings more than a certain level. This can be useful for some empirical studies as now we can apply factor analysis to some applications where not all the factors are pervasive. Examples of these situations are the cases where we can have both regional and global factors. If the regional factors are not too weak, we can still extract them from the whole original data set.

First of all, we emphasise that r is fixed while N will grow to infinity. The number of factors r includes both the strong and weak factors. We should explicitly clarify here that the nature of our model allows factors with different strengths co-exist, hence there must be a clear edge between the weakest factor and the idiosyncratic errors. The fact that r is fixed when $N \to \infty$ is crucial as it implies there must be a lower bound for the strength of the factors, which leads to the clear edge mentioned previously.

Some results in this paper require $N = o(T^2)$, so generally the framework in this paper is regarding to the case that N and T are comparably large.

2.1 Factors identification techniques

2.1.1 Principle Components

The most usual way to estimate the factors and the loadings are via PCA: Given a value for k as a predetermined number of factors, we estimate (Λ, F) by $(\tilde{\Lambda}^k, \tilde{F}^k)$ such that

$$(\tilde{\Lambda}^k, \tilde{F}^k) = \arg\min_{\Lambda^k, F^k} \frac{1}{NT} \sum_{t=1}^{T} (Y_t - \Lambda^k f_t^k)'(Y_t - \Lambda^k f_t^k) \tag{2.1}$$

where $F^k = [f_1^{k'}, ..., f_T^{k'}]$, f_t^k is a $k \times 1$ vector representing a factor value at time t and Λ^k is a $N \times k$ loadings matrix. Recently, Choi (2012) and Bai and Liao (2013) generalise the standard PC method to generalised PCA (GPCA) method that gives benefit of a lower variance in the estimators, i.e. the objective function becomes:

$$\arg\min_{\Lambda^k, F^k} \frac{1}{NT} \sum_{t=1}^{T} (Y_t - \Lambda^k f_t^k)' W (Y_t - \Lambda^k f_t^k). \tag{2.2}$$

In here to keep thing simple all the proofs refer to the traditional PC method, but a generalisation is also possible.

As usual the solution for (2.1) is not unique up a rotation, because clearly if $(\tilde{\Lambda}^k, \tilde{F}^k)$ is a solution of (2.1) then $(\tilde{\Lambda}^k A, \tilde{F}^k A'^{-1})$ is another solution for any invertible $k \times k$ matrix A. However, if we uniquely restrict $(\tilde{\Lambda}^k, \tilde{F}^k)$ so that $\tilde{\Lambda}^{k'} \tilde{\Lambda}^k$ is diagonal and $\tilde{F}^{k'} \tilde{F}^k / T = I_k$, the following pair of solutions of (2.1) can be used: the columns of $\tilde{F}^k$ will contain $\sqrt{T}$ times the eigenvectors corresponding to the largest k eigenvalues of YY', normalized so that $\tilde{F}^{k'} \tilde{F}^k / T = I_k$, then $\tilde{\Lambda}^k = Y' \tilde{F}^k / T$. In this case, it is easy to check that:

$$\tilde{F}^{k'} \tilde{F}^k / T = I_k; \quad \tilde{\Lambda}^{k'} \tilde{\Lambda}^k \text{ is diagonal.} \tag{2.3}$$

Without similar restriction for the true factors and loadings, these estimators can still converge to the true spanning spaces of Λ and F. Unlike the pervasive factor

scenario, in here we must be careful with which version of estimators to choose. For example consider a pair of solution $(\bar{\Lambda}^k, \bar{F}^k)$ where:

$$\bar{\Lambda}^{k\prime}\bar{\Lambda}^k/N = I_k; \quad \bar{F}^{k\prime}\bar{F}^k/T \text{ is diagonal.} \tag{2.4}$$

This is a standard estimator also shown in Bai and Ng (2002) or Bai (2003). The method for finding $(\bar{\Lambda}^k, \bar{F}^k)$ is as follow: if we concentrate out F^k then the columns of $\bar{\Lambda}^k$ will contain the $\sqrt{N}$ times the eigenvectors corresponding to the largest k eigenvalues of $Y'Y$, normalized so that $\bar{\Lambda}^{k\prime}\bar{\Lambda}^k/N = I_k$. Then by standard least square result, $\bar{f}_t^k = (\bar{\Lambda}^{k\prime}\bar{\Lambda}^k)^{-1}\bar{\Lambda}^{k\prime}Y_t = \bar{\Lambda}^{k\prime}Y_t/N$. However, when the factors are weak, it is possible to have $\Lambda'\Lambda/N$ singular, and therefore $\bar{\Lambda}^k$ can not be a consistent estimator for any rotations of Λ.

For that reason, when the factors are not pervasive, it is always better to use $(\tilde{\Lambda}^k, \tilde{F}^k)$ which satisfy (2.3).

2.1.2 Maximum Likelihood Estimator

Assuming the idiosyncratic errors u_t are i.i.d and follow $Gaussian(0, \Sigma_u)$, the objective function for estimating (F, Λ) using conditional quasi log-likelihood (removing all constant terms, and multiplying by -2) is:

$$\frac{1}{N}\log|\det(\Sigma_u)| + \frac{1}{NT}\text{tr}\left((Y - F\Lambda')'(Y - F\Lambda')\Sigma_u^{-1}\right) \tag{2.5}$$

This can be shown to be more efficient than the principle components method, for discussion see Bai and Liao (2013) or Choi (2012). For example, Choi (2012) shows that the asymptotic variances of the estimators for (F, Λ) are smaller when using the objective function in (2.5) than those obtained when using the original principle component objective function.

It is easy to verify that this is exactly same as the GPCA, where the objective function is:

$$\min_{\Lambda^k, f_t^k} \frac{1}{NT} \sum_{t=1}^{T} (Y_t - \Lambda^k f_t^k)' \Sigma_u^{-1} (Y_t - \Lambda^k f_t^k) \tag{2.6}$$

In here, the weighted matrix is Σ_u^{-1}. Notice that in this case GPCA to PCA is an analogy with generalised least square to ordinary least square.

However, this also requires an estimator for Σ_u. The usual way for obtaining estimator for Σ_u is to assume that it is diagonal and its diagonal entries are just the sample variance of a prior fitted factor model. Recently, Bai and Liao (2013) replace the diagonal condition for $\Sigma_u$with sparsity and use the POET estimator. The general idea of their approach is to find the factors and loadings by PCA, then estimate Σ_u from the residuals[1] and plug this estimator of Σ_u into (2.6) to obtain a better version of (F, Λ) estimators.

As already mentioned, GPCA brings some benefits of variance reduction of the estimators. Most of the results in this thesis can be extended to this, but I only stay in the PCA framework to keep the process and notation clearer.

2.2 Notations

In here I introduce some notations that are used in from here onward. As mentioned before, let Σ and Σ_u be the population covariance matrices of Y_t and u_t. Also, let σ_{ij} be the entries of Σ_u. Furthermore, let $\tilde{\Sigma}$ and $\tilde{\Sigma}_u$ be the corresponding sample covariance matrices. Clearly, only $\tilde{\Sigma}$ can be computed from observed data, $\tilde{\Sigma}_u$ is estimated with estimated version of u_t, which are the residuals after fitting in the factors.

Furthermore, for a given value of $k \leq r$, define the following partitions:

$$f_t^{(l:k)} = (f_t^{(l)}, ..., f_t^{(k)})', \; F^{(l:k)} = (f_1^{(l:k)'}, ..., f_T^{(l:k)'})$$

[1] They also apply a thresholding step after computing the residuals sample covariance matrix, as referred to as POET estimator, see Chapter 4 in this thesis.

$$\lambda_i^{(l:k)} = [\lambda_i^{(l)}, ..., \lambda_i^{(k)}]', \ \Lambda^{(l:k)} = [\lambda_1^{(l:k)}, ..., \lambda_N^{(l:k)}]$$

For more convenient, I also use these notations:

$$f_t^k = f_t^{(1:k)}, \ F^k = F^{(1:k)}$$

$$\lambda_i^k = \lambda_i^{(1:k)}, \ \Lambda^k = \Lambda^{(1:k)}$$

$$u_t^k = Y_t - \Lambda^k f_t^k, \ \Sigma_u^k \equiv \left(\sigma_{ij}^k\right) = \text{cov}(u_t^k)$$

Those notations above are clearly not defined for $k > r$, however their estimated version can take any values for k.

We consider estimating the factors by PC analysis, $\tilde{\Lambda}^k, \tilde{F}^k$ are already defined in (2.1) and (2.3), and we will let all the partitions for $\tilde{\Lambda}^k, \tilde{F}^k$similar to the notations for the true ones above. Also, we define:

$$\tilde{u}_t^k = Y_t - \tilde{\Lambda}^k \tilde{f}_t^k, \ \tilde{\Sigma}_u^k = \frac{1}{T}\sum_{t=1}^{T} \tilde{u}_t^k \tilde{u}_t^{k'}$$

as u_t has mean zero.

For a square matrix we will let μ_i be its ith largest eigenvalues. The matrix norms we use in this paper are the operator norm and the Frobenius norm, i.e. $\|A\| = \mu_1(A'A)$, $\|A\|_F = \text{trace}(A'A)$ respectively. As a special case, we also denote $v_i = \mu_i(\Sigma)$ and $\tilde{v}_i = \mu_i(\tilde{\Sigma})$. In addition if $a_n = O(b_n)$ and $b_n = O(a_n)$ when $n \to \infty$, we will write $a \asymp b$. Similarly, replacing O by O_p, we can define $\asymp_p$ in a same manner. Finally, w.l.o.g we assume all the time series Y_t, u_t and f_t have zero means.

2.3 Asymptotic results

Apart from the new assumption about the strength of factors, the rest of our assumptions are very similar to the standard literature. The main theorem in Chapter 2 can be done by using the same assumptions in Bai (2003) and only replace the

pervasive factors conditions. However, we also wish to develop a new method for determining the number of factors in Chapter 3, which makes use of some results in Fan et al. (2011, 2013). For example, the stationary condition in Fan et al. (2013) for all stochastic process is stronger than in Bai (2003), but can be still reasonable in practice. Therefore, in this thesis I adopt the similar set of assumptions as in Fan et al. (2013).

Assumption 1. *There exists a matrix* $D_N = diag(d_1(N), ..., d_r(N))$*, where for all* $1 \leq i \leq r$*,* $d_i(N) \to \infty$ *and* $d_i(N)/N$ *is bounded away from* ∞ *as* $N \to \infty$*, such that* $D_N^{-1}\Lambda'\Lambda \to D$ *and all the eigenvalues of* D *are bounded away from both 0 and infinity.*

This assumption above allows for a more general case where all the factors can have different strengths, which are indicated by the order of $d_i(N)$. If we assume $\Lambda'\Lambda$ is diagonal, another way to state this assumption is $\left\|\Lambda^{(k)}\right\|^2 = \sum_{i=1}^{N}(\lambda_i^{(k)})^2 \asymp d_k(N)$. In addition, to label the factor according to their strengths, we further let $d_1(N) \geq ... \geq d_r(N)$ as $N \to \infty$. Recently, Lam et al. (2011) and Lam and Yao (2012) consider weak factors similar to the ones discussed in this paper. However, there is a key difference with their factor model, which is assumed to capture all the serial correlation in the original time series. This paper considers a factor model that is similar to the one discussed in Bai and Ng (2002), Bai (2003) or Stock and Watson (2002).

Assumption 2. *(i)* $\{u_t, f_t\}$ *is strictly stationary with zero mean and* $\{f_t\}$ *is independent of* $\{u_t\}$*.*

(ii) $\frac{1}{T}\sum_{t=1}^{T} f_t f_t' \to \Sigma_f$ *as* $T \to \infty$*, with* $\mu_1(\Sigma_f)$ *and* $\mu_r(\Sigma_f)$ *are bounded away from both 0 and* ∞*.*

(iii) Σ_u *has bounded row sum of absolute entries, i.e.* $\max_i \sum_j |\sigma_{ij}| = O(1)$*.*

(iv) (exponential-type tails) There exists constants $r_1, b_1 > 0$ *such that for any* $s > 0$ *and* $i \leq N$*:*

$$P\left(|u_{it}| > s\right) \leq \exp\left(-\left(s/b_1\right)^{r_1}\right).$$

Also, there exists constants $r_2, b_2 > 0$ *such that for any* $s > 0$ *and* $j \leq r$*:*

$$P\left(|f_{jt}| > s\right) \leq \exp\left(-\left(s/b_2\right)^{r_2}\right).$$

The next assumption is the mixing condition for the factors and the idiosyncratic errors.

Assumption 3. *(strong mixing condition) Define*

$$\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{T}^{\infty}} |P(A)P(B) - P(A \cap B)|$$

where $A \in \mathcal{F}_{-\infty}^{0}$, $B \in \mathcal{F}_{T}^{\infty}$ *are the* σ*-algebras generated by* $\{(u_t, f_t)\}_{t=-\infty}^{0}$ *and* $\{(u_t, f_t)\}_{t=T}^{\infty}$ *respectively, there exists positive constant* r_3 *and* C *such that for all* $t \in \mathbf{Z}^{+}$, $3r_1^{-1} + 1.5r_2^{-1} + r_3^{-1} > 1$ *and* $\alpha(T) \leq \exp\left(-Ct^{r_3}\right)$.

Furthermore, we also require the following regularity conditions:

Assumption 4. *For all* $i \in (1, .., N)$ *and* $s, t \in (1, ..., T)$*,*

(i) $\lambda_i^{(k)}$ *is bounded from infinity for all* $k \in (0, ..., r)$.

(ii) $\frac{1}{N}\left[u_s' u_t - E(u_s' u_t)\right]^2 = O_p(1)$.

(iii) $\left\| D_N^{-1/2} \Lambda' u_t \right\|^2 = O_p(1)$

Remarks

Assumption 1 in this thesis is modified on the basis of Assumption 1 in Fan et al. (2013) as I extend the result to weak factor model. Apart from that, Assumptions 2, 3 and 4 are generally the same as in Fan et al. (2013), with the only difference is that we need to replace N by $d_r(N)$ in some places where possible, where $d_r(N)$ represents the strengths of the weakest factors, i.e. the order of its number of non-zero loadings.

The stationary condition for the factors and the idiosyncratic errors in Assumption 2 (i, ii, iii) is slightly stronger than what is required in standard literature of latent factor models. This allows us to drop the time dependence for Σ_f and Σ_u, which

will make it simple to present the main result. It is not hard to extend to the case of heteroskedaticity for the factors and errors and prove the result in this chapter. In such case, we need to put the upper bound for the moments of $\{f_t\}$ and $\{u_t\}$ across all time. For example, if follow the assumptions used in Stock and Watson (2002) and Bai (2003), we would replace our assumption 2(iii) as $|\sigma_{ij,t}| \leq |\sigma_{ij}|$ and $\max_i \sum_j |\sigma_{ij}| = O(1)$, where $\sigma_{ij,t} = \text{cov}(u_{it}u_{jt})$.

In addition, we require stationary and exponential-type tails (Assumption 2 (iv)) to apply the large deviation theorem, which are needed when dealing with the idiosyncratic errors covariance matrix estimator later. The strong mixing condition in Assumption 3 is also for this purpose. More discussions can be found in Fan et al. (2013).

Assumption 4 is popular in the factor model literature as used in Fan et al. (2013) and Bai and Liao (2013). However, Assumption 4 (iii) is adapted to our weak factor model, i.e. the standard $N^{-1/2}\Lambda' u_t$ is replaced by $D_N^{-1/2}\Lambda' u_t$ inside the norm. This helps to improve the convergence rate in Theorem 2.1 but still reasonable. It is because the i^{th} row of the $r \times N$ matrix Λ' only has order of $d_i(N)$ non-zero entries. On the other hand, assumption 4 (ii) is taken exactly as in previously mentioned papers because the sum of variance and auto-covariance of noises from N cross-section components is $O_p(N)$, that is even to assume that the auto-covariance vanishes after some finite lags. Therefore, the $\frac{1}{N}$ term in Assumption 4 (ii) can not be replaced by some weaker term, which has some impacts on the convergence rate of Theorem 2.1. Consequently, the strengths of common factors must be asymptotically bounded below by some level in order to achieve consistency for the estimators by sample PCs, also see the discussion after Theorem 2.1 for more details.

In summary, the assumptions in the framework of Fan et al. (2013) is slightly stronger than the usual assumption used in standard factor literature (e.g. in Bai (2003)). However, the results in this chapter can still be proved with the same assumptions as in Bai (2003), with only weakening Assumption B in Bai (2003)

which restricts the model to have strong factors only. The reasons I adopt Fan et al. (2013) assumptions are as follows: firstly in Chapter 3 I will use these assumptions to propose a new method to identify weak factors, and secondly in Chapter 4, I show that even when extracting more than r factors the covariance matrix estimated by POET in Fan et al. (2013) is still consistent.

2.3.1 Factor strengths

From Assumption 1, we can see that the strength of factor i, $1 \leq i \leq r$, is measured by $d_i(N)$. In standard literature, $d_i(N) = N$ for all i, which indicates that all the factors affect the majority of cross-section components. In here we allow for different strengths depending on the factor. However, as we proceed is it required that for identification issue we have to put a lower bound for $d_i(N)$ so that the PCs can consistently estimate the factors. Hence the extra Assumption 5 (i) is very important, as it tells us how much we can relax Assumption 0 (i). Since we have to estimate the PCs from the sampled data, it is expected that the value of T has to be in the lower bound of $d_i(N)$ to link how much we can tolerate the weakness of the factors. In addition, we introduce Assumption 5(ii) for the purpose of determining the number of factors later, the idea is that there must be a gap that separate the strengths of common factors and of idiosyncratic components.

Assumption 5. *(i)* $\frac{N\sqrt{\log N}}{d_i(N)} = o(\sqrt{T})$ *for all* $1 \leq i \leq r$

(ii) There exists a function (i) $g(N) \to \infty$ *such as* $g(N)/d_i(N) \to 0$ *as* $N \to \infty$ *for all* $1 \leq i \leq r$

Notice that Assumption 5 hold immediately in the standard literature when letting $d_i(N) = N$ and $\log N = o(T)$, but they also allow for great flexibility. For example, if $d_r(N) = N^{3/4}$ then we require $N^{1/4}\sqrt{\log N} = o_p(\sqrt{T})$, which is not hard to achieve in practice. Nevertheless, this condition is imposed technically based on some bounds used in the proofs. In the future if smaller bound is achieved than we could loosen this restriction. With this assumption, it is therefore not recommended using sample

PCs to estimate the factors when we think the strength of the factors is less than $O(\sqrt{N \log N})$, because it then requires $N = o(T)$, which is not the high-dimensional setting we want. Having said that, when $N = 1000$, $\sqrt{N \log N}) \approx 55$, and we clearly can assume even the weakest factor can affect more than 55 cross-section components. In addition, Theorem 2.1 requires $\sqrt{N} = o(d_r(N))$ for convergence, which is another lower bound for the factor strengths for identification.

2.3.2 Main theorem

Based on all these assumptions, I can now introduce the main theorem in this chapter, which establishes the consistency of sample PCs for the factors.

Theorem 2.1. *Under assumption 1-5, and if r is the true number of factors and $\tilde{F}^r$ is estimated by PC method as shown in (2.1) and restricted by (2.3), there exists a $r \times r$ matrix H^r and $G^r = (H^r)^{-1}$ such that:*

(i) $\frac{1}{T}\sum_{t=1}^{T}\left\|\tilde{f}_t^r - H^r f_t\right\|^2 = O_p\left(\frac{N}{[d_r(N)]^2} + \frac{N^2}{T[d_r(N)]^2}\right)$,

and if $\frac{N}{[d_r(N)]^2} + \frac{N^2}{T[d_r(N)]^2} = o(1)$, *we also have:*

(ii) $\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\lambda}_i^r - G^{r\prime}\lambda_i\right\|^2 = O_p\left(\frac{N}{[d_r(N)]^2} + \frac{N^2}{T[d_r(N)]^2}\right)$

The proof for this theorem is given at the end of this chapter. It turns out that due to the impacts of N cross-sectional noises, the convergence is only achieved if $\sqrt{N} = o(d_r(N))$. That is to say that the weakest common factor needs to have impacts strictly more than $N^{1/2}$ components. For example, if the weakest common factor only affects N^α components for $\frac{1}{2} < \alpha < 1$, the convergence rate is $O_p\left(N^{1-2\alpha} + T^{-1}N^{2-2\alpha}\right)$. When $\alpha = 1$ as in the strong factor case, we go back to the original rate, which is $O_p(N^{-1} + T^{-1})$.

This result also gives supports for the arguments in Boivin and Ng (2006), in which the authors argue that by having too many cross-section components, the level of accumulated noises can affect the convergence rate of the factors, and therefore it is better to reduce the cross-section size. If one considers a weak factor which affects $d_i(N)$ components and drop all the irrelevant ones on the original N components, we

go back to the strong factor model with cross-section size $d_i(N)$ and the convergence rate is $O_p(d_i(N)^{-1} + T^{-1})$, which is better than working with the whole set of N components and achieve our rate. However, it is not always possible to know which are the relevant sets and the factor structure can be much more complex. In that case, one has to apply PCs to the whole data and have the convergence rate depending on the factor strengths.

One can also link this result with the one in Heaton (2008) and see the similarity. When the factors are not pervasive, or when the idiosyncratic errors are too strongly correlated as in Heaton (2008), the sample PCs achieve a lower convergence rate toward the true factors space. However, the key concluding remark here is that they are still consistent if $d_r(N)$ diverges faster than $\max(\sqrt{N}, N\sqrt{\log N/T})$. When T is as large as N, this is approximately $\sqrt{N \log N}$, therefore the pervasiveness of factors can be relaxed substantially.

2.4 Illustrated simulations

In here, I will use Monte Carlo simulations to illustrate the key point made in this section: the rate of convergence of the factors estimated improves with the pervasiveness of the factors. We consider a following data generation process:

- $Y_t = \Lambda f_t + u_t$.
- Λ is a $N \times r$ matrix.
- $f_t = \alpha f_{t-1} + w_t$, where $w_t \sim N(0, I_r\sqrt{1-\alpha^2})$ and $\alpha^2 < 1$. This guarantees f_t stationary with serial correlation and the cross-section covariance matrix of f_t is I_r, which make the PCs converge to the factors (up to a sign change) so is easier for comparing the error rates of estimation.
- $u_t = \beta u_{t-1} + \epsilon_t$, where $\epsilon_t \sim N(0, \Sigma_u\sqrt{1-\beta^2})$ and $\beta^2 < 1$.

- $(\Sigma_u)_{N\times N}$ is generated as a positive definite matrix with some degree of cross section (but is still sparse). We control for the sparsity level of Σ_u by the following procedure: first we generate positive semi-definite matrix which is computed as AA' for some random $N\times\bullet$ matrix A, we then forcing a significant number of off-diagonal entries of AA' to 0 symmetrically and then adding the identity matrix to it to make sure it is positive definite.

In order to make the weak factor model. we follow the simulation in Yu and Samworth (2013) and generate:

- $Y_t = \gamma\Lambda f_t + u_t$

where γ is a number smaller than 1. For example, γ is taken to be $1/3$ or $1/10$ in Yu and Samworth (2013).

In Figure (2.1), I plot the average of $\left\|F - \tilde{F}^r\right\|$ over 1000 simulations vs. the value of $1/\gamma$. The value in the vertical axis (so called Estimated error mean) represents how the estimated factors are different from the true factors. Inside the norm, F is simulated as described above and $\tilde{F}^r$ is estimated by PCA as shown in Section 2.1.

From this figure, it can be seen that the estimators are more precise when the factors are stronger, especially we observe a sudden change in estimation error when the strong factors just start to be weaker. In Figure (2.2) reports the standard deviation of $\left\|F - \tilde{F}^r\right\|$ in 1000 simulations (so called Estimated standard deviation), which also what can be expected. The variance of the estimator should also depend on the pervasiveness of the factors, although I will leave it for future research.

Figure 2.1: Estimated factor errors vs. pervasiveness (1 corresponds to strong factors)

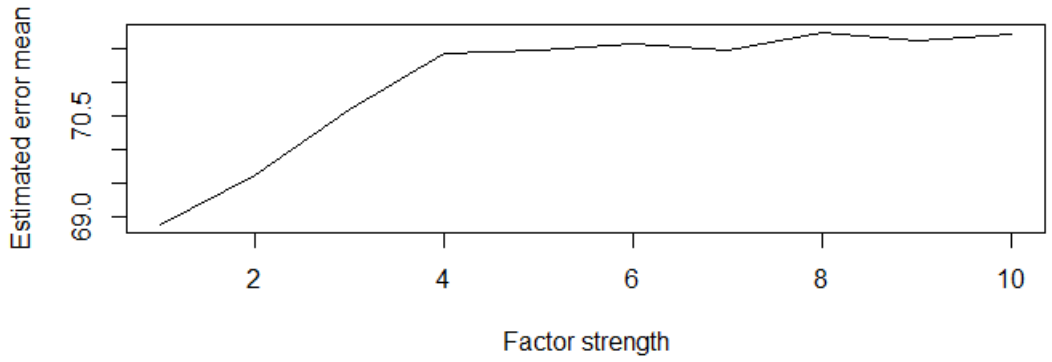
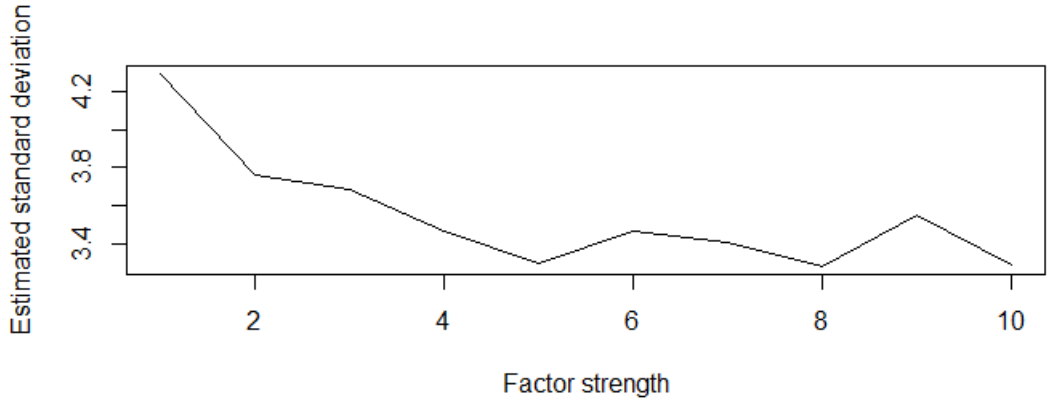


Figure 2.2: Estimated factor errors standard deviation vs. pervasiveness (1 corresponds to strong factors)



Notice that when multiplying the weak factors by γ we decrease the value of all systematic eigenvalues and therefore make the signal-to-noise ratio lower.

2.5 Proofs of results

2.5.1 Proofs of Theorem 2.1

Since $\tilde{F}^r/\sqrt{T}$ is the matrix whose columns are the orthonormal eigenvectors of YY' by definition (the $\sqrt{T}$ term is to make sure that $(\tilde{F}^{r'}\tilde{F}^r/T) = I_r$), we can write:

$$\frac{1}{T}YY'\tilde{F}^r = \tilde{F}^r\tilde{V}$$

where $\tilde{V}$ is the diagonal matrix whose diagonal consists of the r largest eigenvalues of YY'/T. Clearly, the eigenvalues of YY'/T are the same as the eigenvalues of $\tilde{\Sigma} = Y'Y/T$, provided that $r < \min(N,T)$. Therefore, as introduced in the preliminaries of section 2: $\tilde{V} = \text{diag}(\tilde{v}_1, ..., \tilde{v}_r)$.

We use a similar decomposition in Bai (2003): Let $H^r = \tilde{V}^{-1}\tilde{F}^{r'}F\Lambda'\Lambda/T$

$$\begin{aligned}
\tilde{F}^r - FH^{r'} &= \frac{1}{T}YY'\tilde{F}^r\tilde{V}^{-1} - \frac{1}{T}F\Lambda'\Lambda F'\tilde{F}^r\tilde{V}^{-1} \\
&= \left(\frac{YY'}{T} - \frac{F\Lambda'\Lambda F'}{T}\right)\tilde{F}^r\tilde{V}^{-1} \\
&= \left(\frac{YY'}{T} - \frac{F\Lambda'\Lambda F'}{T}\right)\tilde{F}^r D_N^{-1}D_N\tilde{V}^{-1} \\
&= \left(\frac{F\Lambda'U'}{T} + \frac{U\Lambda F'}{T} + \frac{UU'}{T}\right)\tilde{F}^r D_N^{-1}D_N\tilde{V}^{-1}
\end{aligned}$$

Notice that $\frac{1}{T}\sum_{t=1}^{T}\left\|\tilde{f}_s^r - H^r f_t\right\|^2 = \frac{1}{T}\left\|\tilde{F}^r - FH^{r'}\right\|_F^2$, so we can prove that $\frac{1}{T}\left\|\tilde{F}^r - FH^{r'}\right\|_F^2 = O_p(\frac{N}{[d_r(N)]^2} + \frac{N^2}{T[d_r(N)]^2})$. In addition, notice that $\tilde{F}^r - FH^{r'}$ is a $T \times r$ matrix, so it only has r non-zero singular values, where r is a finite number. Therefore, we can use the operator norm in the subsequent proofs, which makes the notation easier.

For part (i) of theorem 2.1, we shall prove that $\frac{1}{T}\left\|\tilde{F}^r - FH^{r'}\right\|^2 = O_p(\frac{N}{[d_r(N)]^2} +$

$\frac{N^2}{T[d_r(N)]^2}$) via Cauchy Schwartz inequality:

$$\frac{1}{T}\left\|\tilde{F}^r - FH^{r'}\right\|^2 \leq \left\|D_N\tilde{V}^{-1}\right\|^2 \frac{1}{T}\left\{\left\|\frac{1}{T}F\Lambda'U'\tilde{F}^r D_N^{-1}\right\|^2 + \left\|\frac{1}{T}U\Lambda F'\tilde{F}^r D_N^{-1}\right\|^2 + \left\|\frac{1}{T}UU'\tilde{F}^r D_N^{-1}\right\|^2\right\}.$$

From here, the (i) part of theorem 2.1 is proven by lemma 2.1, 2.2 and 2.3. The dominated term in these is $\left\|\frac{1}{T}UU'\tilde{F}^r D_N^{-1}\right\|^2$, which has order $O_p(\frac{N}{[d_r(N)]^2} + \frac{N^2}{T[d_r(N)]^2})$, the remaining 2 terms in the big curly bracket has order $O_p(\left\|D_N^{-1}\right\| + \frac{1}{T}) = O_p(\frac{1}{d_r(N)} + \frac{1}{T})$, whereas $\left\|D_N\tilde{V}^{-1}\right\| = O_p(1)$.

For the (ii) part, first we notice that H^r has eigenvalues bounded from both 0 and ∞ by lemma 2.4, hence its inverse G^r exists. Similarly with the first part, we will prove using the operator norm of the matrix $\tilde{\Lambda}^r - \Lambda G^k$ for easier notation. We now use the following decomposition:

$$\begin{aligned}
\tilde{\Lambda}^r &= Y'\tilde{F}^r/T \\
&= \left(\Lambda G^r H^r F' + U'\right)\tilde{F}^r/T \\
&= \Lambda G^r\left(H^r F' - \tilde{F}^{r'} + \tilde{F}^{r'}\right)\tilde{F}^r/T + U'\left(\tilde{F}^r - FH^{r'} + FH^{r'}\right)/T \\
&= \Lambda G^r + \Lambda G^r\left(H^r F' - \tilde{F}^{r'}\right)\tilde{F}^r/T + U'\left(\tilde{F}^r - FH^{r'}\right)/T + U'FH^{r'}/T.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{N}\left\|\tilde{\Lambda}^r - \Lambda G^k\right\|^2 &\leq \frac{1}{NT^2}\left\|\Lambda G^k\left(H^k F' - \tilde{F}^{r'}\right)\tilde{F}^r\right\|^2 \\
&+ \frac{1}{NT^2}\left\|U'\left(\tilde{F}^r - FH^{r'}\right)\right\|^2 \\
&+ \frac{1}{NT^2}\left\|U'FH^{r'}\right\|^2 \\
&\leq \frac{1}{T}\left\|H^k F' - \tilde{F}^{r'}\right\|^2 \left\|\frac{\tilde{F}^{r'}\tilde{F}^r}{T}\right\| \left\|\frac{\Lambda'\Lambda}{N}\right\| \left\|G^k\right\|^2 \\
&+ \left\|\frac{UU'}{NT}\right\| \frac{1}{T}\left\|\tilde{F}^r - FH^{r'}\right\|^2 \\
&+ \frac{1}{NT^2}\left\|U'F\right\|^2 \left\|H^{r'}\right\|^2 \\
&= O_p\left(\frac{N}{[d_r(N)]^2} + \frac{N^2}{T\,[d_r(N)]^2}\right)
\end{aligned}$$

by theorem 2.1, and the fact that $\left\|G^k\right\| = O_p(1)$ if $\frac{1}{T}\left\|H^k F' - \tilde{F}^{r'}\right\|^2 = o_p(1)$, $\left\|H^{r'}\right\| = O_p(1)$. Furthermore, $\left\|\frac{UU'}{NT}\right\|$ is bounded above as U has finite entries and the size of U is $N \times T$. For similar reason, $\frac{1}{NT}\left\|U'F\right\|^2$ is bounded away from infinity, which establishes the final result above.

2.5.2 Technical Lemmas

Lemma 2.1. *Under assumption 1-5, $D_N\tilde{V}^{-1}$ has eigenvalues asymptotically bounded away from both 0 and ∞, where as introduced in the preliminaries: $\tilde{V} = diag(\tilde{v}_1, ..., \tilde{v}_r)$, $\tilde{v}_1$ is the ith largest eigenvalues of $\tilde{\Sigma}$.*

Proof. We have:

$$D_N\tilde{V}^{-1} \equiv \begin{bmatrix} \frac{d_1(N)}{\tilde{v}_1} & & 0 \\ & \ddots & \\ 0 & & \frac{d_r(N)}{\tilde{v}_r} \end{bmatrix}$$

It suffices to prove that $\tilde{v}_i \asymp d_i(N)$ for every $i \in (1, ..., r)$. First of all, we need to state **Weyl's Theorem**: Let $\{a_i\}_{i=1}^N$ be the eigenvalues of A in descending order. Correspondingly, let $\{b_i\}_{i=1}^N$ be the eigenvalues of B. Then for all $i \leq N$, $|a_i - b_i| \leq$

$\|A - B\|$.

If $\mu_i(\Lambda'\Lambda)$ is the ith largest eigenvalue of $\Lambda'\Lambda$ and $\Lambda\Lambda'$. So by Weyl's Theorem, we have:

$$\left|v_i - \mu_i(\Lambda'\Lambda)\right| \leq \left\|\Sigma - \Lambda\Lambda'\right\| = \|\Sigma_u\| = O_p(1).$$

Therefore since $\mu_i(\Lambda'\Lambda) \asymp \left\|\Lambda^{(i)}\right\| \asymp d_i(N)$ as in assumption 3(ii), $v_i \asymp d_i(N)$. Now, again by Weyl's Theorem:

$$\left|\frac{v_i - \tilde{v}_i}{d_i(N)}\right| \leq \left\|\Sigma - \tilde{\Sigma}\right\| / d_i(N) = O_p(\frac{N}{d_i(N)}\sqrt{\frac{\log N}{T}}).$$

The result $\left\|\Sigma - \tilde{\Sigma}\right\|$ can be found for example in Fan et al. (2011). Since by assumption 4(ii) we have that $\frac{N}{d_i(N)}\sqrt{\frac{\log N}{T}} = o(1)$, $v_i \asymp d_i(N)$ implies $\tilde{v}_i \asymp_p d_i(N)$. □

Lemma 2.2. $\frac{1}{T}\left\|\frac{1}{T}F\Lambda'U'\tilde{F}^r D_N^{-1}\right\|^2 = O_p(\|D_N^{-1}\|)$ *and* $\frac{1}{T}\left\|\frac{1}{T}U\Lambda F'\tilde{F}^r D_N^{-1}\right\|^2 = O_p(\|D_N^{-1}\|)$.

Proof. We have that:

$$\begin{aligned}
\frac{1}{T}\left\|\frac{1}{T}F\Lambda'U'\tilde{F}^r D_N^{-1}\right\|^2 &\leq \|D_N^{-1}\|\,\frac{1}{T}\left\|D_N^{-1/2}\Lambda'U'\right\|^2 \left\|\frac{1}{T}\tilde{F}^{r\prime}\tilde{F}^r\right\| \left\|\frac{1}{T}F'F\right\| \\
&\leq \|D_N^{-1}\|\,\frac{1}{T}\sum_{t=1}^{T}\left\|D_N^{-1/2}\Lambda' u_t\right\|^2 \\
&= O_p(\|D_N^{-1}\|)
\end{aligned}$$

as by assumption 4 (iii) $\left\|D_N^{-1/2}\Lambda' u_t\right\|^2 = O_p(1)$. Similarly, $\frac{1}{T}\left\|\frac{1}{T}U\Lambda F'\tilde{F}^r D_N^{-1}\right\|^2 = O_p(\|D_N^{-1}\|)$. □

Lemma 2.3. $\frac{1}{T}\left\|\frac{1}{T}UU'\tilde{F}^r D_N^{-1}\right\|^2 = O_p\left(\frac{N}{[d_r(N)]^2} + \frac{N^2}{T[d_r(N)]^2}\right)$

Proof. We have that:

$$\begin{aligned}
\frac{1}{T}\left\|\frac{1}{T}UU'\tilde{F}^r D_N^{-1}\right\|^2 &\leq \left\|D_N^{-1}\right\|^2 \left\|\frac{1}{T}UU'\right\|^2 \left\|\frac{1}{T}\tilde{F}^{r\prime}\tilde{F}^r\right\| \\
&= \frac{1}{[d_r(N)]^2}\left\|\frac{1}{T}UU'\right\|^2 \\
&\leq \frac{1}{[d_r(N)]^2}\left(\max_s \frac{1}{T}\sum_{t=1}^{T} u_s' t_t\right)^2 \\
&= \frac{1}{[d_r(N)]^2}\left(\max_s \left[\frac{1}{T}\sum_{t=1}^{T} u_s' t_t - E(u_s' t_t) + E(u_s' t_t)\right]\right)^2 \\
&\leq \frac{2}{[d_r(N)]^2}\left(\max_s \left[\frac{1}{T}\sum_{t=1}^{T} u_s' t_t - E(u_s' t_t)\right]\right)^2 \\
&+ \frac{2}{[d_r(N)]^2}\left(\max_s \frac{1}{T}\sum_{t=1}^{T} E(u_s' t_t)\right)^2 \\
&= \frac{2N}{[d_r(N)]^2}\frac{1}{N}\left(\max_{s,t}\left[u_s' t_t - E(u_s' t_t)\right]\right)^2 \\
&+ \frac{2N^2}{T\,[d_r(N)]^2}\left(\max_s \frac{1}{N}\sum_{t=1}^{T} E(u_s' t_t)\right)^2 \\
&= O_p\left(\frac{N}{[d_r(N)]^2} + \frac{N^2}{T\,[d_r(N)]^2}\right)
\end{aligned}$$

where we use assumption 4 (ii) to have $\frac{1}{N}\left(\max_s\left[u_s' t_t - E(u_s' t_t)\right]\right)^2 = O_p(1)$ and lemma 6 in Fan et al. (2013) to have $\max_s \frac{1}{N}\sum_{t=1}^{T}\left|E(u_s' t_t)\right| = O(1)$. □

Lemma 2.4. *H^r has eigenvalues bounded from both 0 and ∞. error mean*

Proof. $H^r = \tilde{V}^{-1}\tilde{F}^{r\prime}F\Lambda'\Lambda/T = D_N\tilde{V}^{-1}\tilde{F}^{r\prime}FD_N^{-1}\Lambda'\Lambda/T$. It is easy to see that $\|H^r\| = O_p(1)$, which is equivalent to all eigenvalues of H^r is strictly less than ∞because:

$$\|H^r\| \leq \left\|D_N\tilde{V}^{-1}\right\|\left\|\frac{1}{T}\tilde{F}^{r\prime}\tilde{F}^r\right\|^{1/2}\left\|\frac{1}{T}F'F\right\|^{1/2}\left\|D_N^{-1}\Lambda'\Lambda\right\| = O_p(1).$$

On the other hand, by a same technique in Bai and Liao (2013), we have:

$$I_r = \tilde{F}^{r'}\tilde{F}^r = \tilde{F}^{r'}\frac{1}{T}\left(\tilde{F}^r - FH^{r'}\right) + \frac{1}{T}\left(\tilde{F}^r - FH^{r'}\right)' FH^{r'} + H^r\frac{F'F}{T}H^{r'}.$$

By theorem 2.1 (i), we already have $\frac{1}{T}\left\|\tilde{F}^r - FH^{r'}\right\|^2 = O_p\left(\frac{N}{[d_r(N)]^2} + \frac{N^2}{T[d_r(N)]^2}\right) = o_p(1)$, so:

$$H^r\frac{F'F}{T}H^{r'} \to H^r\Sigma_f H^{r'} = I_r + o_p(1)$$

hence the eigenvalues of H^r must also be bounded from 0 because Σ_f has eigenvalues bounded from 0 and ∞. □

3 Determining the number of factors

As discussed earlier, it is not easy to adapt the current methods in the literature to determine the number of factors when the factors are weak. For example, consider the Bai and Ng (2002) criteria, i.e. we choose k that maximises the following:

$$\frac{1}{NT}\sum_{t=1}^{T}(Y_t - \tilde{\Lambda}^k \tilde{f}_t^k)'(Y_t - \tilde{\Lambda}^k \tilde{f}_t^k) + kg(N,T) \tag{3.1}$$

where $\tilde{\Lambda}^k$ and $\tilde{f}_t^k$ are estimated by PCA as above, and $g(N,T)$ is a function that converges to 0 at a slower rate than $O(N^{-1} + T^{-1})$. To improve the performance, Alessi et al. (2007) propose to scale $kg(N,T)$ with a tuning parameter. This is shown to refine the estimation with finite samples.

To see why this class of criteria may potentially fails due to the existence of weak factors, considering the following interpretation: as we assume u_t has zero-mean,

$$\begin{aligned}\frac{1}{NT}\sum_{t=1}^{T}(Y_t - \tilde{\Lambda}^k \tilde{f}_t^k)'(Y_t - \tilde{\Lambda}^k \tilde{f}_t^k) &= \frac{1}{NT}\text{trace}(Y - \tilde{F}^k\tilde{\Lambda}^{k'})'(Y - \tilde{F}^k\tilde{\Lambda}^{k'}) \\ &= \frac{1}{N}\text{trace}(\tilde{\Sigma}_u^k)\end{aligned}$$

where $\tilde{\Sigma}_u^k$ is the residuals sample covariance matrix given k PCs are extracted (see again the preliminaries at chapter 2 for the notations). Standard matrix algebra tells us that the trace of a matrix is the sum of all eigenvalues. Therefore, the first term in Bai and Ng (2002) can be seen as the average of all eigenvalues of $\tilde{\Sigma}_u^k$.

Suppose we ignore the sampling error so that $\frac{1}{N}\text{trace}(\Sigma_u^k)$ is used as a criterion

in stead of $\frac{1}{N}\text{trace}(\tilde{\Sigma}_u^k)$. Under Assumption 0, if we do not extract up to r factors by PCA, some systematic eigenvalues with order $\asymp N$ are not extracted and hence $\frac{1}{N}\text{trace}(\Sigma_u^k))$ is bounded away from zero, which then dominates the term $g(N,T)$. On the other hand, if more than r factors are extracted, no more systematic eigenvalues absorbed in Σ_u^k hence $\frac{1}{N}\text{trace}(\Sigma_u^k)$ can be shown to be dominated by $g(N,T)$. The key idea behind the proof in Bai and Ng (2002) is to work out the asymptotic value for $\frac{1}{N}\text{trace}(\tilde{\Sigma}_u^k)$ in place of $\frac{1}{N}\text{trace}(\Sigma_u^k)$, which then must take into account of the sampling size T.

Now in our case, where some eigenvalues of Σ have order $d_i(N)$ which can be $o(N)$, the choice of $g(N,T)$ needs to be adjusted carefully to separate the edge of the rth and the $(r+1)$th largest eigenvalues of Σ. Given that it is possible, I think the Bai and Ng (2002) can be adapted to deal with this problem. However, in this thesis I will propose a different approach which exploit a different property of $\tilde{\Sigma}_u^k$.

Another popular direction in determining the number of factors is by using the eigenvalues of $\tilde{\Sigma}$, i.e. the sample covariance matrix of Y_t. This approach has been originated by the scree test of Cattell (1966). This is a visual test based on the behaviors of the eigenvalues of $\tilde{\Sigma}$. The idea is to plot all the eigenvalues of $\tilde{\Sigma}$ in the descending order and spot the sharpest edge (elbow) between 2 eigenvalues. Recently, Forni et al. (2000) also proposed a visual test based on the behavior of the eigenvalues for determining the number of factors in the context of dynamic factor models. More similar methods are described in Onatski (2010) and Ahn and Horenstein (2013). For example, in Ahn and Horenstein (2013), $\tilde{r}_{ER} = \text{argmax}_{k<kmax}(\mu_k(\tilde{\Sigma})/\mu_{k+1}(\tilde{\Sigma}))$ is used as an estimator for r.

These methods based on the sample eigenvalues are empirically shown to work well under a wide range of factor models, including the case when the idiosyncratic errors have both serial and cross-section correlations. Another advantage of this approach is that we do not need to specify a penalty function and the tuning parameter. However, if we relax the strong factor assumption and replace it with the mix model

of strong and weak factors, the eigenvalues-ratio based will potentially fail if some factors are not strong enough. Particularly, if the noise-to-signal is not higher than the ratio between 2 systematic eigenvalues, $\tilde{r}_{ER}$ can not be consistent.

Technically speaking, there is a strong link between the Bai and Ng (2002) criteria and the eigenvalues based method (see Onatski (2010) and Ahn and Horenstein (2013) for further discussions and comparisons). Hence both suffer from the weaker assumption for the pervasiveness of the factors. On the other hand, we already see that the factors can not be too weak otherwise they can not be estimated consistently (see Theorem 2.1). Therefore, if the strengths of all common factors are assumed accordingly, i.e. $\max(\sqrt{N}, N\sqrt{\log N/T}) = o(d_r(N))$, the ratio between the systematic and idiosyncratic eigenvalues is at least $\asymp \sqrt{N}$. Therefore the performance of the eigenvalues ratio based method should work in most cases, because the ratio of two systematic eigenvalues can not be greater than a magnitude of order $\asymp \sqrt{N}$. This agrees with our simulation results, which show that the eigenvalues ratio method usually performs well if the factors are not too weak.

Recently, Fan et al. (2014) show that the eigenvalues ratio method is robust when the factors are weak. Particularly, they show that if all the factors have strength with order N^{α} for some $\alpha \in (0, 1]$, the eigenvalues ratio method still gives consistent estimator for the number of factors. However, they still assume that all the factors have same strength, i.e. all the systematic eigenvalues scaled by $N^{-\alpha}$ must be bounded away from 0 and ∞. This certainly is more stronger than our assumption in this paper.

In summary, most current methods in the literature do not take into account the case where all the factors have different strengths, some are pervasive and some are weaker. This is a key motivation for the work in this chapter. Particularly, in section 3.1 I discuss a novel approach in determining the number of factors that is consistent even when the factors have different strengths. The key requirement is that there must exist a diverging sequence which is $o(d_r(N))$.

In the next section, a metric called sparsity level is proposed, which measures the amount of pairwise correlations between the idiosyncratic errors. From a model selection point of view, adding more factors to the model should increase the sparsity level in the conditional idiosyncratic components[1]. The final criterion is obtained in the similar manner with the Bai and Ng (2002) criteria or the AIC, where we add a penalty function to the sparsity level.

3.1 Determining the number of factors by sparsity level

To the best of my knowledge, Bickel and Levina (2008) originate the first paper that makes use of the sparsity structure and provides a consistent estimator for a sparse covariance matrix. However, they propose a sparse structure for a general covariance matrix, not the idiosyncratic error covariance matrix[2]. In this chapter, as in Bickel and Levina (2008), I define the sparsity level as follows: for a matrix $A = (a_{ij})$, $i = 1 : m$, $j = 1 : n$,

$$m(A) = \max_i \sum_{j=1}^{n} \mathrm{I}(a_{ij} \neq 0), \tag{3.2}$$

which is the maximum number of non-zero entries across all rows of A. For this measure, the smaller $m(A)$ is, the more sparse A is. A more general measure for sparsity level replaces $\mathrm{I}(a_{ij} \neq 0)$ with $|a_{ij}|^q$ for some $q \in [0, 1]$. Notice that the definition in (3.2) is a special case of $\max_i \sum_{j=1}^{n} |a_{ij}|^q$ for $q = 0$.

Clearly, assuming the covariance matrix of Y_t to be sparse is not realistic since it is possible that all components in Y_t are correlated with each other. However, if we remove the common factors from the original components, the idiosyncratic errors are more likely to be uncorrelated with each other. As a result, imposing the sparsity condition to Σ_u is more reasonable.

This paper is not the first work involving sparse idiosyncratic covariance matrix.

[1] which is the original component subtracting the common component.

[2] The key difference is that the idiosyncratic errors are not observed, see Fan et al. (2011, 2013) for the work regarding the idiosyncratic covariance matrix.

In fact, a standard assumption in the current literature is to have an upper bound for $m(\Sigma_u)$. For example, Fan et al. (2011) propose a sparsity structure for the idiosyncratic error covariance matrix. They assume that for a given observed factor model (such as the Fama-French 3-factor model) the sparsity level of Σ_u must be $o(\sqrt{T/(r^2 \log N)})$. Exploiting this condition, they provide an estimator for Σ_u and thence Σ. Fan et al. (2013) extend the result of their 2011 paper by proposing the sparsity structure for the idiosyncratic errors covariance matrix when a PC factor model is applied to the data. Notice that the estimator for sparse Σ_u is important for some applications other than for estimating Σ. For instance, Bai and Liao (2013) use the estimator for Σ_u (assuming that it is sparse) to compute weighted principle components as more efficient estimator for the latent factors.

There are two key remarks in this chapter. Firstly, we have a stronger sparsity assumption than in Fan et al. (2011, 2013), i.e. $m(\Sigma_u)$ is bounded when N grows to infinity. Secondly, the main focus of this chapter is not to use the sparsity assumption to estimate Σ_u, it is to use the sparsity assumption to select the number of factors in the data by estimating the sparsity level of Σ_u^k for a range of k.

Clearly, by definition, estimating $m(\Sigma_u^k)$ requires methods of identifying the exact zero entries in Σ_u^k, which usually can be done with the thresholding technique. Particularly, Bickel and Levina (2008) apply hard universal thresholding, in which all entries that have magnitude less than a single value are forced to zero. Cai and Liu (2011) proposes the adaptive hard thresholding technique, where the values of threshold vary from entries to entries. More general form of threshold include the smoothly clipped absolute deviation (SCAD), soft thresholding, the adaptive lasso etc., which can be found in Rothman et al. (2009) and the references therein.

In this chapter, I use the adaptive hard thresholding to estimate the sparsity level, which is defined as:

$$\tilde{m}(\Sigma_u^k) = \max_{i \leq N} \sum_{j=1}^{N} \mathbf{I}\left(\left|\tilde{\sigma}_{ij}^k\right| > h_{ij}^k\right) \tag{3.3}$$

where $h_{ij}^k = C_1 \omega_T \sqrt{\tilde{\theta}_{ij}^k}$, in which C_1 is a tuning constant, $\tilde{\theta}_{ij}^k$ is an adaptive parameters that must be asymptotically bounded between 0 and ∞, i.e.:

$$\exists\,(C_L, C_H) \text{ such that } \forall (i,j),\ \mathbf{P}\left(C_L \leq \tilde{\theta}_{ij}^k \leq C_H\right) = 1, \tag{3.4}$$

and

$$\omega_T = \sqrt{\frac{\log N}{T}} + \frac{\sqrt{N}}{[d_r(N)]} + \frac{N}{\sqrt{T}\,[d_r(N)]}. \tag{3.5}$$

Some choices for $\tilde{\theta}_{ij}^k$ is $T^{-1}\sum_{t=1}^{T}(\tilde{u}_{it}^k \tilde{u}_{jt}^k - \tilde{\sigma}_{ij}^k)^2$ as in Cai and Liu (2011) or $\tilde{\sigma}_{ii}^k \tilde{\sigma}_{jj}^k$, which both can be shown to satisfy the asymptotic bounded requirement in (3.4). Notice that if $\tilde{\theta}_{ij}^k = \tilde{\sigma}_{ii}^k \tilde{\sigma}_{jj}^k$, then we just thresholding the sample correlation matrix of $\tilde{u}_t^k$ by a universal thresholding value $C_1\omega_T$. Furthermore, in order to consistently identify the non-zero entry, we will need an extra assumption that require all the non-zero entries of Σ_u^k are not too small, which is shown below.

Assumption 6. *$\forall k \leq r, \sigma_{ij}^k \neq 0 \Leftrightarrow \left|\sigma_{ij}^k\right| > \tau = C'\omega_T$, for a sufficiently large constant C'.*

In words, the assumption above requires all the non-zero entries in Σ_u^k are bounded away from zero at a certain level, which is needed to correctly identify the non-zero entries of Σ_u. Since ω_T is converging toward zero when N and T approaches infinity, this assumption is reasonable for large dimension.

When we apply the sparsity assumption to the idiosyncratic errors covariance matrix, the intuition is that after a correct factor model is specified, the conditional idiosyncratic errors must have low level of cross-section correlation. We formulate that intuition into Lemma 3.1 as follows.

Lemma 3.1. *Under Assumptions 1-6, $m(\Sigma_u^k)$ is bounded as $N \to \infty$ if and only if $k = r$.*

The above lemma is useful in connecting the factor model assumption with the newly introduced sparsity level measure. We exploit this property to find the number

of factors, which should be taken as the first number when the conditional sparsity level is bounded. In this paper, we use the sparsity level as a metric to select the number of factors, instead of the sample eigenvalues ratio or the mean square errors. Using it has some advantages over the others with the first one being that it actually has meaning on its own, so we can directly interpret the criterion value and see how good a chosen factor model is. Secondly, although $m(\Sigma_u^k)$ is not observed so we need to provide its estimator, it turns out in Theorem 3.1 that our estimator has a good rate of convergence toward the true number.

Notice that although u_t^k and Σ_u^k are only defined for $k \leq r$, $\tilde{u}_t^k$ and $\tilde{\Sigma}_u^k$ are available for any non-negative value of k less than $\min(N, T)$. Therefore, we also need to investigate $\tilde{m}(\Sigma_u^k)$ even when $k > r$. This is achieved in Theorem 3.1.

Theorem 3.1. *Under Assumptions 1-6, if $N = o(T^2)$ then $\tilde{m}(\Sigma_u^k)$ is bounded for $k \geq r$ and grows to infinity with rate at least $d_r(N)$ when $k < r$.*

Based on Theorem 3.1, we can estimate the number of factors by adding the penalty function to $\tilde{m}(\Sigma_u^k)$. By Assumption 4 (i), we know that there exists a function $g(N)$ which can be used to separate $d_r(N)$ and a bounded sequence. In this case, r will be estimated using the corollary below.

Corollary 3.1. *Under assumption 1-6, if we define $\tilde{r} = argmin_k \left\{\tilde{m}(\Sigma_u^k) + C\, k\, g(N)\right\}$ for some constant C, then with probability tending to 1, we have $\tilde{r} = r$.*

Intuitively, this criteria is similar to the ones in Bai and Ng (2002), where the term $\tilde{m}(\Sigma_u^k)$ corresponds to how good a model is and $C\, k\, g(N)$ penalises the complexity of the model.

3.2 Choices of threshold and penalty functions

The proposed criterion in this thesis consists of two parts, the estimated level of sparsity $\tilde{m}(\Sigma_u^k)$ and the penalty term $C\, k\, g(N)$. Each of these terms has its own tun-

ing constant and unknown quantity so I will in turn discuss them in the subsequent sections.

3.2.1 Thresholding value

Recall that the thresholding value used here is $h_{ij}^k = C_1\omega_T\sqrt{\tilde{\theta}_{ij}^k}$. The most important work in finding this value is to work out the theoretical bound for ω_T as in (3.5). However, as we do not observe $d_r(N)$ in practice, the choice of ω_T is subjective and should reflect the prior belief regarding the strength of the weakest factor. In this paper I propose to use the thresholding value of $\sqrt{\frac{\log N}{T}} + \frac{1}{N^{1/4}} + \frac{N^{1/4}}{\sqrt{T}}$ (denoted as SC1) which corresponds to the case where $d_r(N) = N^{3/4}$. Due to simulations result, I also suggest using $\sqrt{\frac{\log N}{T}} + \frac{1}{N^{1/4}}$ (denoted as SC2), which although is not quite the theoretically required value, does better in some cases.

Regarding the constant C_1, I use a conservative choice of $\frac{1}{2}$ in SC1 and 1 in SC2 for simulations in this paper. Notice that the asymptotic consistency of the estimators does not depend on the value for C_1. However, clearly a more data-driven way for choosing C_1 is better in practice.

In the literature of thresholding sparse covariance matrix, the choice of C_1 is also discussed, e.g. in Fan et al. (2013), Bickel and Levina (2008) and Cai and Liu (2011). Their focus is to improve the performance of thresholding in order to obtain a closed estimator for the true covariance matrix, which also theoretically improves the estimation for the sparsity level. Therefore, we can apply the following cross-validation procedure suggested by Bickel and Levina (2008) to determine C_1 as follows:

- First we split our residuals $\left\{u_t^k\right\}_{t=1}^T$ into 2 part for cross-validation purpose, the length for each part should be $T_1 = \frac{T}{\log T}$ and $T_2 = T - T_1$.

- We construct the thresholded sample covariance matrix $\tilde{\Sigma}_u^{k,\tau}(T_1)$ based on a trial value of C_1 and the residuals set which has T_1 elements. The rule for

thresholding is the adaptive hard thresholding as we discussed earlier in this paper, with thresholding value $h_{ij}^k = C_1 \omega_T \sqrt{\tilde{\theta}_{ij}^k}$.

- We find C such that the distance between $\tilde{\Sigma}_u^{k,\tau}(T_1)$ and $\tilde{\Sigma}_u^k(T_2)$ is minimal, where $\tilde{\Sigma}_u^k(T_2)$ is the sample covariance matrix of the residuals set which has T_2 elements. . The distance between two matrices is measured by the Frobenius norm.

- To get better result, we should repeat the cross-validation procedure many times and choose the value C_1 such that the average of the Frobenius norms of $\tilde{\Sigma}_u^{k,\tau}(T_1) - \tilde{\Sigma}_u^k(T_2)$ for different splits is minimised.

If one has chosen to use the value of C_1 suggested by the procedure described above, it is possible to see some improvements in the sparsity level estimation, and hence also for the number of factors. The rate of convergence can also be revised with the data-driven value for C_1. However, to keep our simulations simple I do not use this procedure for selecting C_1.

3.2.2 The penalty function

A second part in the criterion is the penalty function, and the tuning parameter going with it. Recall that $g(N)$ needs to grow to infinity at a slower rate than $d_r(N)$. In practice, I suggest to use $\sqrt{N}$ as the value for $g(N)$ and let $C = 1/10$. However, this tuning parameter can be defined in a data driven way, which is left for further research. At the moment, our simulations show these choices achieve relatively good results comparing to the existing methods. Similar approach to work out C can be done as in Alessi et al. (2008), in which the authors also suggest multiplying a penalty function of Bai and Ng (2002) criteria with a constant.

3.3 Monte Carlo Simulations

I start with simulating some standard factor models with dimensions easily seen in practice. The three methods used here are the Sparsity criterion (SC1 and SC2), the eigenvalues ratio (ER) method of Ahn and Horenstein (2013) and the BIC_3 in Bai and Ng (2002)[3]. We summarise all the criteria for choosing the number of factors as follow:

- Sparsity criterion: I use adaptive thresholding with $\tilde{\theta}_{ij}^k = \tilde{\sigma}_{ii}^k \tilde{\sigma}_{jj}^k$, therefore I choose r by the following 2 objective functions, corresponding to each case of the thresholding function:

$$\text{SC1:} \quad \text{argmin}_k \left\{ \max_{i \leq N} \sum_{j=1}^{N} \mathbf{I} \left(\frac{\left|\tilde{\sigma}_{ij}^k\right|}{\left|\tilde{\sigma}_{ii}^k \tilde{\sigma}_{jj}^k\right|} > \frac{1}{2} \left(\sqrt{\frac{\log N}{T}} + \frac{1}{N^{1/4}} + \frac{N^{1/4}}{\sqrt{T}} \right) \right) + \frac{k\, N^{1/2}}{10} \right\} \tag{3.6}$$

$$\text{SC2:} \quad \text{argmin}_k \left\{ \max_{i \leq N} \sum_{j=1}^{N} \mathbf{I} \left(\frac{\left|\tilde{\sigma}_{ij}^k\right|}{\left|\tilde{\sigma}_{ii}^k \tilde{\sigma}_{jj}^k\right|} > \left(\sqrt{\frac{\log N}{T}} + \frac{1}{N^{1/4}} \right) \right) + \frac{k\, N^{1/2}}{10} \right\} \tag{3.7}$$

- Eigenvalues ratio: we choose r by

$$\text{argmax}_{k<kmax}(\mu_k(\tilde{\Sigma})/\mu_{k+1}(\tilde{\Sigma})) \tag{3.8}$$

  with the dummy case for $k = 0$ is set as $\mu_0(\tilde{\Sigma}) = \left(\sum_{k=1}^{kmax} \mu_k(\tilde{\Sigma})\right) / \log N$.

- BIC_3: we choose r by

$$\frac{1}{NT} \sum_{t=1}^{T} (Y_t - \tilde{\Lambda}^k \tilde{f}_t^k)'(Y_t - \tilde{\Lambda}^k \tilde{f}_t^k) + k\hat{\sigma}^2 \left(\frac{(N+T-k)\log(NT)}{NT} \right) \tag{3.9}$$

  where $\hat{\sigma}^2 = \sum_{t=1}^{T} (Y_t - \tilde{\Lambda}^{kmax} \tilde{f}_t^{kmax})'(Y_t - \tilde{\Lambda}^{kmax} \tilde{f}_t^{kmax})$.

[3] The reason we choose the BIC_3 is because it usually outperforms the rest of BN criteria, as shown in Bai and Ng (2002). Also, we see the note that Prof. Juhsan Bai recommends using for comparisons made in Ahn and and Horenstein (2013).

3.3.1 Simulated Scenarios for Comparing

3.3.1.1 Weakening signal-to-noise ratio

I simulate a mix model between strong and weak model by the following data generation process:

- $Y_t = \Lambda^{(1:m)} f_t^{(1:m)} + \gamma \Lambda^{(m+1:r)} f_t^{(m+1:r)} + u_t$

The generation Λ, f_t, u_t are described in Section 2.4 of Chapter 2. In here, some of the control parameters are:

- α: the serial correlation level for the factors.
- β: the serial correlation level for the idiosyncratic error.
- γ: the strength of the factors (which controls the signal-to-noise ratio), so to weaken the signal-to-noise ratio we decrease γ. $\gamma = 1$ corresponds to the all-strong-factor case.
- m: the number of strong factors.
- r: the total strong and weak factors (which we fix to be 5).

So in general there are m strong factors and $r - m$ weak factors. Different scenarios for α, β, γ and m are simulated with different values of cross-section and sample sizes. Tables 3.1-3.6 report the results of the mixture models of strong and weak factors ($m = 2$). Section 3.6 has additional tables for the case of all-strong ($m = 5$) or all-weak factors ($m = 0$). For each value of N, T in each scenario (table), the result reported is the average number of factors estimated by different methods from 500 repeated simulations.

3.3.1.2 Regional factors

Notice that when multiplying the weak factors by γ we decrease the value of all systematic eigenvalues and therefore make the signal-to-noise ratio lower. On the

other hand, it is more interested to consider the case that is more likely to come across in practice. We go back to the aforementioned case in the literature review (Section 1.1.1.2, Figure 1.4) where $r = 3$ and

$$\Lambda = \begin{bmatrix} \pi_1 & 0 & 0 \\ 0 & \pi_2 & 0 \\ 0 & 0 & \pi_3 \end{bmatrix}$$

where π_i is the $N_i \times 1$ vectors of loadings, and $N_1 + N_2 + N_3 = N$. For this simulation we choose $N = 200$, $N_1 = 100$, $N_2 = 50$, $N_3 = 50$, i.e. after simulate random $N \times r$ loadings matrix Λ, we force parts of the entries to 0 as shown in the matrix Λ above.

Then we generate f_t and u_t as previous section and let $Y_t = \Lambda f_t + u_t$. It means that the first 100 components of Y_t are generated from the first factor, the next 50 components of Y_t are generated by the second factor and so on. In this case I increase the sample size from 100 up to 500 while keeping the cross-section size of 200, this is shown in Tables 3.7 and 3.8.

3.3.2 Comparisons between methods

3.3.2.1 Weakening signal-to-noise ratio

For all cases regarding the parameter γ, we include in our simulations the sub-cases with and without serial correlations in f_t and u_t. We provide our results in the tables shown below. The interesting cases and most relevant to our model are reported in Tables 3.1-3.6, where we include both strong and weak factors in the model. In most cases, our criterion correctly determines five factors, although only the first two are strong. When some factors are very weak ($\gamma = 1/10$), it can be seen from Table 3.6 that under reasonable values of N and T, only our criterion selects the true number of factors whereas other methods fail to capture the weak factors.

In Section 3.6, I show additional tables for simulation results without mix-strength factors. When the factors are all strong, the BIC_3 criterion is extremely accurate in

determining the number of factors. However, its performance drops when the factors get less pervasive. For the case $\gamma < 1/5$, BIC_3 consistently results in zero factors. Therefore it is not recommended to use BIC_3 when the factors may not be pervasive. On the other hand, ER is much more robust to weak factors as it still be able to pick the near the true number when $\gamma = 1/10$. In fact, when the factors all have the same strengths, ER can work well even the strength of the factors grow slower than N.

In theory, I have shown that the number of factors estimated by $SC1$ and $SC2$ should be consistent. This is verified with the simulations, although the performance of these is slightly worse than ER when the factors have same strengths. However, when the factors have various strengths, $SC1$ and $SC2$ perform much better than BIC_3 and ER.

Table 3.1: Strong and weak factors ($m = 2$, $r = 5, \gamma = 1/3$), $kmax = 8$, ($\alpha = \beta = 0$). The number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations.

N	T	no serial correlations in f_t and u_t							
		SC1		SC2		BIC_3		ER	
100	20	5.016	0.126	5.004	0.063	4.318	0.591	4.994	0.134
100	40	5.014	0.118	5.000	0.000	3.888	0.603	5.000	0.000
100	60	5.006	0.077	5.000	0.000	3.610	0.585	5.000	0.000
200	60	5.004	0.063	5.000	0.000	3.128	0.587	5.000	0.000
500	60	5.000	0.000	5.000	0.000	2.614	0.556	5.000	0.000
100	100	5.008	0.089	5.000	0.000	3.228	0.584	5.000	0.000
200	100	5.002	0.045	5.000	0.000	2.682	0.556	5.000	0.000
500	100	5.000	0.000	5.000	0.000	2.182	0.391	5.000	0.000
10	100	3.446	2.076	3.368	2.005	4.888	0.316	3.872	1.411
20	100	5.434	0.674	5.304	0.623	4.366	0.526	4.936	0.410
40	100	5.300	0.532	5.040	0.196	3.880	0.564	5.000	0.000
60	100	5.240	0.459	5.026	0.159	3.654	0.596	5.000	0.000
60	200	5.242	0.451	5.006	0.077	3.248	0.586	5.000	0.000
60	500	5.310	0.524	5.018	0.133	2.794	0.559	5.000	0.000

Table 3.2: Strong and weak factors ($m = 2$, $r = 5, \gamma = 1/3$), $kmax = 8$, ($\alpha = \beta = 0.5$). The number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations.

N	T	serial correlations in f_t and u_t ($\alpha = \beta = 0.5$)							
		SC1		SC2		BIC_3		ER	
100	20	5.144	0.363	5.056	0.230	4.562	0.539	4.964	0.327
100	40	5.136	0.387	5.034	0.181	4.100	0.599	5.000	0.000
100	60	5.090	0.313	5.020	0.140	3.768	0.575	5.000	0.000
200	60	5.084	0.292	5.032	0.187	3.434	0.575	5.000	0.000
500	60	5.006	0.077	5.032	0.176	3.028	0.580	5.000	0.000
100	100	5.070	0.255	5.008	0.089	3.436	0.592	5.000	0.000
200	100	5.034	0.181	5.010	0.100	3.016	0.604	5.000	0.000
500	100	5.004	0.063	5.004	0.063	2.534	0.523	5.000	0.000
10	100	3.384	2.122	3.364	2.049	4.914	0.288	3.912	1.427
20	100	5.470	0.694	5.310	0.631	4.382	0.570	4.928	0.446
40	100	5.462	0.661	5.098	0.298	4.024	0.607	5.000	0.000
60	100	5.466	0.665	5.134	0.347	3.752	0.619	5.000	0.000
60	200	5.380	0.580	5.054	0.226	3.348	0.579	5.000	0.000
60	500	5.342	0.549	5.040	0.196	2.814	0.562	5.000	0.000

Table 3.3: Strong and weak factors ($m = 2$, $r = 5, \gamma = 1/5$), $kmax = 8$, ($\alpha = \beta = 0$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations.

N	T	no serial correlations in f_t and u_t							
		SC1		SC2		BIC_3		ER	
100	20	5.002	0.045	5.002	0.100	2.570	0.531	3.266	1.484
100	40	5.012	0.109	5.000	0.000	2.006	0.077	4.418	1.187
100	60	5.004	0.063	5.000	0.000	2.000	0.000	4.664	0.947
200	60	5.004	0.063	5.000	0.000	2.000	0.000	4.982	0.232
500	60	5.000	0.000	5.000	0.000	2.000	0.000	5.000	0.000
100	100	5.006	0.077	5.000	0.000	2.000	0.000	4.952	0.377
200	100	5.004	0.063	5.000	0.000	2.000	0.000	5.000	0.000
500	100	5.000	0.000	5.000	0.000	2.000	0.000	5.000	0.000
10	100	3.784	1.405	3.588	1.301	4.426	0.567	2.106	0.769
20	100	5.388	0.628	5.232	0.554	2.772	0.611	2.722	1.302
40	100	5.344	0.571	5.042	0.201	2.056	0.230	3.980	1.423
60	100	5.228	0.457	5.020	0.140	2.002	0.045	4.556	1.066
60	200	5.288	0.519	5.008	0.089	2.000	0.000	4.820	0.713
60	500	5.226	0.481	5.022	0.147	2.000	0.000	4.952	0.377

Table 3.4: Strong and weak factors ($m = 2$, $r = 5, \gamma = 1/5$), $kmax = 8$, ($\alpha = \beta = 0.5$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations.

N	T	serial correlations in f_t and u_t ($\alpha = \beta = 0.5$)							
		SC1		SC2		BIC_3		ER	
100	20	5.118	0.347	5.022	0.160	3.108	0.604	2.728	1.300
100	40	5.136	0.349	5.022	0.147	2.192	0.399	3.518	1.501
100	60	5.118	0.323	5.024	0.153	2.018	0.133	4.058	1.394
200	60	5.104	0.325	5.040	0.196	2.000	0.000	4.544	1.078
500	60	5.008	0.089	5.024	0.166	2.000	0.000	4.820	0.713
100	100	5.058	0.242	5.002	0.045	2.000	0.000	4.736	0.851
200	100	5.052	0.231	5.006	0.077	2.000	0.000	4.958	0.353
500	100	5.002	0.045	5.004	0.063	2.000	0.000	5.000	0.000
10	100	3.682	1.481	3.602	1.285	4.438	0.565	2.220	0.882
20	100	5.530	0.742	5.288	0.585	2.906	0.602	2.640	1.245
40	100	5.402	0.655	5.108	0.311	2.118	0.329	3.482	1.501
60	100	5.386	0.618	5.124	0.330	2.012	0.109	4.154	1.351
60	200	5.328	0.530	5.058	0.234	2.000	0.000	4.688	0.917
60	500	5.372	0.588	5.028	0.165	2.000	0.000	4.922	0.478

Table 3.5: Strong and weak factors ($m = 2$, $r = 5, \gamma = 1/10$), $kmax = 8$, ($\alpha = \beta = 0$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations.

N	T	no serial correlations in f_t and u_t							
		SC1		SC2		BIC_3		ER	
100	20	4.812	0.507	4.324	0.888	2.000	0.000	2.000	0.000
100	40	5.012	0.109	5.000	0.000	2.000	0.000	2.000	0.000
100	60	5.006	0.077	5.000	0.000	2.000	0.000	2.000	0.000
200	60	5.002	0.045	5.000	0.000	2.000	0.000	2.000	0.000
500	60	5.000	0.000	5.000	0.000	2.000	0.000	2.000	0.000
100	100	5.004	0.063	5.000	0.000	2.000	0.000	2.000	0.000
200	100	5.000	0.000	5.000	0.000	2.000	0.000	2.000	0.000
500	100	5.000	0.000	5.000	0.000	2.000	0.000	2.000	0.000
10	100	2.964	0.934	2.596	0.784	3.196	0.662	1.978	0.147
20	100	5.194	0.823	4.352	1.013	2.000	0.000	2.000	0.000
40	100	5.354	0.577	4.994	0.233	2.000	0.000	2.000	0.000
60	100	5.250	0.447	5.008	0.089	2.000	0.000	2.000	0.000
60	200	5.236	0.470	5.012	0.109	2.000	0.000	2.000	0.000
60	500	5.288	0.515	5.008	0.089	2.000	0.000	2.000	0.000

Table 3.6: Strong and weak factors ($m = 2$, $r = 5, \gamma = 1/10$), $kmax = 8$, ($\alpha = \beta = 0.5$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations.

N	T	serial correlations in f_t and u_t ($\alpha = \beta = 0.5$)							
		SC1		SC2		BIC_3		ER	
100	20	4.838	0.699	4.484	0.839	2.008	0.089	2.000	0.000
100	40	5.130	0.365	5.004	0.228	2.000	0.000	2.000	0.000
100	60	5.122	0.351	5.018	0.133	2.000	0.000	2.000	0.000
200	60	5.082	0.296	5.038	0.191	2.000	0.000	2.000	0.000
500	60	5.004	0.063	5.026	0.159	2.000	0.000	2.000	0.000
100	100	5.078	0.276	5.008	0.089	2.000	0.000	2.000	0.000
200	100	5.032	0.176	5.010	0.100	2.000	0.000	2.000	0.000
500	100	5.004	0.063	5.006	0.077	2.000	0.000	2.000	0.000
10	100	2.964	0.892	2.698	0.803	3.260	0.676	1.968	0.198
20	100	5.228	0.870	4.460	1.040	2.000	0.000	1.998	0.045
40	100	5.464	0.691	5.122	0.379	2.000	0.000	2.000	0.000
60	100	5.464	0.637	5.130	0.343	2.000	0.000	2.000	0.000
60	200	5.376	0.579	5.064	0.245	2.000	0.000	2.000	0.000
60	500	5.322	0.524	5.038	0.191	2.000	0.000	2.000	0.000

3.3.2.2 Regional factors

In Tables 3.7 and 3.8 are the performances of all the criteria in the case of regional 3-factor. We also include the results obtained when we forcing ER and BIC_3 identify at least 1 factor. As it can be seen, the sparsity criterion usually produces better results in this case, due to the fact that all the factors are not too strong. Even when we ensure that ER and BIC_3 select at least one factor, they still under-perform comparing to SC1 and SC2. This shows great support for SC1 and SC2, as regional factors can be very common in practice.

Table 3.7: Regional factors, $r = 3$, no serial correlations in f_t and u_t, $kmax = 8$, the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations. We also include the case where we remove the zero factor case for the ER and BIC_3.

N	T	ER (no zero)		BIC_3 (no zero)		SC1	
200	100	2.32	0.91	1.00	0.00	2.88	0.33
200	150	2.61	0.78	1.00	0.00	2.98	0.13
200	200	2.74	0.65	1.00	0.00	3.00	0.06
200	250	2.82	0.56	1.00	0.00	3.00	0.04
200	300	2.85	0.51	1.00	0.00	3.00	0.04
200	350	2.90	0.42	1.00	0.00	3.00	0.00
200	400	2.91	0.42	1.00	0.00	3.00	0.00
200	450	2.91	0.41	1.00	0.00	3.00	0.06
200	500	2.96	0.27	1.00	0.00	3.00	0.00

N	T	ER		BIC_3		SC2	
200	100	0.00	0.00	0.00	0.00	2.61	0.55
200	150	0.00	0.00	0.00	0.00	2.82	0.39
200	200	0.00	0.04	0.00	0.00	2.92	0.27
200	250	0.16	0.68	0.00	0.00	2.95	0.23
200	300	0.65	1.23	0.00	0.00	2.96	0.19
200	350	1.59	1.49	0.00	0.00	2.97	0.17
200	400	2.16	1.34	0.00	0.00	2.99	0.09
200	450	2.55	1.05	0.00	0.00	2.99	0.11
200	500	2.78	0.77	0.00	0.00	3.00	0.04

Table 3.8: Regional factors, $r = 3$, with serial correlations in f_t and u_t ($\alpha = \beta = 0.5$) $kmax = 8$, the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations. We also include the case where we remove the zero factor case for the ER and BIC_3.

N	T	ER (no zero)		BIC_3 (no zero)		SC1	
200	100	1.72	0.88	1.00	0.00	2.81	0.49
200	150	2.09	0.94	1.00	0.00	2.98	0.27
200	200	2.29	0.92	1.00	0.00	3.02	0.14
200	250	2.42	0.89	1.00	0.00	3.01	0.09
200	300	2.63	0.76	1.00	0.00	3.01	0.09
200	350	2.73	0.66	1.00	0.00	3.02	0.14
200	400	2.73	0.67	1.00	0.00	3.01	0.12
200	450	2.77	0.63	1.00	0.00	3.00	0.06
200	500	2.85	0.50	1.00	0.00	3.01	0.08

N	T	ER		BIC_3		SC2	
200	100	0.00	0.00	0.00	0.00	2.50	0.64
200	150	0.00	0.04	0.00	0.00	2.80	0.43
200	200	0.02	0.20	0.00	0.00	2.93	0.25
200	250	0.04	0.33	0.00	0.00	2.97	0.18
200	300	0.42	1.02	0.00	0.00	2.98	0.16
200	350	1.12	1.43	0.00	0.00	2.99	0.11
200	400	1.91	1.40	0.00	0.00	2.98	0.13
200	450	2.38	1.14	0.00	0.00	2.99	0.10
200	500	2.70	0.83	0.00	0.00	3.00	0.00

3.4 Remarks

In overall, we can see that the proposed criteria for selecting the number of factors SC1 and SC2 work well in most cases. They show more advantage over other competing criteria in the case where both strong and weak factors exist in the model. Notice that these simulations are based on conservative choice for the thresholding constant as discussed in Section 3.2. Therefore, we can even improve the selection with a data-driven method for choosing the constant. However, this cross-validation process may require significant computational time and therefore will be drawback for moderate system.

Another remark worth mentioning this chapter is the fact that the estimated idiosyncratic covariance matrix from the factor model will also be sparse even when we extract more than r factors. This is a foundation for our criterion to work. Interestingly, the reason for this is that when we extract more than r factors, the sample covariance $\tilde{\sigma}_{ij}^k$ still converge to the true σ_{ij}. It turns out as this result leads to some extension theory in the large covariance matrix estimation literature. More on this will be discussed in Chapter 4.

3.5 Proofs of results

3.5.1 Proofs of Lemma 3.1

We consider the case when $k = r$ first. In this case we will drop the superscript k in Σ_u^k, so we need to prove that $m(\Sigma_u)$ is bounded, which is immediately stated in assumption 2(iii).

Now for the case when $k < r$,

$$\tilde{u}_t^k = Y_t - \Lambda^{(1:k)} f_t^{(1:k)} = \Lambda^{(k+1)} f_t^{(k+1)} + ... + \Lambda^{(r)} f_t^{(r)} + u_t$$

Hence, if we consider the case when $\Sigma_f = I_r$ for convenient notation (otherwise the result is still valid)

$$\begin{aligned}\Sigma_u^k = \text{cov}(Y_t - \Lambda^{(1:k)} f_t^{(1:k)}) &= \text{cov}(\Lambda^{(k+1)} f_t^{(k+1)} + ... + \Lambda^{(r)} f_t^{(r)}) + \Sigma_u \\ &= \Lambda^{(k+1:r)} \Lambda^{(k+1:r)'} + \Sigma_u\end{aligned}$$

It is clear to see that with the pervasive condition of factors, the eigenvalues of $\Lambda^{(k+1:r)} \Lambda^{(k+1:r)'}$ diverge at least at rate $d_{k+1}(N)$ whereas the eigenvalues of Σ_u are bounded. Hence, $\left\|\Sigma_u^k\right\|$ diverges at least at rate $d_{k+1}(N)$ because

$$\|\Sigma_u\| = \left\|\Lambda^{(k+1:r)} \Lambda^{(k+1:r)'} - \Sigma_u^k\right\| \geq \left\|\Lambda^{(k+1:r)} \Lambda^{(k+1:r)'}\right\| - \left\|\Sigma_u^k\right\|$$

Therefore, the maximum row sum of Σ_u^k must diverge as well. Since all the entries of Σ_u^k are finite, diverging maximum row sum implies that the number of non-zero entries must be unbounded, since clearly all the entries of Σ_u^k are finite.

3.5.2 Proofs of Theorem 3.1

The proof of this theorem will make use of Lemmas 3.8 and 3.9. We will divide the proof of this theorem to 2 cases: when $k \geq r$ and $k < r$.

When $k \geq r$

Define $m_i(\Sigma_u) = \sum_{j=1}^{N} \mathbf{I}\{\sigma_{ij} \neq 0\}$ and $\tilde{m}_i(\Sigma_u^k) = \sum_{j=1}^{N} \mathbf{I}\left\{\left|\tilde{\sigma}_{ij}^k\right| > h_{ij}^k\right\}$, then $m(\Sigma_u) = \max_i m_i(\Sigma_u)$ and $\tilde{m}(\Sigma_u^k) = \max_i \tilde{m}_i(\Sigma_u^k)$. Now:

$$\begin{aligned} m(\Sigma_u) - \tilde{m}(\Sigma_u^k) &= \max_i m_i(\Sigma_u) - \max_i \tilde{m}_i(\Sigma_u^k) \\ &\leq \max_i \left| m_i(\Sigma_u) - \tilde{m}_i(\Sigma_u^k) \right| \end{aligned}$$

Since $m(\Sigma_u)$ is bounded, we only need to show that $\max_i \left| m_i(\Sigma_u) - \tilde{m}_i(\Sigma_u^k) \right| = o_p(1)$. To find the bound for $\max_i \left| m_i(\Sigma_u) - \tilde{m}_i(\Sigma_u^k) \right|$, use Markov's Inequality, i.e. we have that:

$$\forall i,\ \mathbf{P}\left\{\left| m_i(\Sigma_u^k) - \tilde{m}_i(\Sigma_u^k) \right| > \epsilon\right\} < \frac{\mathrm{E}\left\{\left| m_i(\Sigma_u^k) - \tilde{m}_i(\Sigma_u^k) \right|\right\}}{\epsilon}$$

Also,

$$\begin{aligned}
\mathrm{E}\left\{\left|m_i(\Sigma_u^k)-\tilde{m}_i(\Sigma_u^k)\right|\right\} &= \mathrm{E}\left\{\left|\sum_{j=1}^{N}\mathbf{I}\left\{\sigma_{ij}\neq 0\right\}-\sum_{j=1}^{N}\mathbf{I}\left\{\left|\tilde{\sigma}_{ij}^k\right|>h_{ij}^k\right\}\right|\right\} \\
&= \mathrm{E}\left\{\sum_{j=1}^{N}\mathbf{I}\left\{\left|\tilde{\sigma}_{ij}^k\right|>h_{ij}^k,\sigma_{ij}=0\right\}+\sum_{j=1}^{N}\mathbf{I}\left\{\left|\tilde{\sigma}_{ij}^k\right|\leq h_{ij}^k,\sigma_{ij}\neq 0\right\}\right\} \\
&= \sum_{j=1}^{N}\mathbf{P}\left\{\left|\tilde{\sigma}_{ij}^k\right|>h_{ij}^k,\sigma_{ij}=0\right\}+\sum_{j=1}^{N}\mathbf{P}\left\{\left|\tilde{\sigma}_{ij}^k\right|\leq h_{ij}^k,\sigma_{ij}\neq 0\right\} \\
&\leq \sum_{j=1}^{N}\mathbf{P}\left\{\left|\tilde{\sigma}_{ij}^k\right|>h_{ij}^k|\sigma_{ij}=0\right\}+\sum_{j=1}^{N}\mathbf{P}\left\{\left|\tilde{\sigma}_{ij}^k\right|\leq h_{ij}^k|\sigma_{ij}\neq 0\right\} \\
&\leq \sum_{j=1}^{N}O\left(\frac{1}{N^2}+\frac{1}{T^2}\right)=O\left(\frac{1}{N}+\frac{N}{T^2}\right)
\end{aligned}$$

The last step above is from lemma 3.8. Therefore, choosing ϵ such that $\frac{1}{N}+\frac{N}{T^2}=o(\epsilon)$, then clearly $\forall i,\ \mathbf{P}\left\{|m_i-\hat{m}_i|>\epsilon\right\}\to 0$, Notice that if $N,T\to\infty$ and $N=o(T^2)$, then $\epsilon\to 0$ and $\max_i\left|m_i(\Sigma_u)-\tilde{m}_i(\Sigma_u^k)\right|\to 0$, which proves Theorem 3.1 for the case $k\geq r$.

When $k<r$

By lemma 3.9 , we have the same result required to show that $\left|m_i(\Sigma_u^k)-\tilde{m}_i(\Sigma_u^k)\right|\to 0$. Since $m_i(\Sigma_u^k)$ diverges at rate at least $d_{k+1}(N)$ as shown in lemma 3.1, we can establish the claim.

3.5.3 Proofs of Corollary 3.1

We need to prove that in both case where $k>r$ and $k<r$, then

$$\mathbf{P}\left\{\tilde{m}(\Sigma_u^k)+C\,k\,g(N)>\tilde{m}(\Sigma_u^r)+C\,r\,g(N)\right\}\to 1$$

or

$$\mathbf{P}\left\{\tilde{m}(\Sigma_u^k)-\tilde{m}(\Sigma_u^r)+C\ (k-r)\ g(N)>0\right\}\to 1$$

Using theorem 3.1, when $k < r$, $\tilde{m}(\Sigma_u^k)$ grows to infinity at rate $d_k(N)$, $\tilde{m}(\Sigma_u^r)$ is $O_p(1)$, $g(N)$ grows at a slower rate then $d_k(N)$, therefore the dominating term is $\tilde{m}(\Sigma_u^k)$, which is positive.

On the other hand, when $k > r$, $\tilde{m}(\Sigma_u^k)$ and $\tilde{m}(\Sigma_u^r)$ are both $O_p(1)$ so the dominated term is $C\ (k-r)\ g(N)$, which are also positive.

Hence, $\mathbf{P}\left\{\tilde{m}(\Sigma_u^k) - \tilde{m}(\Sigma_u^r) + C\ (k-r)\ g(N) > 0\right\} \to 1$ for $k \neq r$.

3.5.4 Technical Lemmas

Lemma 3.2. *For $k \geq r$, let $H^k = (D_N^k)^{-1}\tilde{F}^{k\prime}F\Lambda'\Lambda/T$ which is a $k \times r$ matrix, where*

$$D_N^k = \begin{bmatrix} D_N & 0 \\ 0 & NI_{k-r} \end{bmatrix}.$$

Furthermore, for $k \geq r$, $\tilde{V}^k = diag(\tilde{v}_1, ..., \tilde{v}_k)$ and $\hat{F}^k = \tilde{F}^k\tilde{V}^k(D_N^k)^{-1} = \frac{1}{T}YY'\tilde{F}^k(D_N^k)^{-1}$. Under assumption 1-5, there exists a constant C such that:

$$\mathbf{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{f}_t^k - H^k f_t\right\|^2 > C\left[\frac{N}{[d_r(N)]^2} + \frac{N^2}{T\,[d_r(N)]^2}\right]\right) \leq O\left(\frac{1}{T^2}\right).$$

Proof. Similarly to the proof of theorem 2.1, we consider the operator norm of the matrix $\hat{F}^k - FH^{k\prime}$ and a similar decomposition:

$$\begin{aligned}
\hat{F}^k - FH^{k\prime} &= \frac{1}{T}YY'\tilde{F}^k(D_N^k)^{-1} - \frac{1}{T}F\Lambda'\Lambda F'\tilde{F}^k(D_N^k)^{-1} \\
&= \left(\frac{F\Lambda'U'}{T} + \frac{U\Lambda F'}{T} + \frac{UU'}{T}\right)\tilde{F}^k(D_N^k)^{-1} \\
&= \begin{bmatrix} \left(\frac{F\Lambda'U'}{T} + \frac{U\Lambda F'}{T} + \frac{UU'}{T}\right)\tilde{F}^r(D_N)^{-1} & 0 \\ 0 & \left(\frac{F\Lambda'U'}{NT} + \frac{U\Lambda F'}{NT} + \frac{UU'}{NT}\right)\tilde{F}^{(k+1:r)} \end{bmatrix}
\end{aligned}$$

Therefore, for (i) we have:

$$\frac{1}{T}\left\|\hat{F}^k - FH^{k'}\right\|^2 \leq \max\left(\underbrace{\frac{1}{T}\left\|\left(\frac{F\Lambda'U'}{T}+\frac{U\Lambda F'}{T}+\frac{UU'}{T}\right)\tilde{F}^r(D_N)^{-1}\right\|^2}_{A}, \underbrace{\frac{1}{T}\left\|\left(\frac{F\Lambda'U'}{NT}+\frac{U\Lambda F'}{NT}+\frac{UU'}{NT}\right)\tilde{F}^{(k+1:r)}\right\|^2}_{B}\right).$$

We work out the result for each term A and B separately:

$$\begin{aligned}A &\leq \frac{1}{T}\left\|\left(\frac{1}{T}F\Lambda'U'\tilde{F}^r(D_N)^{-1}\right)\right\|^2 + \frac{1}{T}\left\|\frac{1}{T}U\Lambda F'\tilde{F}^r(D_N)^{-1}\right\|^2 + \frac{1}{T}\left\|\frac{1}{T}UU'\tilde{F}^r(D_N)^{-1}\right\|^2 \\ &\leq 2\left\|D_N^{-1}\right\|\frac{1}{T}\left\|(D_N)^{-1/2}\Lambda'U'\right\|^2\left\|\frac{1}{T}\tilde{F}^{r'}\tilde{F}^r\right\|\left\|\frac{1}{\sqrt{T}}F\right\|^2 \\ &+ \left\|\frac{1}{d_r(N)T}UU'\right\|^2\left\|\frac{1}{T}\tilde{F}^{r'}\tilde{F}^r\right\|.\end{aligned}$$

Since $\left\|\frac{1}{T}\tilde{F}^{r'}\tilde{F}^r\right\|$, $\frac{1}{T}\left\|(D_N)^{-1/2}\Lambda'U'\right\|^2$ are all $O_p(1)$, we need to show that:

$\mathbf{P}\left(\left\|\frac{1}{\sqrt{T}}F\right\|^2 > C\right) = O\left(\frac{1}{T^2}\right)$: This follows by Lemma B.1 (i) in Fan et al. (2011) under the common assumptions with this thesis. See also the proof of lemma 3.1 (ii) in Fan et al. (2011).

$\mathbf{P}\left(\left\|\frac{1}{d_r(N)T}UU'\right\|^2 > C\left[\frac{N}{[d_r(N)]^2}+\frac{N^2}{T[d_r(N)]^2}\right]\right) = O\left(\frac{1}{T^2}\right)$: This is done by the same decomposition as in Lemma 2.3, where we have:

$$\begin{aligned}\left\|\frac{1}{d_r(N)T}UU'\right\|^2 &\leq \frac{2N}{[d_r(N)]^2}\frac{1}{N}\left(\max_{s,t}\left[u_s't_t - E(u_s't_t)\right]\right)^2 \\ &+ \frac{2N^2}{T\left[d_r(N)\right]^2}\left(\max_s\frac{1}{N}\sum_{t=1}^{T}E(u_s't_t)\right)^2\end{aligned}$$

and since $\frac{1}{N}\left(\max_{s,t}\left[u_s't_t - E(u_s't_t)\right]\right)^2 = O_p(1)$ by assumption 4 (ii), there exist a constant C such that $\mathbf{P}\left(\frac{1}{N}\left(\max_{s,t}\left[u_s't_t - E(u_s't_t)\right]\right)^2 > C\right)$ is arbitrarily small. For similar reason, since $\max_s\frac{1}{N}\sum_{t=1}^{T}E(u_s't_t) = O(1)$, there exists a constant C such

that $\mathbf{P}\left(\left(\max_s \frac{1}{N}\sum_{t=1}^{T} E(u_s' t_t)\right)^2 > C\right)$ is arbitrarily small. Hence,

$$\mathbf{P}\left(A > C\left[\frac{N}{[d_r(N)]^2} + \frac{N^2}{T\,[d_r(N)]^2}\right]\right) = O\left(\frac{1}{T^2}\right).$$

Notice that we omit the term $\left\|D_N^{-1}\right\| = \frac{1}{d_r(N)}$ that comes from the first part of A because it is dominated by $\frac{N}{[d_r(N)]^2}$. For B,

$$\begin{aligned} B &\leq \frac{1}{T}\left\|\left(\frac{1}{NT}F\Lambda' U'\tilde{F}^{(k+1:r)}\right)\right\|^2 + \frac{1}{T}\left\|\frac{1}{T}U\Lambda F'\tilde{F}^{(k+1:r)}\right\|^2 + \frac{1}{T}\left\|\frac{1}{T}UU'\tilde{F}^{(k+1:r)}\right\|^2 \\ &\leq 2\frac{1}{T}\left\|\frac{1}{NT}\Lambda' U'\right\|^2 \left\|\frac{1}{T}\tilde{F}^{(k+1:r)'}\tilde{F}^{(k+1:r)}\right\| \left\|\frac{1}{\sqrt{T}}F\right\|^2 \\ &+ \left\|\frac{1}{NT}UU'\right\|^2 \left\|\frac{1}{T}\tilde{F}^{(k+1:r)'}\tilde{F}^{(k+1:r)}\right\|. \end{aligned}$$

By similar result with term A (replacing $d_r(N)$ with N), we have: there exist a constant C such that,

$$\mathbf{P}\left(B > C\left[\frac{N}{[d_r(N)]^2} + \frac{N^2}{T\,[d_r(N)]^2}\right]\right) \leq \mathbf{P}\left(B > C\left[\frac{1}{N} + \frac{1}{T}\right]\right) = O\left(\frac{1}{T^2}\right).$$

Hence the result follows. □

Remark 3.1. Lemma 3.2 establishes the result when one estimate more than r factors by PCs. It turns out that we can not have both the estimated factors and the loadings consistent. One can prove that although $\hat{F}^k$ is consistent up to a rotation (even for all k), $\hat{\Lambda}^k = Y'\hat{F}^k(\hat{F}^{k\prime}\hat{F}^k)^{-1}$ will not be a consistent estimator when $k > r$. We do not use this result in this paper so we leave it aside. However, interestingly, as shown in lemma 3.3, the product of factors and loadings, i.e. the estimated idiosyncratic errors $\tilde{u}_{it}^k$ is consistent for u_{it}. The rationale behind redefine the matrix H^k and $\hat{F}^k$ is that we want to at least get the factors consistent when $k > r$, which is needed for the proof of lemma 3.2.

Lemma 3.3. *Recall that* $\omega_T = \sqrt{\frac{\log N}{T}} + \frac{\sqrt{N}}{[d_r(N)]} + \frac{N}{\sqrt{T}[d_r(N)]}$. *Under assumption 1-5:*

if $k \geq r$ then: There exists a constant C such that for all $c > C$,

(i) $\mathbf{P}\Big(\max_{i\leq N} \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{u}_{it}^{k} - u_{it}\right)^2 > c\omega_T^2\Big) = O(\frac{1}{N^2} + \frac{1}{T^2})$

(ii) $\mathbf{P}\left(\max_{i,j}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{u}_{it}^{k}\tilde{u}_{jt}^{k} - u_{it}u_{jt}\right)\right| > c\omega_T\right) = O\left(\frac{1}{N^2} + \frac{1}{T^2}\right).$

Proof. First of all, for $k > r$, we let: $\hat{\Lambda}^k = Y'\hat{F}^k(\hat{F}^{k\prime}\hat{F}^k)^{-1}$. An important identity that we exploit is $\hat{\Lambda}^k\hat{F}^k = \tilde{\Lambda}^k\tilde{F}^k$. Furthermore, we define G^k such that $G^kH^k = I_r$. Now to prove (i), fix $i \leq N$ and consider the expanding of $\tilde{u}_{it}^k - u_{it}$:

$$
\begin{aligned}
&\tilde{u}_{it}^k - u_{it} \\
&= \lambda_i' f_t - \tilde{\lambda}_i^{k\prime}\tilde{f}_t^k = \lambda_i' f_t - \hat{\lambda}_i^{k\prime}\hat{f}_t^k \\
&= \lambda_i' G^k H^k f_t - \lambda_i' G^k \hat{f}_t^k + \lambda_i' G^k \hat{f}_t^k - \hat{\lambda}_i^{k\prime}\hat{f}_t^k \\
&= \lambda_i' G^k\left(H^k f_t - \hat{f}_t^k\right) + \left(\lambda_i' G^k - \hat{\lambda}_i^{k\prime}\right)\hat{f}_t^k \\
&= \lambda_i' G^k\left(H^k f_t - \hat{f}_t^k\right) + \left(\lambda_i' G^k - (\lambda_i' F' + u_i')\hat{F}^k(\hat{F}^{k\prime}\hat{F}^k)^{-1}\right)\hat{f}_t^k \\
&= \lambda_i' G^k\left(H^k f_t - \hat{f}_t^k\right) + \lambda_i' G^k\hat{F}^{k\prime}\hat{F}^k(\hat{F}^{k\prime}\hat{F}^k)^{-1}\hat{f}_t^k \\
&+ \lambda_i' G^k H^k F'\hat{F}^k(\hat{F}^{k\prime}\hat{F}^k)^{-1}\hat{f}_t^k + \sum_{t=1}^{T}\hat{f}_t^{k\prime}\left(\hat{F}^{k\prime}\hat{F}^k\right)^{-1}\hat{f}_t^k u_{it} \\
&= \lambda_i' G^k\left(H^k f_t - \hat{f}_t^k\right) + \lambda_i' G^k\left(\hat{F}^{k\prime} - H^k F'\right)\hat{F}^k\left(\hat{F}^{k\prime}\hat{F}^k\right)^{-1}\hat{f}_t^k + \sum_{t=1}^{T}\hat{f}_t^{k\prime}\left(\hat{F}^{k\prime}\hat{F}^k\right)^{-1}\hat{f}_t^k u_{it}
\end{aligned}
$$

For lemma 3.3 (i):

$$\begin{aligned}\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{u}_{it}^{k}-u_{it}\right)^{2} &\leq \max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}\left|\lambda_{i}^{\prime}G^{k}\left(H^{k}f_{t}-\hat{f}_{t}^{k}\right)\right|^{2}\\ &+\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}T\left|\lambda_{i}^{\prime}G^{k}\left(\hat{F}^{k^{\prime}}-H^{k}F^{\prime}\right)\hat{F}^{k}\left(\hat{F}^{k^{\prime}}\hat{F}^{k}\right)^{-1}\hat{f}_{t}^{k}\right|^{2}\\ &+\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}\left|\sum_{t=1}^{T}\hat{f}_{t}^{k^{\prime}}\left(\hat{F}^{k^{\prime}}\hat{F}^{k}\right)^{-1}\hat{f}_{t}^{k}u_{it}\right|^{2}\\ &\leq \frac{1}{T}\sum_{t=1}^{T}\left\|H^{k}f_{t}-\hat{f}_{t}^{k}\right\|^{2}\max_{i\leq N}\left\|\lambda_{i}^{\prime}G^{k}\right\|^{2}\\ &+\frac{1}{T}\left\|H^{k}F^{\prime}-\hat{F}^{k^{\prime}}\right\|^{2}\max_{i\leq N}\left\|\lambda_{i}^{\prime}G^{k}\right\|^{2}\left\|\hat{F}^{k}\left(\hat{F}^{k^{\prime}}\hat{F}^{k}\right)^{-1}\hat{F}^{k^{\prime}}\right\|^{2}\\ &+\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}|u_{it}|^{2}\left\|\hat{F}^{k}\left(\hat{F}^{k^{\prime}}\hat{F}^{k}\right)^{-1}\hat{F}^{k^{\prime}}\right\|^{2}\end{aligned}$$

We use $\mathbf{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{f}_{t}^{k}-H^{k}f_{t}\right\|^{2}>C\left[\frac{N}{[d_{r}(N)]^{2}}+\frac{N^{2}}{T[d_{r}(N)]^{2}}\right]\right)=O\left(\frac{1}{T^{2}}\right)$, $\left\|\hat{F}^{k}\left(\hat{F}^{k^{\prime}}\hat{F}^{k}\right)^{-1}\hat{F}^{k^{\prime}}\right\|=O_{p}(1)$, $\left\|\lambda_{i}^{\prime}G^{k}\right\|=O_{p}(1)$ and $\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}|u_{it}|=\sqrt{\frac{\log N}{T}}$ as shown below:

By the Bernstein inequality for weakly dependent process (see Merlevède et al. (2011)), for some positive constants C_1, C_2, C_3, C_4 and C_5 and for any $i\in(1,...,N)$, since $\mathrm{E}u_{it}=0$,

$$\begin{aligned}\left(\frac{1}{T}\sum_{t=1}^{T}|u_{it}|>s\right)&\leq T\exp\left(-\frac{(Ts)^{\gamma}}{C_{1}}\right)+\exp\left(-\frac{(Ts)^{2}}{C_{2}(1+TC_{3})}\right)\\ &+\exp\left(-\frac{(Ts)^{2}}{C_{4}T}\exp\left(\frac{(Ts)^{\gamma(1-\gamma)}}{C_{5}(\log Ts)^{\gamma}}\right)\right).\end{aligned}$$

Using Bonferroni's method and choosing $s=\sqrt{\frac{\log N}{T}}$ yields:

$$\mathbf{P}\left(\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}|u_{it}|^{2}>\frac{\log N}{T}\right)=O\left(\frac{1}{N^{2}}\right)$$

Therefore, the desired result is obtained by combining all the results above (note

that $\omega_T^2 = O(\frac{\log N}{T} + \frac{N}{[d_r(N)]^2} + \frac{N^2}{T[d_r(N)]^2})$.

Part (ii) follows from part (i), by the same argument as in the proof of lemma A.3 in Fan et al. (2011). □

Remark 3.2. Lemma 3.3 (i) is a similar result from Fan et al. (2013). The reason that we show it again here is because we emphasise that it is true even for the case when the factors are weak and even for $k > r$, not only $k = r$ as in Fan et al. (2013). This leads to the immediate consistency for the POET estimator under weaker factor assumption and any $k \geq r$.

Lemma 3.4. *Suppose that the random variables Z_1, Z_2 both satisfy the exponential-type tail condition: There exists r_1, $r_2 \in (0,1)$ and b_1, $b_2 > 0$, such that $\forall s > 0$,*

$$\mathbf{P}\left(|Z_i| > s\right) \leq exp\left(1 - (s/b_i)^{r_i}\right), \quad i = 1, 2.$$

Then $Z_1 + Z_2$ also satisfy the exponential-type tail condition, i.e. for some $r_3 \in (0,1)$ and $b_3 > 0$, $\forall s > 0$ we have:

$$\mathbf{P}\left(|Z_1 + Z_2| > s\right) \leq exp\left(1 - (s/b_3)^{r_3}\right)$$

Proof. Let $b = 2\max(b_1, b_2)$ and $r = \min(r_1, r_2)$, then $\forall s > 0$ we have:

$$\begin{aligned}
\mathbf{P}\left(|Z_1 + Z_2| > s\right) &\leq \mathbf{P}\left(|Z_1| + |Z_2| > s\right) \\
&\leq \mathbf{P}\left(|Z_1| > s/2\right) + \mathbf{P}\left(|Z_2| > s/2\right) \\
&\leq \exp\left(1 - (s/(2b_1))^{r_1}\right) + \exp\left(1 - (s/(2b_2))^{r_2}\right) \\
&\leq 2\exp\left(1 - (s/(b))^{r}\right).
\end{aligned}$$

The rest of the proof is similar to the proof of Lemma A.2 in Fan et al. (2011) and hence omitted. □

Lemma 3.5. *Under assumption 1-5, there exists a constant C such that for all $c > C$, $k < r$:*

$$\mathbf{P}\left(\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{u}_{it}^{k}-u_{it}^{k}\right)^{2}>c\omega_{T}^{2}\right)=O(\frac{1}{N^{2}}+\frac{1}{T^{2}}) \tag{3.10}$$

Proof. Let $H^k = (D_N^k)^{-1}\tilde{F}^{k\prime}F^k\Lambda^{k\prime}\Lambda^k/T$ which is a $k \times k$ matrix, where

$$D_N^k = \begin{bmatrix} d_1(N) & & 0 \\ & \ddots & \\ 0 & & d_k(N) \end{bmatrix}.$$

We first need to prove that: for some constant C:

$$\mathbf{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{f}_{t}^{k}-H^{k}f_{t}^{k}\right\|^{2}>C\left[\frac{N}{\left[d_{r}(N)\right]^{2}}+\frac{N^{2}}{T\left[d_{r}(N)\right]^{2}}\right]\right)=O\left(\frac{1}{T^{2}}\right). \tag{3.11}$$

Now since, $Y = F\Lambda' + U = F^k\Lambda^{k\prime} + F^{(k+1:r)}\Lambda^{(k+1:r)\prime} + U$

$$YY'-F^{k}\Lambda^{k\prime}\Lambda^{k}F^{k\prime}=\frac{F\Lambda'U'}{T}+\frac{U\Lambda F'}{T}+\frac{UU'}{T}+F^{k}\Lambda^{k\prime}\Lambda^{(k+1:r)}F^{(k+1:r)\prime}+F^{(k+1:r)}\Lambda^{(k+1:r)\prime}\Lambda^{k}F^{k\prime}$$

So,

$$\begin{aligned}\hat{F}^{k}-F^{k}H^{k\prime} &= \frac{1}{T}YY'\tilde{F}^{k}(D_{N}^{k})^{-1}-\frac{1}{T}F^{k}\Lambda^{k\prime}\Lambda^{k}F^{k\prime}\tilde{F}^{k}(D_{N}^{k})^{-1}\\ &= \left(\frac{F\Lambda'U'}{T}+\frac{U\Lambda F'}{T}+\frac{UU'}{T}\right)\tilde{F}^{k}(D_{N}^{k})^{-1}\\ &\quad+\left(\frac{F^{k}\Lambda^{k\prime}\Lambda^{(k+1:r)}F^{(k+1:r)\prime}}{T}+\frac{F^{(k+1:r)}\Lambda^{(k+1:r)\prime}\Lambda^{k}F^{k\prime}}{T}\right)\tilde{F}^{k}(D_{N}^{k})^{-1}\\ &= I+II\end{aligned}$$

By lemma 3.2, we have already shown that $\mathbf{P}\left(\frac{1}{T}\left\|I\right\|^{2}>C\left[\frac{N}{[d_r(N)]^2}+\frac{N^2}{T[d_r(N)]^2}\right]\right)=$

$O\left(\frac{1}{T^2}\right)$. Now for II,

$$\begin{aligned}
\frac{1}{T}\|II\|^2 &\leq \frac{1}{T}\left\|\frac{F^k\Lambda^{k'}\Lambda^{(k+1:r)}F^{(k+1:r)'}\tilde{F}^k(D_N^k)^{-1}}{T}\right\|^2 \\
&\quad + \frac{1}{T}\left\|\frac{F^{(k+1:r)}\Lambda^{(k+1:r)'}\Lambda^k F^{k'}\tilde{F}^k(D_N^k)^{-1}}{T}\right\|^2 \\
&\leq \frac{2}{T}\left\|\Lambda^{k'}\Lambda^{k'}(D_N^k)^{-1}\right\|\left\|(D_N^k)^{-1}\right\|\left\|\Lambda^{(k+1:r)'}\Lambda^{(k+1:r)}\right\|\left\|\frac{\tilde{F}^{k'}\tilde{F}^k}{T}\right\|\left\|\frac{1}{\sqrt{T}}F^k\right\|\left\|\frac{1}{\sqrt{T}}F^{(k+1:r)}\right\| \\
&\leq \frac{2}{T}\left\|\Lambda^{k'}\Lambda^{k'}(D_N^k)^{-1}\right\|\left\|\frac{\Lambda^{(k+1:r)'}\Lambda^{(k+1:r)}}{d_k(N)}\right\|\left\|\frac{\tilde{F}^{k'}\tilde{F}^k}{T}\right\|\left\|\frac{1}{\sqrt{T}}F^k\right\|\left\|\frac{1}{\sqrt{T}}F^{(k+1:r)}\right\|.
\end{aligned}$$

Using the same results as in lemma 3.2 regarding $\left\|\frac{1}{\sqrt{T}}F^k\right\|$ and $\left\|\frac{1}{\sqrt{T}}F^{(k+1:r)}\right\|$, we have

$$\mathbf{P}\left(\frac{1}{T}\|II\|^2 > C\left[\frac{1}{T}\right]\right) = O\left(\frac{1}{T^2}\right).$$

Combining, the result for I and II, we have proven (3.11). Now, using

$$\begin{aligned}
\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{u}_{it}^k - u_{it}^k\right)^2 &\leq \frac{1}{T}\sum_{t=1}^{T}\left\|H^k f_t^k - \hat{f}_t^k\right\|^2 \max_{i\leq N}\left\|\lambda_i' G^k\right\|^2 \\
&\quad + \frac{1}{T}\left\|H^k F^{k'} - \hat{F}^{k'}\right\|^2 \max_{i\leq N}\left\|\lambda_i' G^k\right\|^2 \left\|\hat{F}^k\left(\hat{F}^{k'}\hat{F}^k\right)^{-1}\hat{F}^{k'}\right\|^2 \\
&\quad + \max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}\left|u_{it}^k\right|^2\left\|\hat{F}^k\left(\hat{F}^{k'}\hat{F}^k\right)^{-1}\hat{F}^{k'}\right\|^2.
\end{aligned}$$

We have already worked our the part for the first 2 terms above in result (3.11). For the bound of $\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}\left|u_{it}^k\right|$, consider:

$$u_{it}^k = \sum_{l=k+1}^{r}\lambda_i^{(l)}f_t^{(l)} + u_{it}.$$

From lemma 3.4, we know that the sum of 2 exponential-type tail condition variables is an exponential-type tail condition variable. Also, it is clear that an exponential-type tail condition variable also preserves its condition under scaling by a constant. Therefore, u_{it}^k satisfies the exponential-type tail condition. Hence, similar to part of

the proof in lemma 3.3, we have that for some constant C:

$$\mathbf{P}\left(\max_{i\leq N}\frac{1}{T}\sum_{t=1}^{T}\left|u_{it}^{k}\right|^{2}>C\frac{\log N}{T}\right)=O\left(\frac{1}{N^{2}}\right).$$

Therefore, combining everything, we finally reach the result (3.10). □

Lemma 3.6. *Under assumption 1-5, there exists a constant C such that for all $c > C$, $k < r$:*

(i) $\mathbf{P}\left(\max_{i,j}\left|\frac{1}{T}\sum_{t=1}^{T}\left(u_{it}^{k}u_{jt}^{k}-\sigma_{ij}^{k}\right)\right|>c\sqrt{\frac{\log N}{T}}\right)=O\left(\frac{1}{N^{2}}\right)$

(ii) $\mathbf{P}\left(\max_{i,j}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{u}_{it}^{k}\tilde{u}_{jt}^{k}-u_{it}^{k}u_{jt}^{k}\right)\right|>c\omega_{T}\right)=O(\frac{1}{N^{2}}+\frac{1}{T^{2}}).$

Proof. To prove part (i), we just need to recall that u_{it}^{k} satisfies the exponential-type tail condition, as shown in part of the proof of lemma 3.5. The rest is exactly the same as in the proof of lemma A.3 (i) in Fan et al. (2001).

Part (ii) follows from lemma 3.5, by the same argument as in the proof of lemma A.3 (ii) in Fan et al. (2001). □

Lemma 3.7. *Under assumption 1-5, we have the following results:*

(i) $\forall(i,j):\ \mathbf{P}\left(\left|\tilde{\sigma}_{ij}^{k}-\sigma_{ij}\right|>h_{ij}^{k}\right)\leq O\left(\frac{1}{N^{2}}+\frac{1}{T^{2}}\right)$ for $k\geq r$.

(ii) $\forall(i,j):\ \mathbf{P}\left(\left|\tilde{\sigma}_{ij}^{k}-\sigma_{ij}^{k}\right|>h_{ij}^{k}\right)\leq O\left(\frac{1}{N^{2}}+\frac{1}{T^{2}}\right)$ for $k<r$.

Proof. Fix $k \geq r$. For (i), we first use the following results in Fan et al. (2011, 2013): Under assumptions 1-4, there exists constant C_1 such that for all $c > C_1$:

$$\mathbf{P}\left(\max_{i,j}\left|\frac{1}{T}\sum_{t=1}^{T}\left(u_{it}u_{jt}-\sigma_{ij}\right)\right|>c\sqrt{\frac{\log N}{T}}\right)=O\left(\frac{1}{N^{2}}\right)\tag{3.12}$$

Result (3.12) intuitively shows a convergence of the average of week dependent data, which in this case is $u_{it}u_{jt}$. Also, by lemma 3.3(iii) for $k \geq r$ there exists a constant C_2 such that for all $c > C_2$

$$\mathbf{P}\left(\max_{i,j}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{u}_{it}^{k}\tilde{u}_{jt}^{k}-u_{it}u_{jt}\right)\right|>c\omega_{T}\right)=O\left(\frac{1}{N^{2}}+\frac{1}{T^{2}}\right).\tag{3.13}$$

Together, (3.12)) and (3.13) yield the following: if $c > \max(C_1, C_2)$:

$$\mathbf{P}\left(\max_{i,j}\left|\tilde{\sigma}_{ij}^k - \sigma_{ij}\right| \leq c\omega_T\right) \geq 1 - O\left(\frac{1}{N^2} + \frac{1}{T^2}\right). \tag{3.14}$$

Furthermore, by our assumption that $\tilde{\theta}_{ij}^k$ is asymptotically bounded between 2 constants,

$$\exists\,(C_L, C_H) \text{ such that } \forall(i,j),\ \mathbf{P}\left(C_L \leq \tilde{\theta}_{ij}^k \leq C_H\right) \geq 1 - O\left(\frac{1}{N^2} + \frac{1}{T^2}\right). \tag{3.15}$$

Hence if $h_{ij}^k = C\omega_T\sqrt{\tilde{\theta}_{ij}^k}$ for some constant C then

$$\mathbf{P}\left(\left|\tilde{\sigma}_{ij}^k - \sigma_{ij}\right| > h_{ij}^k\right) \leq O\left(\frac{1}{N^2} + \frac{1}{T^2}\right)$$

for $k \geq r$ as required.

For the case $k < r$ in (ii), lemma (3.6) establishes the similar results as in (3.12) and (3.13), which lead to the required result. □

Lemma 3.8. *Under assumption 1-5, we have the following results:* for $k \geq r$,

(i) $\forall(i,j):\ \mathbf{P}\left(\left|\tilde{\sigma}_{ij}^k\right| > h_{ij}^k \mid \sigma_{ij} = 0\right) \leq O\left(\frac{1}{N^2} + \frac{1}{T^2}\right)$.

Under assumption 1-6, we have: for $k \geq r$,

(ii) $\forall(i,j):\ \mathbf{P}\left(\left|\tilde{\sigma}_{ij}^k\right| < h_{ij}^k \mid \sigma_{ij} \neq 0\right) \leq O\left(\frac{1}{N^2} + \frac{1}{T^2}\right)$

Proof. Lemma 3.7 (i) directly implies Lemma 3.8 (i), since $k \geq r$ given that $\sigma_{ij} = 0$, event $\left\{\left|\tilde{\sigma}_{ij}^k - \sigma_{ij}\right| > h_{ij}^k\right\}$ is equivalent to $\left|\tilde{\sigma}_{ij}^k\right|$ is greater than h_{ij}, which has asymptotic probability tending to 0 as well. Similarly for the case $k < r$.

Now, to prove (ii) we will first use assumption 6: $\forall(i,j)$,

$$\mathbf{P}\left(\left|\tilde{\sigma}_{ij}^k\right| > h_{ij} \mid \sigma_{ij} \neq 0\right) = \mathbf{P}\left(\left|\tilde{\sigma}_{ij}^k\right| > h_{ij} \mid |\sigma_{ij}| > \tau\right)$$

Let us defining 2 events:

$$\text{E1} = \left\{ \max_{i,j} \left| \tilde{\sigma}_{ij}^k - \sigma_{ij} \right| \leq c\omega_T \right\} \quad \text{for all } c > \max(C_1, C_2),$$

$$\text{E2} = \left\{ \forall (i,j), \quad C_L \leq \tilde{\theta}_{ij}^k \leq C_H \right\}.$$

By lemma 3.7 (i), $\mathbf{P}(\text{E1}) \geq 1 - O\left(\frac{1}{N^2} + \frac{1}{T^2}\right)$. By the assumption that $\tilde{\theta}_{ij}^k$ must be asymptotically bounded, $\mathbf{P}(\text{E2}) = 1$. Therefore, $\mathbf{P}(\text{E1} \cap \text{E2}) \geq 1 - O\left(\frac{1}{N^2} + \frac{1}{T^2}\right)$. Under event E2 and assumption 6, if we have C' sufficiently large such that $C' > C\sqrt{C_H}$ then $\max_{i,j} h_{ij} \leq C\sqrt{C_H}\omega_T < \tau$. Hence $\forall (i,j)$,

$$\begin{aligned}
\mathbf{P}\left(\left| \tilde{\sigma}_{ij}^k \right| > h_{ij} \mid |\sigma_{ij}| > \tau \right) &\geq \mathbf{P}\left(\left| \tilde{\sigma}_{ij}^k - \sigma_{ij} \right| \leq \tau - h_{ij} \right) \\
&\geq \mathbf{P}\left(\left| \tilde{\sigma}_{ij}^k - \sigma_{ij} \right| \leq \tau - C\sqrt{C_H}\omega_T \right) \\
&\geq \mathbf{P}\left(\left| \tilde{\sigma}_{ij}^k - \sigma_{ij} \right| \leq C'\omega_T - C\sqrt{C_H}\omega_T \right) \\
&\geq \mathbf{P}\left(\left| \tilde{\sigma}_{ij}^k - \sigma_{ij} \right| \leq \left(C' - C\sqrt{C_H} \right) \omega_T \right)
\end{aligned}$$

Given event E1, if we have C' sufficiently large such that $C' - C\sqrt{C_H} > c$ then $\forall (i,j) : \left| \tilde{\sigma}_{ij}^k - \sigma_{ij} \right| \leq \left(C' - C\sqrt{C_H} \right) \omega_T$. Hence,

$$\forall (i,j) : \mathbf{P}\left(\left| \tilde{\sigma}_{ij}^k - \sigma_{ij} \right| \leq \left(C' - C\sqrt{C_H} \right) \omega_T \right) \geq \mathbf{P}(\text{E1} \cap \text{E2})$$

and therefore:

$$\forall (i,j) : \mathbf{P}\left(\left| \tilde{\sigma}_{ij}^k \right| > h_{ij} \mid \sigma_{ij} \neq 0 \right) \geq 1 - O\left(\frac{1}{N^2} + \frac{1}{T^2} \right),$$

which establishes our required result. □

Lemma 3.9. *Under assumption 1-5, we have the following results:* for $k < r$,

(i)$\forall (i,j) : \mathbf{P}\left(\left| \tilde{\sigma}_{ij}^k \right| > h_{ij}^k \mid \sigma_{ij}^k = 0 \right) \leq O\left(\frac{1}{N^2} + \frac{1}{T^2}\right)$

Under assumption 1-6, we have: for $k < r$,

(ii) $\forall(i,j):\ \mathbf{P}\left(\left|\tilde{\sigma}_{ij}^{k}\right| < h_{ij}^{k} \mid \sigma_{ij}^{k} \neq 0\right) \leq O\left(\frac{1}{N^2} + \frac{1}{T^2}\right)$

Proof. Similarly to the proof of lemma 3.8, lemma 3.7 (ii) directly proves result (i) required above, and result (ii) above can be proven by similar argument as in the proof of 3.8 (ii). □

3.6 Additional Tables and Figures

Table 3.9: Strong factors only ($r = 5$), $kmax = 8$, ($\alpha = \beta = 0$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations

N	T	no serial correlations in f_t and u_t							
		SC1		SC2		BIC_3		ER	
100	20	5.008	0.089	5.004	0.063	5.000	0.000	5.000	0.000
100	40	5.004	0.063	5.000	0.000	5.000	0.000	5.000	0.000
100	60	5.006	0.077	5.002	0.045	5.000	0.000	5.000	0.000
200	60	5.000	0.000	5.000	0.000	5.000	0.000	5.000	0.000
500	60	5.000	0.000	5.000	0.000	5.000	0.000	5.000	0.000
100	100	5.006	0.077	5.000	0.000	5.000	0.000	5.000	0.000
200	100	5.002	0.045	5.000	0.000	5.000	0.000	5.000	0.000
500	100	5.000	0.000	5.000	0.000	5.000	0.000	5.000	0.000
10	100	2.048	2.332	1.472	2.134	5.000	0.000	4.956	0.224
20	100	5.408	0.653	5.218	0.524	5.000	0.000	5.000	0.000
40	100	5.318	0.545	5.054	0.226	5.000	0.000	5.000	0.000
60	100	5.198	0.451	5.028	0.165	5.000	0.000	5.000	0.000
60	200	5.226	0.455	5.014	0.118	5.000	0.000	5.000	0.000
60	500	5.216	0.445	5.010	0.100	5.000	0.000	5.000	0.000

Table 3.10: Strong factors only ($r = 5$), $kmax = 8$, ($\alpha = \beta = 0.5$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations

N	T	serial correlations in f_t and u_t ($\alpha = \beta = 0.5$)							
		SC1		SC2		BIC_3		ER	
100	20	5.106	0.327	5.060	0.238	5.000	0.000	5.000	0.000
100	40	5.112	0.352	5.032	0.187	5.000	0.000	5.000	0.000
100	60	5.118	0.347	5.014	0.118	5.000	0.000	5.000	0.000
200	60	5.074	0.270	5.026	0.171	5.000	0.000	5.000	0.000
500	60	5.004	0.063	5.032	0.176	5.000	0.000	5.000	0.000
100	100	5.078	0.283	5.006	0.077	5.000	0.000	5.000	0.000
200	100	5.052	0.240	5.006	0.077	5.000	0.000	5.000	0.000
500	100	5.006	0.077	5.008	0.089	5.000	0.000	5.000	0.000
10	100	1.870	2.297	1.374	2.088	5.000	0.000	4.960	0.225
20	100	5.520	0.750	5.292	0.589	5.000	0.000	5.000	0.000
40	100	5.462	0.673	5.136	0.349	5.000	0.000	5.000	0.000
60	100	5.432	0.634	5.124	0.336	5.000	0.000	5.000	0.000
60	200	5.432	0.625	5.072	0.266	5.000	0.000	5.000	0.000
60	500	5.346	0.575	5.032	0.176	5.000	0.000	5.000	0.000

Table 3.11: Weak factors only ($r = 5, \gamma = 1/3$), $kmax = 8$, ($\alpha = \beta = 0$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations

N	T	no serial correlations in f_t and u_t							
		SC1		SC2		BIC_3		ER	
100	20	5.008	0.089	5.000	0.000	3.846	0.619	5.000	0.000
100	40	5.012	0.109	5.004	0.063	3.180	0.679	5.000	0.000
100	60	5.004	0.063	5.000	0.000	2.768	0.683	5.000	0.000
200	60	5.004	0.063	5.000	0.000	2.160	0.651	5.000	0.000
500	60	5.000	0.000	5.000	0.000	1.454	0.642	5.000	0.000
100	100	5.012	0.109	5.000	0.000	2.248	0.651	5.000	0.000
200	100	5.004	0.063	5.000	0.000	1.430	0.615	5.000	0.000
500	100	5.000	0.000	5.000	0.000	0.676	0.576	5.000	0.000
10	100	1.588	1.975	1.030	1.623	4.796	0.403	4.522	0.909
20	100	5.366	0.611	5.252	0.552	3.914	0.599	4.994	0.077
40	100	5.302	0.532	5.044	0.205	3.226	0.663	5.000	0.000
60	100	5.226	0.446	5.014	0.118	2.746	0.677	5.000	0.000
60	200	5.240	0.459	5.008	0.089	2.282	0.657	5.000	0.000
60	500	5.330	0.523	5.018	0.133	1.688	0.687	5.000	0.000

Table 3.12: Weak factors only ($r = 5, \gamma = 1/3$), $kmax = 8$, ($\alpha = \beta = 0.5$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations

N	T	serial correlations in f_t and u_t ($\alpha = \beta = 0.5$)							
		SC1		SC2		BIC_3		ER	
100	20	5.120	0.355	5.062	0.257	4.168	0.587	4.992	0.109
100	40	5.142	0.387	5.050	0.218	3.486	0.653	5.000	0.000
100	60	5.112	0.334	5.032	0.176	3.084	0.665	5.000	0.000
200	60	5.084	0.278	5.028	0.165	2.574	0.652	5.000	0.000
500	60	5.002	0.045	5.030	0.182	1.996	0.614	5.000	0.000
100	100	5.066	0.256	5.006	0.077	2.500	0.662	5.000	0.000
200	100	5.048	0.232	5.006	0.077	1.884	0.669	5.000	0.000
500	100	5.000	0.000	5.000	0.000	1.194	0.611	5.000	0.000
10	100	1.630	2.040	1.190	1.711	4.770	0.431	4.436	0.944
20	100	5.468	0.717	5.282	0.575	3.914	0.622	4.982	0.204
40	100	5.436	0.656	5.104	0.312	3.380	0.670	5.000	0.000
60	100	5.428	0.624	5.150	0.357	2.950	0.661	5.000	0.000
60	200	5.358	0.571	5.060	0.238	2.370	0.659	5.000	0.000
60	500	5.348	0.569	5.044	0.205	1.798	0.631	5.000	0.000

Table 3.13: Weak factors only ($r = 5, \gamma = 1/5$), $kmax = 8$, ($\alpha = \beta = 0$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations

N	T	no serial correlations in f_t and u_t							
		SC1		SC2		BIC_3		ER	
100	20	5.008	0.109	4.996	0.063	1.356	0.631	4.986	0.118
100	40	5.018	0.147	5.002	0.045	0.154	0.361	5.000	0.000
100	60	5.004	0.063	5.000	0.000	0.012	0.109	5.000	0.000
200	60	5.002	0.045	5.000	0.000	0.000	0.000	5.000	0.000
500	60	5.000	0.000	5.000	0.000	0.000	0.000	5.000	0.000
100	100	5.004	0.063	5.000	0.000	0.000	0.000	5.000	0.000
200	100	5.000	0.000	5.000	0.000	0.000	0.000	5.000	0.000
500	100	5.000	0.000	5.000	0.000	0.000	0.000	5.000	0.000
10	100	1.706	1.713	1.204	1.388	4.134	0.649	3.656	1.468
20	100	5.402	0.649	5.130	0.869	1.722	0.665	4.928	0.308
40	100	5.336	0.558	5.028	0.165	0.304	0.482	5.000	0.000
60	100	5.214	0.434	5.022	0.147	0.024	0.153	5.000	0.000
60	200	5.248	0.472	5.008	0.089	0.002	0.045	5.000	0.000
60	500	5.280	0.492	5.014	0.118	0.000	0.000	5.000	0.000

Table 3.14: Weak factors only ($r = 5, \gamma = 1/5$), $kmax = 8$, ($\alpha = \beta = 0.5$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations

N	T	serial correlations in f_t and u_t ($\alpha = \beta = 0.5$)							
		SC1		SC2		BIC_3		ER	
100	20	5.112	0.357	5.014	0.387	2.172	0.654	4.896	0.444
100	40	5.160	0.398	5.028	0.165	0.698	0.579	5.000	0.000
100	60	5.112	0.340	5.018	0.133	0.158	0.371	5.000	0.000
200	60	5.078	0.276	5.028	0.165	0.020	0.140	5.000	0.000
500	60	5.004	0.063	5.050	0.218	0.000	0.000	5.000	0.000
100	100	5.060	0.246	5.008	0.089	0.010	0.100	5.000	0.000
200	100	5.044	0.249	5.008	0.089	0.000	0.000	5.000	0.000
500	100	5.002	0.045	5.004	0.063	0.000	0.000	5.000	0.000
10	100	1.652	1.655	1.184	1.362	4.174	0.610	3.658	1.455
20	100	5.472	0.747	5.080	0.965	1.870	0.686	4.900	0.427
40	100	5.516	0.729	5.136	0.343	0.518	0.553	4.998	0.045
60	100	5.442	0.666	5.114	0.324	0.100	0.300	5.000	0.000
60	200	5.374	0.554	5.050	0.218	0.004	0.063	5.000	0.000
60	500	5.366	0.611	5.050	0.218	0.000	0.000	5.000	0.000

Table 3.15: Weak factors only ($r = 5, \gamma = 1/10$), $kmax = 8$, ($\alpha = \beta = 0$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations

N	T	no serial correlations in f_t and u_t							
		SC1		SC2		BIC_3		ER	
100	20	3.606	1.784	1.628	1.661	0.000	0.000	3.838	1.918
100	40	5.004	0.063	4.954	0.310	0.000	0.000	4.998	0.045
100	60	5.006	0.077	4.998	0.045	0.000	0.000	4.996	0.063
200	60	5.000	0.000	5.000	0.000	0.000	0.000	5.000	0.000
500	60	5.000	0.000	5.000	0.000	0.000	0.000	5.000	0.000
100	100	5.004	0.063	5.000	0.000	0.000	0.000	5.000	0.000
200	100	5.004	0.063	5.000	0.000	0.000	0.000	5.000	0.000
500	100	5.000	0.000	5.000	0.000	0.000	0.000	5.000	0.000
10	100	1.300	1.173	0.724	0.926	2.562	0.766	3.010	2.006
20	100	4.586	1.234	2.458	1.598	0.004	0.063	3.472	1.477
40	100	5.284	0.494	4.880	0.535	0.000	0.000	4.940	0.359
60	100	5.246	0.449	5.010	0.134	0.000	0.000	5.000	0.000
60	200	5.252	0.474	5.010	0.100	0.000	0.000	5.000	0.000
60	500	5.280	0.492	5.016	0.126	0.000	0.000	5.000	0.000

Table 3.16: Weak factors only ($r = 5, \gamma = 1/10$), $kmax = 8$, ($\alpha = \beta = 0.5$), the number of factors reported is averaged out of 500 simulations, on the right side are the standard deviations

N	T	serial correlations in f_t and u_t ($\alpha = \beta = 0.5$)							
		SC1		SC2		BIC_3		ER	
100	20	3.944	1.452	2.578	1.708	0.038	0.191	1.702	2.053
100	40	5.118	0.390	4.950	0.309	0.000	0.000	4.496	1.369
100	60	5.124	0.359	5.016	0.154	0.000	0.000	4.994	0.077
200	60	5.064	0.268	5.026	0.159	0.000	0.000	5.000	0.000
500	60	5.006	0.077	5.034	0.181	0.000	0.000	5.000	0.000
100	100	5.070	0.271	5.004	0.063	0.000	0.000	5.000	0.000
200	100	5.048	0.214	5.008	0.089	0.000	0.000	5.000	0.000
500	100	5.000	0.000	5.000	0.000	0.000	0.000	5.000	0.000
10	100	1.408	1.149	0.816	0.951	2.562	0.753	2.750	1.987
20	100	4.602	1.413	2.628	1.720	0.008	0.089	3.284	1.491
40	100	5.458	0.691	4.972	0.558	0.000	0.000	4.858	0.561
60	100	5.530	0.680	5.114	0.330	0.000	0.000	4.980	0.260
60	200	5.318	0.538	5.056	0.230	0.000	0.000	5.000	0.000
60	500	5.356	0.578	5.034	0.181	0.000	0.000	5.000	0.000

4 Applications of weak factor model in large dimensional covariance matrix estimation

4.1 Introduction

Suppose we want to estimate the covariance matrix Σ of a homoskedaticity multivariate process Y_t, the sample covariance matrix constructed from T observations are very ill-behaved when the cross-section dimension N is as large as T. The literature discussing the issue with large-dimensional sample covariance matrix is extremely large, e.g. some can be found in Ledoit and Wolf (2004), Fan et al. (2011, 2013) and the references therein.

As discussed in Section 1.2.2, proposing factor structure gives great advantage in estimating large covariance matrix. Recall that if we assume Y_t has factor structure, i.e. $Y_t = \Lambda f_t + u_t$, then we have $\Sigma = \Lambda \Sigma_f \Lambda' + \Sigma_u$. In here, Σ_f is the $r \times r$ covariance matrix of f_t and Σ_u is the covariance matrix of u_t. While Σ_f can be estimated by the covariance matrix of f_t (assuming that r is small), Σ_u requires more attention as it is still a $N \times N$ matrix.

Originally, it is assuming to be a diagonal matrix, with i-diagonal entry estimated by the sample variance of $\tilde{u}_{it} = y_{it} - \tilde{\lambda}_i' f_t$ (or $= y_{it} - \tilde{\lambda}_i' \tilde{f}_t$ if the factors are not observed and need to be extracted from Y_t)[1]. However, after approximate factor

[1] See Chapter 2 for notations and the factors identification techniques.

model is introduced, it is more reasonable to relax the diagonal restriction and only assume it is sparse.

Therefore, as the final part of constructing the estimator for Σ, we need to have good estimators for Σ_u. In this chapter I focus on the recently proposed estimator for Σ_u in high-dimensional setting, which is discussed in the principle orthogonal complement thresholding (POET) estimators proposed by Fan et al. (2013).

4.2 The POET estimators for Σ and Σ_u

Recently, the POET estimators proposed by Fan et al. (2013) has provided a very useful technique for estimating the covariance matrix of large multivariate series. The main idea of this method is to decompose the covariance matrix into a low rank and a sparse components, which is implied by proposing the approximate factor structure into the observed data.

In fact, this is not the only attempt to identify the decomposition of Σ into a low rank ($\Lambda\Sigma_f\Lambda'$) and a sparse matrix (Σ_u), for example see Wright et al. (2009), Lin et al. (2009), etc. Comparing to these approaches, the POET requires a stronger assumption for the low rank part, i.e. the systematic eigenvalues grow with rate $\asymp N$. When this assumption is satisfied, we can identify exactly the factors and loadings space. This assumption is standard in factor analysis literature, where we not only require $\Lambda\Sigma_f\Lambda'$ but we also need the factors (and loadings) values. In this case, the factors and loadings can be estimated consistently by PCA techniques.

The factor structure adopted by Fan et al. only includes strong factors, and hence the large signal-to noise ratio helps to improve the rate of convergence of the POET estimator. In the discussion of Fan et al. (2013), Yu and Samworth (2013) point out that this condition can be loosen in some certain cases. However, Yu and Samworth do not discuss about the effect of this condition to factors estimation, which is now shown in our Theorem 2.1. Furthermore, a problem arises for estimating the number of PCs (or factors) because now the gap that separate the eigenvalues of the factors

part and the idiosyncratic errors part is narrow. In this section, I contribute some discussion regarding to the POET estimator, particularly with results in Chapter 2 and 3 we can show that the POET estimator is still consistent under our weak factor model. Furthermore, the number of factors is not an important matter for the consistency of POET. In fact, any numbers of PCs greater than or equal to r will make the POET estimator consistent.

4.2.1 Steps for constructing POET estimator

The construction of POET estimator can be summarised in the following steps:

1. Suggest the number of factors k to be some known values, or estimate by some given criteria.

2. Given k, estimate $(\tilde{\Lambda}^k, \tilde{F}^k)$ by PCA as in Chapter 2.

3. Then the sample residuals covariance matrix are construed: $\tilde{\Sigma}_u^k = (\tilde{\sigma}_{ij}^k)_{N\times N}$

4. Applying thresholding operator to $\tilde{\Sigma}_u^k$ to obtain an estimator for Σ_u, i.e.

$$\tilde{\Sigma}_u^{k,\tau} = \left(\tilde{\sigma}_{ij}^{k,\tau}\right)_{N\times N} \quad \text{and} \quad \begin{cases} \tilde{\sigma}_{ij}^{k,\tau} = \tilde{\sigma}_{ij}^k & \text{for } i = j \\ \tilde{\sigma}_{ij}^{k,\tau} = s_{ij}^k\left(\tilde{\sigma}_{ij}^k\right) & \text{for } i \neq j \end{cases} \tag{4.1}$$

   The operator $s_{ij}(.)$ is the adaptive thresholding. For example, the adaptive hard thresholding is the one used above when estimating the sparsity level:

$$s_{ij}^k(\tilde{\sigma}_{ij}^k) = \tilde{\sigma}_{ij}^k \mathbf{I}\left(\left|\tilde{\sigma}_{ij}^k\right| > h_{ij}^k\right)$$

   Other thresholding rules are the soft thresholding, smoothly clipped absolute deviation and the adaptive lasso (see Rothman et al, (2009) for more discussions).

5. Finally, the POET estimator for Σ is constructed as: $\tilde{\Sigma}^{k,\tau} = \tilde{\Lambda}^k \tilde{\Lambda}^{k'} + \tilde{\Sigma}_u^{k,\tau}$ since

$\tilde{F}^{k\prime}\tilde{F}^k/T = I_k$ by our restriction in PCA.

4.2.2 Spiked eigenvalues and the choice for the number of factors

There are a large number of discussants contributing comments to the POET method of Fan et al. (2013). Many of them gave their concerns about the spiked eigenvalues, which is implied by Assumption 0 (i) in Chapter 1. For example, Yu and Samworth (2013) suggest a weaker version of the pervasive condition, which is a specific case of our proposed model where the factors have strengths N^α for $\alpha \in (0,1)$. Therefore, the results we have in this paper confirm that the main results of Fan et al. (2013) still go through, if the strength of the weakest factor grows at least faster than $\max(\sqrt{N}, N\sqrt{\log N/T})$. Of course, the rate of convergence of $\tilde{\Sigma}^{k,\tau}$ will depend on the estimated values for the factors and loadings, and therefore will be slower when the factors are not pervasive.

In addition, when building up to our main result, Lemma 3.3 implies that the POET method can be used with any values of $k \geq r$. Consequently, even if one can not determine a reliable number of factors, choosing a relative large value to start with is recommended. Based on our developed lemmas, the following theorem can be derived.

Theorem 4.1. *Under assumptions 1-5, if $k \geq r$ then*

(i) $\left\|\tilde{\Sigma}_u^{k,\tau} - \Sigma_u\right\| = O_p(\sqrt{\frac{\log N}{T}} + \frac{\sqrt{N}}{[d_r(N)]} + \frac{N}{\sqrt{T}[d_r(N)]})$

(ii) $\left\|\left(\tilde{\Sigma}_u^{k,\tau}\right)^{-1} - \Sigma_u^{-1}\right\| = O_p(\sqrt{\frac{\log N}{T}} + \frac{\sqrt{N}}{[d_r(N)]} + \frac{N}{\sqrt{T}[d_r(N)]})$

Apart from the idiosyncratic covariance matrix, using $\tilde{\Sigma}^{k,\tau}$ as an estimator for Σ is not consistent under the same matrix norm as above. However, as shown in Fan et al. (2013), the entropy loss matrix norm $\left\|\tilde{\Sigma}^{k,\tau} - \Sigma\right\|_\Sigma$ converges to 0, whereas $\left\|\tilde{\Sigma} - \Sigma\right\|_\Sigma$ does not converge if $N > T$.

4.2.3 Simulated examples for demonstration

In this section, I demonstrate the idea behind Theorem 4.1 with a simulated experiment. Consider the same data generation processes as in Section 3.3 of Chapter 3, i.e.

$$Y_t = \Lambda^{(1:m)} f_t^{(1:m)} + \gamma \Lambda^{(m+1:r)} f_t^{(m+1:r)} + u_t.$$

For simplicity I ignore the serial correlation parameters α and β as in Chapter 3. The weakness of the factors are obtained by letting $\gamma = \frac{1}{5}$. In addition, I choose $r = 10$ which means that the model will have a total of 10 factors. Three scenarios (all-strong, all-weak, mix-strong-and-weak factor models) will be applied to demonstrate the results. In each scenario setting, I try $k = 1$ to 20 (k is the number of principle components extracted to estimate the idiosyncratic error covariance matrix) and verify that for any $k \geq 10$ the estimators $\tilde{\Sigma}_u^{k,\tau}$ are all very closed to Σ_u. This consolidates the result we have in Theorem 4.1, that is POET will work as long as we extract more than r principle components.

- In the case of **strong factors**, Figure 4.1 reports $\left\|\tilde{\Sigma}_u^{k,\tau} - \Sigma_u\right\|$ for $k = 1$ to 20 for 20 simulated models. Σ_u is known as we use this to generate u_t, and $\tilde{\Sigma}_u^{k,\tau}$ is estimated as described above. It can be seen from there that when $k > r$ the gain in consistency of $\tilde{\Sigma}_u^{k,\tau}$ is negligible.

- In the case of **mix-strong-and-weak factors** (I let $m = 4$ to represent 4 strong factors and 6 week factors), once all the strong factors are extracted, $\tilde{\Sigma}_u^{k,\tau}$ is reasonably closed to Σ_u, which is what we expect. However, better estimator is obtained if we use at least the true number of total factors, see Figure 4.2 for results on 20 simulated models.

- In the case of **all-weak factors**, we can see that the change of $\left\|\tilde{\Sigma}_u^{k,\tau} - \Sigma_u\right\|$ right at the value where $k = r$ is more gradual (see Figure 4.3), which is due to the fact that now the factors are less separated from the idiosyncratic errors.

In all cases, it can be seen that for any $k \geq 10$ the estimators $\tilde{\Sigma}_u^{k,\tau}$ should be consistent for Σ_u. This agrees with the main idea behind the method for selecting the number of factors in Chapter 3, which states that once we extract more than r principle components, the sparsity level of the sample idiosyncratic covariance matrix will not significantly change.

Figure 4.1: $\left\|\tilde{\Sigma}_u^{k,\tau} - \Sigma_u\right\|$, $k = 1 : 20$ for 20 different strong factor models, $T = 200$, $N = 200$ and $r = 10$

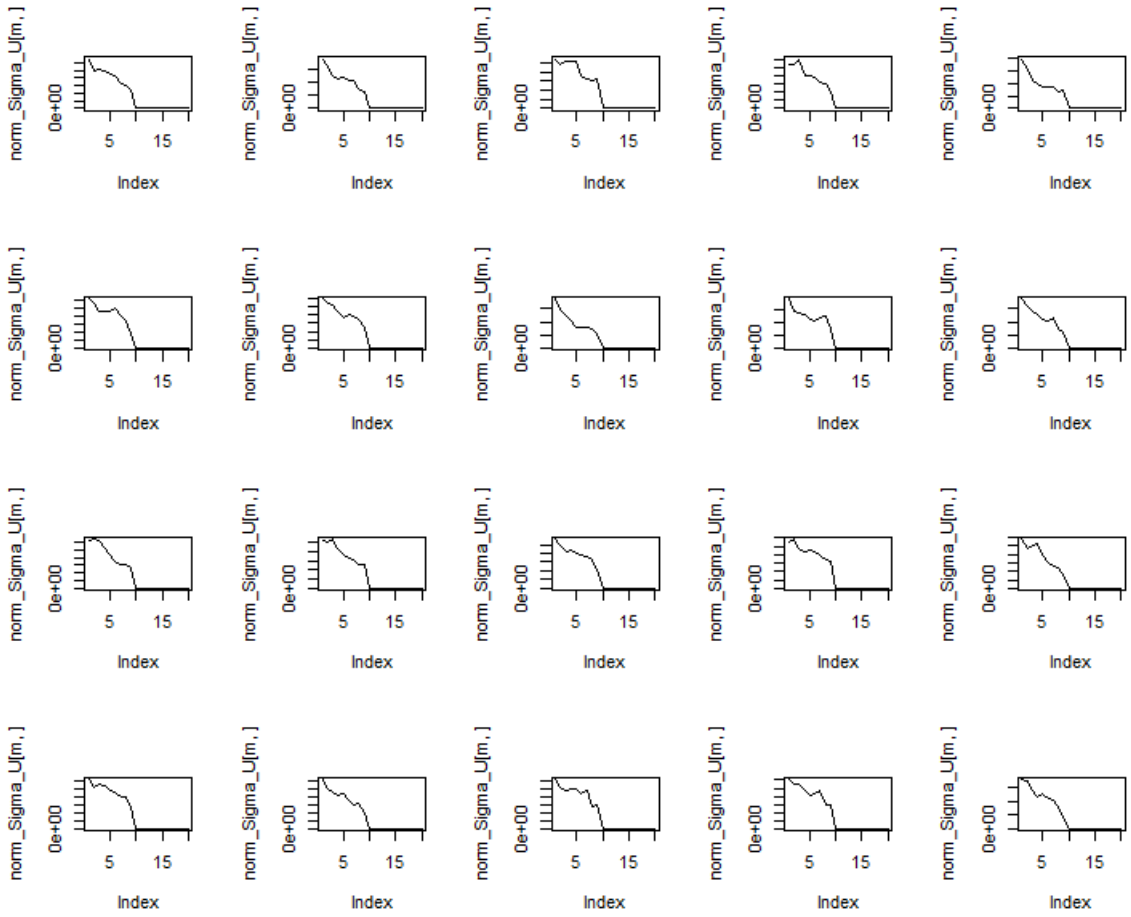


Figure 4.2: $\left\|\tilde{\Sigma}_u^{k,\tau} - \Sigma_u\right\|$, $k = 1 : 20$ for 20 different mixture strong and weak factor models, $T = 200$, $N = 200$ and $r = 10$, in which the first 4 factors are strong ($\gamma = \frac{1}{5}$).

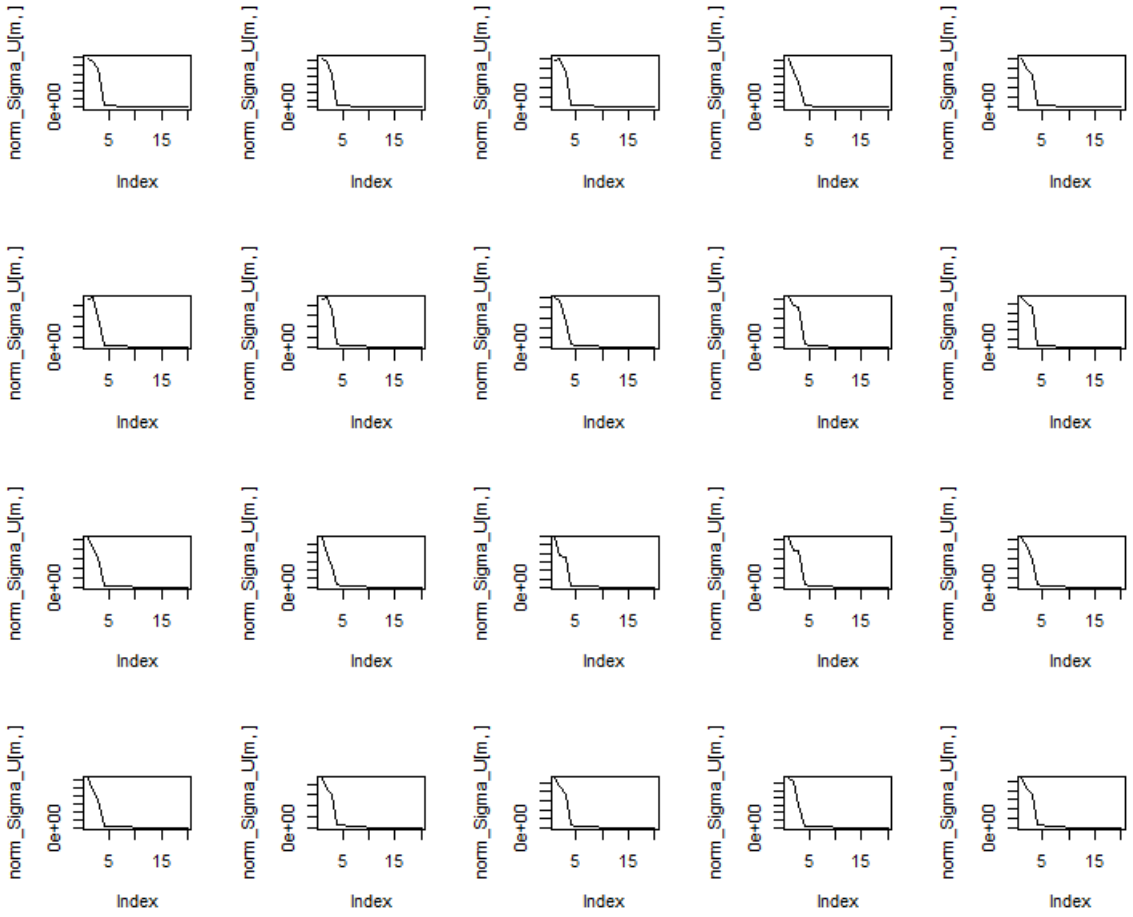
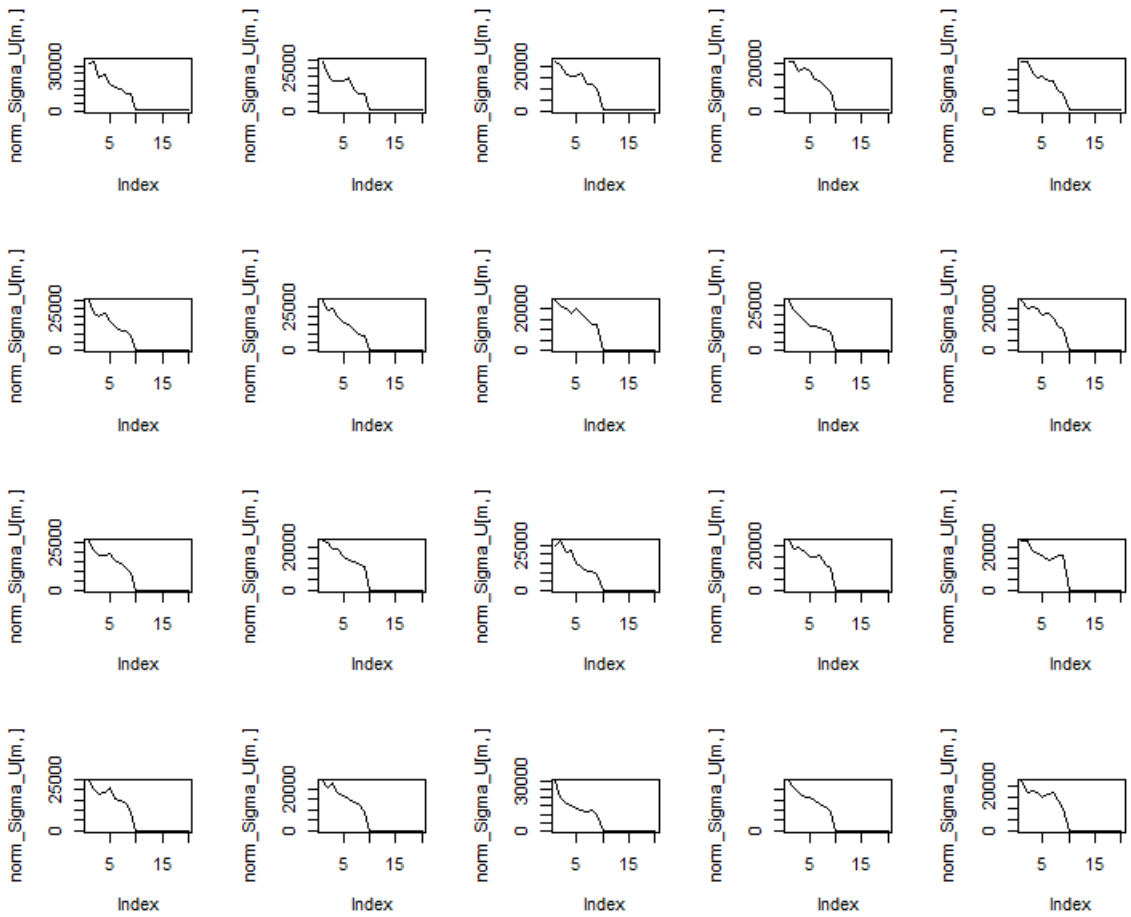


Figure 4.3: $\left\|\tilde{\Sigma}_u^{k,\tau} - \Sigma_u\right\|$, $k = 1 : 20$ for 20 different weak factor models ($\gamma = \frac{1}{5}$), $T = 200$, $N = 200$ and $r = 10$



4.3 Remarks

The POET estimator is ultimately to obtain an estimator for Σ. However, in the path of constructing this, we need to obtain a consistent estimator for the idiosyncratic errors covariance matrix Σ_u, assuming it is sparse. The result we show in this chapter is useful in two aspects. Firstly, we can confirm the process is still valid even when the factors are not all pervasive. Secondly, it makes the estimator for Σ_u less dependent on the number of factors.

Finally, it is interesting to note that the estimator for Σ_u also has many applica-

tions in practice. For example, the statistics used in the asset pricing theory require an estimator for Σ_u^{-1}. More precisely, suppose we have a multivariate linear factor model:

$$Y_t = \alpha + \Lambda f_t + u_t$$

and we wish to test if the vector α is zero. This will support the argument of Ross (1976) for the Arbitrage Pricing Theory, that is the expected excessive return of any financial asset i at time t (y_{it}) should equal the expected excessive returns of some risk factors (f_t) times the loading (λ_i), if the market is frictionless. Then when Σ_u^{-1} is known, the Wald statistics includes the terms $\hat{\alpha}'\Sigma_u^{-1}\hat{\alpha}$ (see Sentana (2009) for a survey in various tests). Therefore an estimator for Σ_u^{-1} will be useful for such applications.

In addition, the rate of convergence obtained in Chapter 2 for weak factors can lead to a possible improvement when estimating the covariance matrix if segmentation can be applied to the original data set. Particularly, in this case the original time series can be divided into several sub-vectors (region), and those sub-vectors are generated by regional factors that are both contemporaneously and serially uncorrelated across regions. Therefore, they can be modeled separately. However, out-of-sector dependence is still possible, due to the idiosyncratic shocks that are not included in the factors. This factor model has a loading matrix similar to the one in (1.4), so e.g. in the case of 3 regional factors for 3 regions, we can have the following representation:

$$\begin{bmatrix} Y_{1t} \\ Y_{2t} \\ Y_{3t} \end{bmatrix} = \begin{bmatrix} \Lambda_1 & 0 & 0 \\ 0 & \Lambda_2 & 0 \\ 0 & 0 & \Lambda_3 \end{bmatrix} \begin{bmatrix} f_t^{(1)} \\ f_t^{(2)} \\ f_t^{(3)} \end{bmatrix} + u_t. \tag{4.2}$$

In here (4.2), Λ_i and Y_{1t} are $N_i \times 1$ vector, and $N_1 + N_2 + N_3 = N$. In order to differentiate between a factor and a idiosyncratic errors, we assume that the regional factors are strongly pervasive in each region, and $N_i \to \infty$, $\forall i$. In this case, the $N \times N$ covariance matrix Σ of Y_t can be decompose into the following (assuming

$\text{cov}(f_t) = I_3$):

$$\Sigma = \begin{bmatrix} \Lambda_1\Lambda_1' & 0 & 0 \\ 0 & \Lambda_2\Lambda_2' & 0 \\ 0 & 0 & \Lambda_3\Lambda_3' \end{bmatrix} + \Sigma_u. \tag{4.3}$$

It can be seen from Chapter 2 that if can segmentate Y_t into Y_{it} for $i = 1, 2, 3$ and estimate the factors and loadings for each sub vector than the rate of convergence can be improved. This can be promising for a future research.

4.4 Proofs of results

4.4.1 Proofs of Theorem 4.1

Both parts of this theorem can be proved from the following results in Lemma (3.7) of Chapter 3. Particularly, recall that we have the following two results: for $k \geq r$,

$$\mathbf{P}\left(\max_{i,j}\left|\tilde{\sigma}_{ij}^k - \sigma_{ij}\right| \leq c\omega_T\right) \geq 1 - O\left(\frac{1}{N^2} + \frac{1}{T^2}\right),$$

and

$$\exists\,(C_L, C_H) \text{ such that } \forall(i,j),\ \mathbf{P}\left(C_L \leq \tilde{\theta}_{ij}^k \leq C_H\right) \geq 1 - O\left(\frac{1}{N^2} + \frac{1}{T^2}\right).$$

These two results are equivalent to the probability of events A1 and A2 approaching 1 in the proof of Theorem A.1 of Fan et al. (2013). As a result, Theorem 4.1 follows directly.

5 Factor models selections

5.1 Observed or un-observed factors model

In this chapter I wish to discuss about a general method that can be used for factor models selection. Due to the advantage of capturing a large proportion of movements in big data, factor analysis rapidly becomes more popular in practice. Parallel to this, researchers nowadays can face a problem with choosing between many potential factors for a same data set, e.g. in asset pricing, returns of financial assets can be explained by many types of factors. In this case, one will have to make a decision of whether to use observed factors (such as Fama-French, Macro factors, etc.) or latent factors estimated by PCs. There are some studies that attempt to link the factors statistically extracted to the observed ones, in order to provide more meaningful insight[1]. However, if they are not statistically identical then one needs to decide which factors fit better to the observed data.

In this thesis, Chapter 3 studies a criterion for choosing the number of latent factors, which is equivalently to select an optimal model of unobserved factors estimated by PCA. More existing criteria for choosing the number of factors can be found in the discussion in there. On the other hand, observed factors models also have their long establishments in asset pricing, with many factors models proposed, including the well-known Fama-French 3-factor model.

To choose the best observed linear factor model, some well-known methods are present, for example Sparks et al. (1983) generalise Mallow (1973) C_p criterion to

[1] e.g. see Bai and Ng (2006)

the multivariate model. The criterion for a model F with k factors and residuals sample covariance $\tilde{\Sigma}_u^k$ (conditional on F) is:

$$C_p = (T - k_{max}) \left(\tilde{\Sigma}_u^{kmax} \right)^{-1} \tilde{\Sigma}_u^k + (2k - T) I_N \tag{5.1}$$

In here, there can be some confusion in the notation, because for observed factors model, k may not be different across models, so it should be understood that $\tilde{\Sigma}_u^k$ is conditional of the set of factors used. In addition,$\tilde{\Sigma}_u^{kmax}$ refers to the case where all the available factors are included.

Other well-known criteria for variables selection are the AIC and BIC under the framework of maximum log-likelihood and a penalty function. The multivariate versions of them for observed factor models are as follow:

$$AIC = \log \left(\left| \tilde{\Sigma}_u^k \right| \right) + \frac{[2kN + N(N+1)]}{T} \tag{5.2}$$

and

$$BIC = T \log \left(\left| \tilde{\Sigma}_u^k \right| \right) + \left(k(N+1) + \frac{[N(N+1)]}{2} \right) \log(T) \tag{5.3}$$

Notice that these are different than the AIC and BIC for unobserved factors in Choi and Jeong (2013).

We already see how the sparsity level is estimated for the case where the factors are unobserved. The key parameter in the thresholding value is ω_T, which is obtained from the convergence rate of $\max_{i \leq N} \frac{1}{T} \sum_{t=1}^{T} \left(\tilde{u}_{it}^k - u_{it}^k \right)^2$ for $k < r$ and $\max_{i \leq N} \frac{1}{T} \sum_{t=1}^{T} \left(\tilde{u}_{it}^k - u_{it} \right)^2$ for $k \geq r$ where $\left\{ \tilde{u}_{it}^k \right\}$ are the residuals after subtracting the estimated factors. For the observed case, this quantity will change due to the fact that we no longer need to estimate the factors themselves. This is derived in

Fan et al. (2011). Therefore, the thresholding function for each case is as follows:

$$\begin{cases} \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}} & \text{for estimated factors} \\ \sqrt{\frac{\log N}{T}} & \text{for observed factors} \end{cases}$$

Notice that the above values for ω_T correspond to the case with strong factor only, when the factors are not all strong, we can use the values for ω_T as shown in the criteria SC1 and SC2. If one concerns about a situation where some observed factors are weak, some modification can be done to modify ω_T in this case, but for simple illustration I will not pursue it.

To estimate the sparsity level, the hard thresholding procedure as in (3.3) is still applied. As the other part of the whole criterion, the choice of penalty function has already been discussed, and it should not be sensitive to whether factors are strong or weak. Particularly, I will use the following criterion:

$$\text{SC:} \quad \tilde{m}(\Sigma_u^k) + \frac{k\, N^{1/2}}{10} \tag{5.4}$$

where the estimated sparsity level $\tilde{m}(\Sigma_u^k)$ is defined as:

$$\tilde{m}(\Sigma_u^k) = \begin{cases} \max_{i \leq N} \sum_{j=1}^{N} \mathbf{I}\left(\frac{|\tilde{\sigma}_{ij}^k|}{|\tilde{\sigma}_{ii}^k \tilde{\sigma}_{jj}^k|} > C\left(\sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}} \right) \right) & \text{for estimated factors} \\ \max_{i \leq N} \sum_{j=1}^{N} \mathbf{I}\left(\frac{|\tilde{\sigma}_{ij}^k|}{|\tilde{\sigma}_{ii}^k \tilde{\sigma}_{jj}^k|} > C\left(\sqrt{\frac{\log N}{T}} \right) \right) & \text{for observed factors.} \end{cases} \tag{5.5}$$

However, a key problem arising when evaluating both observed and unobserved factors at the same time is whether it is fair to use different thresholding parameter for each case. Some prior study shows that for the same $\tilde{\Sigma}_u^k$, the thresholding parameters can have big impact on the sparsity level estimated. If we over- or under- estimate the sparsity level, it would not be sensible to evaluate different models based on the sparsity level. Therefore, it is really important that data-driven choice for C is applied to all of our thresholding procedure, to minimise the risk of mis-specify the

sparsity level.

5.2 Empirical Analysis in the FTSE 100 market

In this empirical analysis, I use the multivariate data containing returns of 66 stocks in the FTSE 100. Most of the empirical studies and simulations use data from the US markets to validate the Fama-French model, therefore I wish to try using data in the UK market to extend the study in a larger scale.

5.2.1 Models description

The observed factor models used in this section are the 1-factor CAPM model, the 3-factor Fama-French model, and the 4-factor Carhart model (1997). These factors include the market returns minus the risk-free rate (Rm-Rf), Small-Minus-Big (SMB), High-minus-low (HML) and the momentum factors (UMD). Rm-Rf is the benchmark describing the premium return of the whole FTSE 100 market over the risk-free rate, which is exactly the factor we have in the well-known CAPM model. In this case, it is the value-weight return on all FTSE 100 stocks minus the one-month US Treasury bill rate (obtained from Ibbotson Associates). SMB factor represents the excess returns of stock with small capitalisation to stock with big capitalisation. HML factor represents the excess returns of stock with high book-to-market ratio to stock with low book-to-market ratio. The reason for including these two factors is that Fama and French (1993) observe that asset with small capitalisation and high book-to-market ratio (value stock) tends to give higher return than the rest. Momentum factor measures the excess of high return stocks and low return stocks recently, because it is observed that stock which recently perform well can keep its momentum[2]. Momentum factor is added here to see if it is in fact a good factor for returns in the UK market. We usually refer to these 4 factors as Fama-French type. The values of these factors are obtained from The Xfi Centre for Finance and

[2]See Carhart (1997) for more discussion.

Investment at the university of Exeter. More descriptions on how to construct the values of the factors can be found on Kenneth French's website.

I use the monthly returns of 66 stocks ($N = 66$) in the FTSE 100 available from 1 Jan 2003 to 31 Dec 2010 ($T = 95$). The returns are calculated by the logarithm of the ratios of prices between the first date of any two consecutive months in this period. This sample period includes the time when the global financial crisis happens in 2008. We expect to have high volatility and good amount of cross-section correlation in the idiosyncratic errors in the asset returns.

5.2.2 Empirical Results

In Table 5.1 I report the results from all the criteria for the observed model selection, including the estimated level of sparsity $\tilde{m}(\Sigma_u^k)$ and the sparsity criterion (SC) $\tilde{m}(\Sigma_u^k) + k\, g(N)/10$. In here, the data-driven method is used to select the constant when estimating $\tilde{m}(\Sigma_u^k)$.

We can see that all the criteria perform very differently, for example the AIC suggests market return, SMB, HML as the 3 factors in the best model whereas the BIC suggests market return only. Other criteria also give different result. Therefore further examination should be taken when choosing the model, such as cross-validation.

For the unobserved factors estimated by PCs, we also try different criteria to examine how many factors for the FTSE 100 asset returns, see Table 5.2. Notice that the criterion SC1 and SC2 in Chapter 3 have the adjusted value of ω_T for the possibility of weak factors, but they still work if the factors are all strong. In this case, it is suggested that 4 and 3 factors exist among 66 components by the SC1 and SC2 respectively.

For better comparison with results in table 5.1, we also apply the sparsity criterion under the assumption that all factors are strong. In this case, we use $\omega_T = \sqrt{\frac{\log N}{T}} + \frac{1}{\sqrt{N}}$, and the constant of thresholding are chosen by data-driven method. This is shown in Table 5.3. Based only on the column SC, it can be seen from table 5.1 and

5.3 that only 2 factors should be included in the model, which are market returns and SMB in the case of observed factors. However, using the 2 PCs as the estimators for the latent factors yield slightly better result, as the SC value of PCs factors is lower.

Also, these numbers have meaning if we ignore the penalty part and look at $\tilde{m}(\Sigma_u^k)$ only. In this case, this indicates what conditional on these factors, any idiosyncratic errors are at most correlated with 5 or 3 others in the cross-section, so these factors capture quite well the amount of correlations for returns of 66 companies in the FTSE 100 during this period.

Table 5.1: Some criteria for each observed factor model

Factors included	$\tilde{m}(\Sigma_u^k)$	SC	AIC	BIC	$\|C_p\|$	$\left\|\tilde{\Sigma}_u^k\left(\frac{T+k}{T-k}\right)\right\|$
Rm-Rf	12	12.8124	-351.01	**-27526.22**	16248.08	0.2488442
SMB	61	61.8124	-348.1545	-27254.95	31729.10	0.2688127
HML	39	39.8124	-348.3557	-27274.06	26276.41	0.2779500
UMD	60	60.8124	-347.0888	-27153.70	35949.29	0.3419941
Rm-Rf, SMB	5	**6.6248**	-351.7284	-27421.36	580.49	**0.2071120**
Rm-Rf, HML	32	33.6248	-351.7803	-27426.28	421.04	0.2347538
Rm-Rf, UMD	33	34.6248	-350.7663	-27329.96	**231.76**	0.2507485
SMB, HML	57	58.6248	-348.8321	-27146.21	26150.12	0.2529516
SMB, UMD	66	67.6248	-347.8230	-27050.34	31049.13	0.2702243
HML, UMD	61	62.6248	-348.0539	-27072.28	25390.25	0.2758726
Rm-Rf, SMB, HML	6	8.4372	**-352.2783**	-27300.49	632.88	0.2095658
Rm-Rf, SMB, UMD	6	8.4372	-351.4926	-27225.85	607.40	0.2100846
Rm-Rf, HML, UMD	7	9.4372	-351.5913	-27235.22	449.09	0.2359544
SMB, HML, UMD	36	38.4372	-348.5143	-26942.91	25434.83	0.2538530
Rm-Rf, SMB, HML, UMD	7	10.2496	-352.0831	-27108.84	653.6978	0.2123541

Table 5.2: Number of latent factors suggested by different criteria

ER (no zero)	BIC_3 (no zero)	SC1	SC2	BIC_3	ER
3	1	4	3	0	0

Table 5.3: Sparsity levels and sparsity criterion after each number of factors extracted, assuming that all factors are strong.

Number of factors	$\tilde{m}(\Sigma_u^k))$	$\tilde{m}(\Sigma_u^k) + {}^{kg(N)}/_{10}$
0	61	61
1	5	5.8124
2	3	**4.6248**
3	3	5.4372
4	3	6.2496
5	1	5.0620
6	1	5.8744
7	1	6.6868
8	1	7.4992
9	1	8.3116
10	1	9.1240

5.3 Remarks

In this chapter, I propose to use adaptive thresholding procedure directly applied to the covariance matrix of idiosyncratic errors. However, unlike the standard use of thresholding to provide the estimated covariance matrix, I only use thresholding to estimate the level of sparsity of the true covariance matrix. Ultimately, estimating the level of sparsity of the idiosyncratic errors covariance matrix is very useful for constructing the SC value, which in a way measures the goodness of factor models. Based on the empirical results, it can be seen that the SC provides informative values which can be used to compare the observed and unobserved models.

However, one challenge in using SC is that estimating the level of sparsity requires good estimation for the value of C. For comparing factor models, choosing a right value of C needs to be careful. A relatively small value of C does not make enough entries to zero, and the large value of C forces everything to zero. In both case it is hard to differentiate the level of sparsity between two thresholded covariance matrices. Therefore, I have to use the data-driven method for selecting C in this Chapter, which takes considerably longer time than traditional model selection methods such as AIC or BIC.

6 Concluding remarks and further directions

The research pursued in this thesis is mainly regarding to estimating the factor models in the case where not all factors are strongly pervasive. As discussed in several applications, factor model is a powerful empirical tool in Economics and Finance, and therefore our findings can be a useful contribution to the current rich literature.

6.1 The findings of the thesis

In Chapter 2 it is shown that under some regularity condition, the factors space can still be consistently estimated if the weakest factor has up to a certain strength. As we can see, in Theorem 2.1, the convergence rate is significantly affected when the pervasiveness condition is relaxed up to a level $d_r(N)$. Particularly, this rate is $\frac{N}{[d_r(N)]^2} + \frac{N^2}{T[d_r(N)]^2}$ so in order to identify the factors, we need $\sqrt{N} = o(d_r(N))$ and $N/d_r(N) = o(\sqrt{T})$. In addition, due to some lemmas used in the proof of theorem 2.1, we also require $N\sqrt{\log N}/d_r(N) = o(\sqrt{T})$, hence the lower bound for $d_r(N)$ is $\max(\sqrt{N}, N\sqrt{\log N/T})$. When T is as large as N, this is approximately $\sqrt{N \log N}$. Therefore, if $d_r(N)$ achieve the rate N^α for some $\alpha \in (\frac{1}{2}, 1)$, the consistency of sample PCs as estimators for the factors space can be assured, although clearly stronger factors are easier to identify.

In addition, in Chapter 3 I propose a new way to select the number of factors,

which can work when the factors are weak. Based on simulations, it is recommended to use $SC1$ and $SC2$ when we think the factors can have various strengths. Monte Carlo simulations verify the performance of this criterion under weak factor model, however in some cases the performances are not always stable. In the near future, more study regarding this class of sparsity criterion will be pursued.

One of the direct consequences of our findings is the consistency of the POET covariance matrix estimator, even when the underlying factors model is not as strong as originally assumed. Moreover, it is shown that the number of factors (or orthogonal PCs) should not play a crucial role in the POET. However, estimating more factors than what is required may lose efficiency of the covariance matrix estimated.

A final contribution in this paper is factor model selection, in which a common problem of which factor model to choose is tackled. This should be useful in practice, because once factor model is considered as a successful tool, we will often face the difficulty of choosing a best model: observed factors or unobserved factors. It is also straightforward to apply to the case where one can have a mix model between observed and unobserved factors, in which the key component in thresholding function should be used as in the unobserved case.

6.2 Future research

A first extension to chapter 2 would be to examine the asymptotic distribution of the factors estimated in the weak model. Furthermore, the performance of the sparsity criterion is sensitive to the practical choices of the thresholding parameters and penalty functions, which are only theoretically shown to satisfied some conditions. Therefore, more practical way of choosing these values in order to improve the performance of the sparsity criterion should be further studied.

One of the important applications of factor model in Economics is the factor-augmented model, which is also called forecasting with diffusion indexes. For such model, an extra level of convergence needs to be derived for the estimated coefficients

in the main regressions equation. The results for the case of strong factor model is well established, but an adaption to our case needs to be developed in the future.

Similarly to the weak factor model is a sparse VAR model, because in both models some variables typically on the right-hand side do not affect the majority of cross-sectional components. However, I focus on the case where the cross-sectional components may not be clusterised into uncorrelated subgroups and we have to extract factors directly from the original data. The consistency of the factors and loadings estimated by principle components (PCs) then depends on the strength of the weakest factor. It is worth noting here that after the factors are identified (under certain restrictions), we can apply the LASSO method for estimating the loadings to exactly identify the zero cases. This can be a promising area, as one need to show the convergence of the LASSO estimated loadings to the true space. This is unlikely to be straightforward because of the high dimension and the weakly pervasiveness of the factors.

In the other case of the where it is possible to segmentate all the cross-section components into regions based on their dependence structure, a weak factor can be interpreted as a regional factor. A possibly better approach in here is to the extract the regional factor from each region, instead of from the whole original data. If natural segmentation in practice exists such as industries in the market or regions in the global economy, empirical work can be done to examine to improve the factors estimated from each region, comparing to from the whole data set. More important, the clear next step is to find an automated way for segmentation and work out the rate of convergence of the regional factors estimated from the estimated regions.

Bibliography

[1] Alessi, L., Barigozzi, M., & Capasso, M. (2007). A robust criterion for determining the number of static factors in approximate factor models (No. 2007/19). LEM Working Paper Series.

[2] Anderson, T. W. (1963). The use of factor analysis in the statistical analysis of multiple time series. Psychometrika, 28(1), 1-25.

[3] Ahn, S. C., & Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. Econometrica, 81(3), 1203-1227.

[4] Amengual, D., & Watson, M. W. (2007). Consistent estimation of the number of dynamic factors in a large N and T panel. Journal of Business & Economic Statistics, 25(1), 91-96.

[5] Bai, J. (2003). Inferential theory for factor models of large dimensions. Econometrica, 71(1), 135-171.

[6] Bai, J., & Liao, Y. (2013). Generalized Principal Components for Panel Data and Factor Models. arXiv preprint arXiv:1307.2662.

[7] Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. Econometrica, 70(1), 191-221.

[8] Bai, J., & Ng, S. (2006). Evaluating latent and observed factors in macroeconomics and finance. Journal of Econometrics, 131(1), 507-537.

[9] Bai, J., & Ng, S. (2007). Determining the number of primitive shocks in factor models. Journal of Business & Economic Statistics, 25(1).

[10] Bai, J., & Ng, S. (2013). Principal components estimation and identification of static factors. Journal of Econometrics, 176(1), 18-29.

[11] Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. Annals of Statistics 36: 2577-2604.

[12] Boivin, J., & Ng, S. (2006). Are more data always better for factor analysis?. Journal of Econometrics, 132(1), 169-194.

[13] Boivin, J., Giannoni, M. P., & Stevanovi, D. (2013). Dynamic effects of credit shocks in a data-rich environment. Manuscript

[14] Breitung, J., & Pigorsch, U. (2013). A canonical correlation approach for selecting the number of dynamic factors. Oxford Bulletin of Economics and Statistics, 75(1), 23-36.

[15] Brillinger, D. R. (1981). Time Series: Data Analysis and Theory, 2nd ed. Holden-Day, Oakland, CA.

[16] Cai, T., & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. Journal of the American Statistical Association, 106(494), 672-684.

[17] Cattell, R. B. (1966). The scree test for the number of factors. Multivariate behavioral research, 1(2), 245-276.

[18] Chen, Y. H., & Härdle, W. K. (2012). Common factors in credit defaults swaps markets (No. 2012-063). SFB 649 Discussion Paper.

[19] Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. Econometrica, 51, 1305–1324.

[20] Choi, I. (2012). Efficient estimation of factor models. Econometric Theory, 28(2), 274.

[21] Choi, I. and Jeong, H.(2013). Model Selection for Factor Analysis: Some New Criteria and Performance Comparisons. Working Papers 1209, Research Institute for Market Economy, Sogang University

[22] Diebold, F. X., & Li, C. (2006). Forecasting the term structure of government bond yields. Journal of econometrics, 130(2), 337-364.

[23] Diebold, F. X., Rudebusch, G. D., & Boragan Aruoba, S. (2006). The macroeconomy and the yield curve: a dynamic latent factor approach. Journal of econometrics, 131(1), 309-338.

[24] Elton, E. J., Gruber, M. J., Agrawal, D., & Mann, C. (2001). Explaining the rate spread on corporate bonds. The Journal of Finance, 56(1), 247-277.

[25] Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of financial economics, 33(1), 3-56.

[26] Fan, J., Liao, Y., & Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. Annals of statistics, 39(6), 3320.

[27] Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75: 603–680.

[28] Gilchrist, S., Yankov, V., & Zakrajšek, E. (2009). Credit market shocks and economic fluctuations: Evidence from corporate bond and stock markets. Journal of Monetary Economics, 56(4), 471-493.

[29] Gregory, A. W., Head, A. C., & Raynauld, J. (1997). Measuring world business cycles. International Economic Review, 677-701.

[30] Heaton, C. (2008). Factor analysis of high dimensional time series. PhD thesis (University of New South Wales).

[31] Lam, C., & Yao, Q. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. The Annals of Statistics, 40(2), 694-726.

[32] Lam, C., Yao, Q., & Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. Biometrika, 98(4), 901-918.

[33] Lawley, D. N., & Maxwell, A. E. (1971). Factor analysis as a statistical method.

[34] Ledoit, O., & Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. The Journal of Portfolio Management, 30(4), 110-119.

[35] Lin, Z., Ganesh, A., Wright, J., Wu, L., Chen, M. I. N. M. I. N. G., & Ma, Y. (2009). Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 61.

[36] Mallows, C.L. (1973). Some comments on Cp. Technometrics, 15, pp. 661-675.

[37] Merlevède, F., Peligrad, M., & Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. Probability theory and related fields, 151(3-4), 435-474.

[38] Onatski, A. (2009). Testing hypotheses about the number of factors in large factor models, Econometrica, 77, 1447 – 1479.

[39] Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. The Review of Economics and Statistics, 92(4), 1004-1016.

[40] Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. Journal of Econometrics, 168(2), 244-258.

[41] Pan, J., & Yao, Q. (2008). Modelling multiple time series via common factors. Biometrika, 95(2), 365-379.

[42] Peña, D., & Box, G. E. (1987). Identifying a simplifying structure in time series. Journal of the American Statistical Association, 82(399), 836-843.

[43] Priestley, M., Rao, T., & Tong, H. (1974). Applications of principal component analysis and factor analysis in the identification of multivariable systems. Automatic Control, IEEE Transactions on, 19(6), 730-734.

[44] Ross, S. A. (1976). The arbitrage theory of capital asset pricing. Journal of economic theory, 13(3), 341-360.

[45] Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. Journal of the American Statistical Association, 104, 177–186.

[46] Schneeweiss, H. (1997). Factors and principal components in the near spherical case. Multivariate Behavioral Research, 32(4), 375-401.

[47] Sentana, E. (2009). The econometrics of mean-variance efficiency tests: a survey. The Econometrics Journal, 12(3), C65-C101.

[48] Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The journal of finance, 19(3), 425-442.

[49] Sparks, R.S., Coutsourides, D. and Troskie, L. (1983). The multivariate Cp. Communications in Statistics-Theory and Methods, 7, pp. 13-26.

[50] Spearman, C. (1904). " General Intelligence," Objectively Determined and Measured. The American Journal of Psychology, 15(2), 201-292.

[51] Stock, J. H., & Watson, M. W. (1998). Diffusion indexes (No. w6702). National bureau of economic research.

[52] Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. Journal of the American statistical association, 97(460), 1167-1179.

[53] Wang, P. (2008). Large dimensional factor models with a multi-level factor structure: identification, estimation and inference. Unpublished manuscript, New York University. Chicago

[54] Wright, J., Ganesh, A., Rao, S., Peng, Y., & Ma, Y. (2009). Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In Advances in neural information processing systems (pp. 2080-2088).

[55] Yu, Y., & Samworth, R. J. (2013). Discussion of Large Covariance Estimation by Thresholding Principal Orthogonal Complements by Fan, Liao and Mincheva.