Inferential Theory for Factor Models of Large Dimensions under Monotone-Missing Data

Ercument Cahan

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Eric Zivot, Chair

Thomas Gilbert

Ji Hyung Lee

Program Authorized to Offer Degree:
Economics

University of Washington

Abstract

Inferential Theory for Factor Models of Large Dimensions under Monotone-Missing Data

Ercument Cahan

Chair of the Supervisory Committee:
Title of Chair Eric Zivot
Department of Chair

In this dissertation we investigate the inferential theory for factor models with large cross-section (N) and time series (T) dimensions under monotone-missing data. The major contribution of the dissertation is the development and testing of an intuitive and parsimonious factor-based imputation (FBI) algorithm that preserves the desirable asymptotic properties of the standard (complete-data) factor models. FBI uses the principal component (PC) estimator obtained from balanced subpanels of the data set to fill in the missing values. Well-behaved asymptotic properties of the factor model estimators obtained from FBI-imputed data sets would decrease the need for truncation, and make it possible to enjoy larger data sets and to carry out factor model inference with more confidence. We also provide the asymptotic distribution of the imputation error born out of FBI. In addition, we compare the small sample performance of FBI with that of Expectation-Maximization (EM) algorithm and show that FBI outperforms EM in every measure we consider.

Recent advances in information technology allowed researchers to access myriad of economic time series over an increasingly long span and at a reasonable cost. While the increase in the availability of data made it possible to test and understand the economic phenomena better, it also led to the problem of organizing the data in an easy to interpret form. Factor analysis, being a useful method for summarizing the information in data-rich environments, has received increasing attention, and the econometric analysis of large dimensional factor models has become a heavily researched topic.Recently, factor models have received particular attention in the macroeconomic forecast literature.

New generation approximate factor models allows the number of observations to be large in both cross-section and time series dimensions. Stock and Watson (2002b) showed that the principal components are consistent estimators of the true latent factors when both N and T approach infinity without imposing any restriction on their relative rates of increase. Bai (2003) proved that estimated factors are consistent and in general asymptotically normal in presence of serial correlation and het-

eroskedasticity, while Bai and Ng (2002) studied the consistent estimation of number of factors under large N and T assumption.

A method of rising interest is to estimate common factors by principal components from large data panels, and then to augment an otherwise standard regression with estimated factors. Stock and Watson (2002a) showed that the feasible forecasts constructed from the estimated factors together with the estimated coefficients converge to the infeasible forecast that would be obtained if the factors and coefficients were known. Bai and Ng (2006) determined the limiting distributions of forecast errors and least squares estimates obtained from factor augmented regressions. They showed that least squares estimates are $\sqrt{T}$-consistent and asymptotically normal if $\sqrt{T}/N \to \infty$.

The studies cited above obtained their results under the assumption that the data panel from which factors are estimated is *balanced*, i.e. it does not suffer from any missingness. However, in practice most data panels are unbalanced. When missing data is present, factor model estimates cannot be obtained directly. In order to estimate them, one should first transform the unbalanced panel at hand to a balanced one. This is achieved by either truncation or imputation. When using a truncated data panel one can rely on standard large sample theory since the resulting data set is balanced and contains no estimation error. However, the remaining data after truncation may not be representative for the entire population (e.g. survivalship bias). Therefore, small sample performance of truncated data sets may become questionable. On the other hand, when missing data is imputed standard large sample theory is no longer valid if the imputation algorithm does not take into account the estimation error incurred during imputation. If the level of missingness (i.e the number of missing cells in the data panel) is kept fixed while the dimensions of the data panel are allowed to increase indefinitely, the effect of missingness on estimation eventually dies out, and standard large sample theory applies as in truncation case. To prevent this bias, large sample theory for factor models that accounts for the missing data should allow missingness to grow indefinitely together with N and T, and determine the asymptotic properties of the factor model estimators. This is the path we take in this study: We consider the large sample properties of factor model estimators extracted from monotone-missing data sets that are imputed with the FBI algorithm. We refer to these estimators as the imputed data (ID) estimators. To our knowledge, our research is the first to focus on the asymptotic properties of the factor model estimators obtained with PC from large data sets that have considerable missing data problem.

The organization the dissertation is as follows: Chapter I starts with introducing the large dimensional factor models, their estimation, forecast and standard large sample theory. Then, it discusses

the missing data mechanisms. Next, we briefly introduce the FBI algorithm and discuss how it exploits the balanced subpanels of the unbalanced data panel to impute the missing values. We refer to the factor model estimators obtained from imputed balanced data sets as *imputed data* (ID) estimators. We show that ID estimators are consistent and aymptotically normal via an extensive Monte Carlo study. This indicates that FBI algorithm preserves the desirable properties of consistency and asymptotic normality of the factor model estimators obtained under complete data. We also consider the large sample behavior of different partitions of the factor estimator. We find that partitioned factor estimators that are exposed to missingness more converge slower and have a higher asymptotic variance.

In Chapter II, we focus on the workings and main statistical properties of FBI algorithm. First, we study the statistical properties of the auxiliary (interim) factor estimators and show that they are consistent and asymptotically normal. These results are instrumental in determining the large sample distribution of the imputation error. Then, we derives the asymptotic distribution of imputation error under FBI, and show that it is consistent and asymptotically normal. Finally, we express the partitioned CID factor estimators in terms of the observed and imputed components of completed data set $\hat{X}$. This analysis reveal that differences in convergence rate and asymptotic variances across different partitioned estimators found in Chapter 1 can be explained with partitions' exposure to missingness. That is, the higher the cross section missingness in a partition, the slower the converge and the bigger the asymptotic variance and vice versa.

Chapter III chapter serves mainly two purposes: (i) characterization of correlation structure under factor model, (i) how the correlation structure affects the relative performance of FBI and Expectation-Maximization algorithms in small sample. To this end, we first establish the relation between factor and correlation structures. We start with a very general factor model under very weak assumptions and determine the analytical form of absolute correlation coefficients in terms of factor model components. Then, assuming that all factor variances are equal we derive the exact probability density function of absolute pairwise correlations. Next, we propose average absolute correlation (denoted by $\mu_{|\rho|}$) as a summary statistic measuring linear comovement among the series in the data set. Utilizing the derived absolute correlation density, we show that there is a negative relation between the number of factors and the average absolute correlation (denoted by $\mu_{|\rho|}$) among the series that admit a factor representation. This is a key finding for imputation purposes since all imputation methods exploit the correlation structure of the data set in one way or another. In the second part of the chapter, we compare the small sample performance of FBI and EM algorithms from various perspectives. To

this end, we develop various performance metrics for that measure imputation accuracy from different dimensions. Using these measures, we show that FBI outperforms EM algorithm under many different scenarios. We also show that FBI is more robust to higher r (lower $\mu_{|\rho|}$) values than EM and should be preferred for data sets admitting a factor structure even if data is multivariate normal.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I am extremely grateful to my dissertation committee, Eric Zivot, R. Douglas Martin, Michael D. Perlman, Thomas Gilbert and Ji Hyung Lee for their advice and encouragement. I am especially thankful to graduate classes that I took from Eric Zivot and R. Douglas Martin who have been tremendous models for me as teachers and researchers. I am grateful to Jushan Bai for his great suggestions and guidance, and Serena Ng for letting me follow her invaluable econometrics classes. I would like to thank Charles R. Nelson and Dick Startz fo their very helpful comments and suggestions.

I am thankful to Sridhar Gollamudi, Sebastian Fossati and Sebastian Pueblas for their constuctive critique that provided my research with new perspectives. I appreciate my conversations with Joel Seabold and Gregory Teplow which have contributed a lot to my dissertation. In addition, I would like to thank my managers and colleagues at Bloomberg Quant Team for their enormous support throughout my dissertation journey. I am also thankful to research seminar participants in Platinum Grove Asset Management, Allianz GIC and Bloomberg LP for their comments and suggestions.

Finally, I would like to thank Gul Ipek Tunc and Alper Guzel for their encouragement and trust in me during my switch from political science to economics, and Muhiddin Uguz for letting me follow his great advanced calculus classes.

DEDICATION

to my dear mom who beat cancer twice...

...and to my lovely wife who has been with me in every step of the way

1

A FACTOR-BASED IMPUTATION ALGORITHM AND THE ASYMPTOTIC BEHAVIOR OF ESTIMATORS IN LARGE DIMENSIONAL FACTOR MODELS WITH MONOTONE MISSING DATA

1.1 Introduction

Recent advances in information technology allowed researchers to access myriad of economic time series over an increasingly long span and at a reasonable cost. While the increase in the availability of data made it possible to test and understand the economic phenomena better, it also led to the problem of organizing the data in an easy to interpret form. Factor analysis, being a useful method for summarizing the information in data-rich environments, has received increasing attention, and the econometric analysis of large dimensional factor models has become a heavily researched topic in recent macroeconomics literature.

Let N and T be the cross-section and time series dimensions of a data set X, respectively. Chamberlain and Rothschild (1983) considered an approximate factor model, which weakens the classical factor model assumption that correlations between idiosyncratic errors are zero and showed that principal components method is equivalent to factor analysis as N goes to infinity. Connor and Korajczyk (1986) extended their results and showed that when N is much larger than T, factor model can be estimated by applying the principal components method to the $T \times T$ covariance matrix. They also proved that the estimated factors are consistent when T is fixed.

New generation of approximate factor models allowed the number of observations to be large in both cross-section and time series dimensions. Stock and Watson (2002a) showed that the principal components are consistent estimators of the true latent factors when both N and T approach infinity without imposing any restriction on their relative rates of increase. Bai (2003) proved that estimated factors are consistent and in general asymptotically normal in presence of serial correlation and heteroskedasticity, while Bai and Ng (2002) studied the consistent estimation of number of factors under large N and T assumption.

The studies cited above obtain their results under the assumption that the data panel from which factors are estimated is *balanced*, i.e. it does not suffer from missing data. However, in practice most data panels are unbalanced. When missing data is present factor estimates cannot be obtained directly.

In order to estimate them, one should first transform the unbalanced panel at hand to a balanced one.

To this end, in the first part of this chapter we develop a parsimonious, intuitive and easy-to-implement factor-based imputation (FBI) algorithm aimed for data sets that admit a factor structure and suffer from monotone missingness. Following Stock and Watson (2002a) and Bai (2003), we employ the principal component estimator since it is easy to calculate and asymptotically equivalent to maximum likelihood estimators. Basically, FBI exploits the balanced sub-blocks of the unbalanced data set to obtain the common component estimates of the missing values. Note that the central promise of FBI is to utilize the information embedded the factors structure of the data set efficiently, which a non-factor based algorithm (e.g. widely used unconditional mean imputation or Expectation-Maximization algorithm under the assumption of multivariate normality) would not fully exploit.

After the implementation of FBI, one obtains the completed data set, $\hat{X}$. In the second part of this chapter, we analyze the asymptotic properties of the factor estimators (which we refer to as *imputed data* (ID) estimators) obtained from $\hat{X}$. We mainly investigate whether the nice asymptotic properties (namely consistency and asymptotic normality) of the factor model estimators[1] still hold under FBI algorithm. To our knowledge, our research is the first to focus on the asymptotic properties of factor estimators obtained with the method of principal components (PC) from large data sets that have considerable missing data problem. To this end, we first characterize the conditions under which the large sample properties of ID estimators should be studied. In a nutshell, if the level of missingness is kept fixed while the dimensions of the data panel are allowed to increase indefinitely, the effect of missingness on estimation eventually dies out and the resulting asymptotic results cannot be used for inference purposes under missing data. With this intuition in mind, though various simulation experiments we show that the large sample theory for factor models with completed data sets should allow missingness to grow at the same rate with N and T.

Utilizing this approach we find that unless there is factor structure shift, ID estimators are consistent and asymptotically normal just like their standard complete-data counterparts. Next, we show that the rate of convergence of the ID estimators are slower and the asymptotic variance of the ID estimators is larger compared to the complete-data estimators. We also find that asymptotic variance of the ID estimators increase with the level of missingness. Finally, we discover that portions of ID estimators that expose to different level of missingness behave differently even if those of their true counterparts behave the same.

---

[1]Throughout this dissertation, factor model estimators refer to factor, loading and common component estimators of a factor model.

The organization of the paper is as follows: Section 1.2 introduces factor models under complete data, discusses the major findings in recent large dimensional factor model literature and briefly studies their estimation via the method of PC. The problem of missing data and its possible remedies, missingness patterns and mechanisms are studied in Section 1.4. Factor models under missing data and our factor-based imputation method are considered in Section 1.5. Section 1.6 carries out an extensive simulation experiment to find the right approach in studying the asymptotics of ID estimators, and to test whether FBI preserves the useful asymptotic features of the standard complete-data factor model estimators. Results are evaluated in Section 1.7. Finally, concluding remarks are provided in Section 1.8.

1.2 Factor Models under Complete Data

Factor models mainly decompose a series into common and idiosyncratic components. A *static* factor model under complete data has the following representation [2]

$$X_{it} = \lambda_i' F_t + e_{it} \tag{1.1}$$

where X_{it} is the i^{th} cross-section unit at time t, for $i = 1, .., N$ and $t = 1, ..., T$, F_t is an $r \times 1$ vector of common factors, λ_i is a $r \times 1$ vector of factor loadings (sensitivities), and e_{it} is the idiosyncratic (specific) component of X_{it}. In Equation (1.1), the term $\lambda_i' F_t$ is called the common component of X_{it}. On the other hand, a dynamic factor model allows for dynamic loadings and is defined as

$$X_{it} = \lambda_i(L) F_t + e_{it}$$

where $\lambda_i(L) = (1 - \lambda_{i1}L - \lambda_{i2}L^2 ... - \lambda_{ip}L^p)$ is vector of dynamic loadings of order p. Dynamic models allow lags of the factors enter the process, and are commonly used in macroeconomic forecasting literature. They are typically estimated by eigenvalue decomposition of the spectral density matrix. In addition, every dynamic factor can be expressed as static model with $r(p+1)$ factors. In this paper, we solely consider the static factor models.

In matrix form, a static factor model can be defined as

$$X = F\Lambda' + e \tag{1.2}$$

where $X = (X_1, X_2, ..., X_T)'$, $F = (F_1, F_2, ..., F_T)'$, $\Lambda = (\lambda_1, \lambda_2, ..., \lambda_N)'$, and $e = (e_1, e_2, ..., e_T)'$.

[2] Throughout this paper we will follow the notation developed in Bai (2003)

X can also be expressed in vector form. Depending on the dimension we consider, there are two different ways to do so: The t^{th} cross-section sample of X is given by

$$X_t = \Lambda F_t + e_t \tag{1.3}$$

where for $t = 1, 2, ..., T$. Similarly the i^{th} variable (column) can be expressed as

$$\underline{X}_i = F\lambda_i + \underline{e}_i \tag{1.4}$$

where $\underline{X}_i$ and $\underline{e}_i$ are the i^{th} column of X and e, respectively.

Assuming F and e are uncorrelated and have zero mean and Λ is non-stochastic, the following covariance structure is obtained

$$\Sigma = \Lambda\Sigma_F\Lambda' + \Omega \tag{1.5}$$

where Σ, Σ_F and Ω are the population covariance matrices of X_t, F_t and e_t. Strict (classical) factor model assumes that idiosyncratic errors are uncorrelated (i.e. Ω is diagonal) and the time series dimension is larger than the cross-section dimension. However, classical factor analysis cannot consistently estimate the common factors, which in general are of direct interest in applications. On the other hand, *approximate* factor models allow weak cross-section and serial correlation in the idiosyncratic component. In addition, under large N and T assumptions, factors and loadings in approximate factor models can be consistently estimated, and their limiting distributions can be obtained.

Factor models arise often in economics and finance. For example, in economics X_{it} can be a macroeconomic series such as GDP or unemployment while F_t is a vector of common drivers such as business cycle, and λ_i is the sensitivity of series i to common drivers. Likewise, in finance X_{it} can be the return for asset i at time t while F_t is the fundamental firm characteristics (factor returns) such as value or momentum, and e_{it} is the idiosyncratic returns. Although strict factor models may be applicable in certain areas (psychology, physics, climate studies etc.), its assumptions are unlikely to be satisfied in most macroeconomic and financial applications.

1.2.1 Estimation

Factors, loadings and idiosyncratic errors in a factor model are all unobserved and need to be estimated. When N is small, factor models can be expressed in state space form, normality is assumed, and parameters are estimated by maximum likelihood (ML).

On the other hand, for large data sets computational difficulties prevent the use of ML, and the method of principal components (PC) arises as a computationally feasible and desirable alternative.

Method of PC is a non-parametric method that does not require any distributional assumptions. The number of factors that can be estimated by method of PC is $min(N,T)$, which is much higher than those that can be estimated by employing state space models using a parametric estimation method. In addition, PC estimators are asymptotically equivalent to ML estimators. Since our focus is on the behavior of estimated factors and loadings in factor models of large dimensions, in this paper we consider the PC estimator. Consider

$$X = F^{\star}\Lambda^{\star\prime} + e \tag{1.6}$$

where $F^{\star} = FC$, $\Lambda^{\star} = \Lambda C^{-1\prime}$ and C is some $r \times r$ nonsingular matrix. As can be seen (1.6) is observationally equivalent to (1.5). Therefore F and Λ cannot be identified separately, and r^2 restrictions are required to uniquely fix them. The normalizations $F'F/T = I$ and $\Lambda'\Lambda$ being diagonal provides these r^2 restrictions. Alternatively, one can consider the normalizations $\Lambda'\Lambda/N = I$ and $F'F$ being diagonal in order to satisfy the required r^2 restrictions. In both cases, factors and loadings can be uniquely (still up to a sign change) identified.

Assuming there are k factors (k may or may not be equal to r), factor and loading estimates are obtained by solving the following maximization problem

$$\min_{\Lambda^k, F^k} S(k) = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it} - \lambda_i^k F_t^k)^2 \tag{1.7}$$

subject to one of the normalizations on F and Λ discussed above. If $N < T$, we concentrate out F^k. Then the loading estimate is found as $\bar{\Lambda}^k = \sqrt{N}Z$ where Z is the eigenvectors corresponding to the k largest eigenvectors of the $N \times N$ matrix $X'X$. $\bar{F}^k$ is found by simply regressing X on $\bar{\Lambda}^k$, as $\bar{F}^k = X\bar{\Lambda}^k/N$. Common component estimate is obtained simply as $\bar{C} = \bar{F}^k\bar{\Lambda}^{k\prime}$.

On the other hand, if $T < N$, we concentrate out Λ^k and minimizing $S(k)$ becomes identical to maximizing $tr(F^{k\prime}(X'X)F^k)$. In this case, we obtain $\tilde{F}^k = \sqrt{T}Z$ where Z is the eigenvectors corresponding to the k largest eigenvectors of the $T \times T$ matrix XX', and $\tilde{\Lambda}^k = X^{'}\tilde{F}^k/T$. Finally, the common component estimate is obtained simply as $\tilde{C} = \tilde{F}^k\tilde{\Lambda}^{k\prime}$.

Bai and Ng (2008) shows that both $\bar{F}^k$ and $\tilde{F}^k$ estimate the same factor space by showing the following identity

$$\bar{F}^k = \tilde{F}^k(V^k)^{1/2}$$

where V^k denotes the $k \times k$ diagonal matrix consisting of the first k largest eigenvalues of the matrix $XX^{'}/NT$, arranged in decreasing order. Therefore, each column of $\bar{F^k}$ is a scalar multiple of

the corresponding column[3] in $\tilde{F^k}$.

In the rest of this study for both $N < T$ and $N > T$ cases we use a unified notation and denote factor, loading and common component estimators obtained from complete data sets with $\tilde{F}$, $\tilde{\Lambda}$ and $\tilde{C}$, respectively. Which estimator group above we refer to should be obvious from the relative magnitudes of N and T.

When data have missing values, factor model estimators above cannot be obtained directly. In this sense, in missing a data setting we refer to $\tilde{F}$, $\tilde{\Lambda}$ and $\tilde{C}$ as infeasible estimators.

1.2.2 Large Sample Theory

Stock and Watson (2002a) studies the large sample properties of estimated factors and loadings. It shows that estimated factors and loadings consistently estimate the factor and loading spaces, respectively, and provides their convergence rates as follows

Theorem SW1[4]: Suppose that k factors are extracted from X via PCA, where k maybe $\leq$ or $> r$. Let S_i denote a variable with a value of ± 1 and $N, T \to \infty$, then the following holds:

$$\frac{1}{T}\sum_{i=1}^{T}(S_i\tilde{F}_{it} - F_{it})^2 \xrightarrow{p} 0 \quad (for \ \text{i=1,...,r}) \tag{1.8a}$$

$$S_i\tilde{F}_{it} \xrightarrow{p} F_{it} \quad (for \ \text{i=1,...,r}) \tag{1.8b}$$

$$\frac{1}{T}\sum_{i=1}^{T}\tilde{F}_{it}^2 \xrightarrow{p} 0 \quad (for \ \text{i=r+1,...,k}) \tag{1.8c}$$

Bai (2003) goes beyond the above consistency result and provides the conditions for the asymptotic normality of the factor model estimators under very general conditions that allow for correlations and heteroskedasticity in both time series and cross-section dimensions. Below we give the main assumptions[5] and theorems in Bai (2003) without proof.

ASSUMPTION A - Factors: $E\,||F_t||^4 \leq M < \infty$ and $T^{-1}\sum_{t=1}^{T} F_t F_t^{'} \xrightarrow{p} \Sigma_F$ for some $r \times r$ positive definite matrix Σ_F.

---

[3]For the rest of this paper, we drop the superscript k used on factor and loading estimates.

[4]Throughout the text, theorems and proposals with letter 'SW', 'B' and 'BN' would refer to the theorems in Stock and Watson (2002a), Bai (2003) and Bai and Ng (2006), respectively.

[5]We will refer to these assumptions in Chapter II while deriving the asymptotic properties of the imputation error under FBI.

ASSUMPTION B - Factor Loadings: $||\lambda_i|| \leq \bar{\lambda} < \infty$, and $\left|\left|\Lambda'\Lambda/N - \Sigma_\Lambda\right|\right| \to 0$ for some $r \times r$ positive definite matrix Σ_Λ.

ASSUMPTION C - Time and Cross-Section Dependence and Heteroskedasticity: There exist a positive constant $M < \infty$ such that for all N and T we have:

i. $E(e_{it}) = 0$, $E|e_{it}|^8 \leq M$

ii. $E(e_s' e_t/N) = E(N^{-1}\sum_{i=1}^{N} e_{is}e_{it}) = \gamma_N(s,t)$, $|\gamma_N(s,s)| \leq M$ for all s, and

$$T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}|\gamma_N(s,t)| \leq M$$

iii. $E(e_{it}e_{jt}) = \tau_{ij,t}$ with $|\tau_{ij,t}| \leq \tau_{ij}$ for some τ_{ij} and for all t. In addition

$$N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}|\tau_{ij}| \leq M$$

iv. $E(e_{it}e_{js}) = \tau_{ij,ts}$ and $(NT)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{s=1}^{T}\sum_{t=1}^{T}|\tau_{ij,ts}| \leq M$

v. For every (t,s), $E|N^{-1/2}\sum_{i=1}^{N}[e_{is}e_{it} - E(e_{is}e_{it})]|^4 \leq M$

ASSUMPTION D - Weak dependence between factors and idiosyncratic errors:

$$E\left(\frac{1}{N}\sum_{i=1}^{N}\left|\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t e_{it}\right|\right|\right) \leq M$$

ASSUMPTION E - Weak Dependence: There exist $M < \infty$ such that for all T and N, and for every $t \leq T$ and every $i \leq N$:

i. $\sum_{s=1}^{T}|\gamma_N(s,t)| \leq M$

ii. $\sum_{k=1}^{N}|\tau_{ki}| \leq M$

ASSUMPTION F- Moments and Central Limit Theorem: There exist an $M < \infty$ such that for all N and T we have:

i. for each t,

$$E\left|\left|\frac{1}{\sqrt{NT}}\sum_{s=1}^{T}\sum_{k=1}^{N}F_s[e_{ks}e_{kt}-E(e_{ks}e_{kt})]\right|\right|^2 \leq M$$

ii. the $r \times r$ matrix satisfies

$$E\left|\left|\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\sum_{k=1}^{N}F_t\lambda_k^{'}e_{kt}\right|\right|^2 \leq M$$

iii. for each t, as $N \to \infty$,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i e_{it} \xrightarrow{d} N(0,\Gamma_t)$$

where $\Gamma_t = lim_{N\to\infty}(1/N)\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j^{'}E(e_{is}e_{it}$

iv. for each i, as $T \to \infty$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t e_{it} \xrightarrow{d} N(0,\Phi_i)$$

where $\Phi_i = plim_{T\to\infty}(1/T)\sum_{s=1}^{T}\sum_{t=1}^{T}E[F_tF_t^{'}e_{is}e_{it})]$

ASSUMPTION G - The eigenvalues of the $r \times r$ matrix $(\Sigma_\Lambda\Sigma_F)$ are distinct.

ASSUMPTION H- λ_i, F_t and e_it are three groups of mutually independent stochastic variables.

Under the assumptions above, Bai (2003) obtains the following results

Theorem B1: If $\sqrt{N}/T \to 0$, then for each t,

$$\sqrt{N}(\tilde{F}_t - H'F_t) \xrightarrow{d} N(0,\Pi_t)$$

where $\Pi_t = V^{-1}Q\Gamma_tQ'V^{-1}$, $V = diag(\nu_1,\nu_2,...\nu_r)$, $\nu_1 > \nu_2 > ...\nu_r$ are the eigenvalues of $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}$ (Σ_Λ and Σ_F are the covariance matrices of Λ and F, respectively.), Γ_t is the asymptotic covariance matrix of sequence $\{\lambda_i e_{it}\}$, and Q is equal to $plim_{T,N\to\infty}\frac{\tilde{F}'F}{T}$.

Theorem B2: If $\sqrt{T}/N \to 0$, then for each t,

$$\sqrt{T}(\tilde{\lambda_i} - H^{-1}\lambda_i) \xrightarrow{d} N(0, \Theta_i)$$

where $\Theta_i = (Q')^{-1}\Phi_i Q^{-1}$, and Φ_i is the asymptotic covariance matrix of sequence $\{F_t e_{it}\}$.

Theorem B3: Let $C_{it} = F_t\lambda_i$ and $\tilde{C}_{it} = \tilde{F}_t\tilde{\lambda}_i$. Then, for each i and t we have,

$$\left(\frac{1}{N}V_{it} + \frac{1}{T}W_{it})^{-1/2}(\tilde{C}_{it} - C_{it}\right) \xrightarrow{d} N(0,1)$$

where $V_{it} = \lambda_i'\Sigma_\Lambda^{-1}\Gamma_t\Sigma_\Lambda^{-1}\lambda_i$ and $W_{it} = F_t'\Sigma_F^{-1}\Phi_i\Sigma_F^{-1}F_t$.

Proposition B1: For any fixed $k \geq 1$, there exists a $r \times k$ matrix H^k with $rank(H^k) = min(k, r)$, and $\delta_{NT} = min(\sqrt{N}, \sqrt{T})$, such that

$$\delta_{NT}^2\left(\frac{1}{T}\sum_{t=1}^{T}\left|\left|\|\tilde{F}_t - H^{k\prime}F_t\|\right|\right|^2\right) = O_p(1)$$

First three results above show that estimated factors, loadings and common components are asymptotically normal, and determine their convergence rates and asymptotic variances. The last one shows that the mean squared deviation between the estimated and true factor spaces vanish at the rate of $min(N, T)$, therefore $\tilde{F}$ consistently estimates the true latent factor space.

1.2.3 Consistent Estimation of Asymptotic Covariance Matrices

Bai (2003) derives consistent estimators for the asymptotic covariance matrices in Theorems B1-B3 in the previous section.

a) Covariance matrix of estimated factors: Γ_t in Theorem B1 above depends on the cross-section correlation of idiosyncratic errors and loadings. Unlike time-series variables, there is not a natural order for cross-section variables. Therefore, a HAC-type (Newey-West) estimator is not feasible. In order to overcome this order problem, Bai (2003) assumes cross-sectional independence for e_{it}. Replacing the factors and loadings in Π_t in Theorem B2 with their estimates, and using $\tilde{F}'\tilde{F}/T = I$, the following is obtained

$$\tilde{\Pi}_t = \tilde{V}^{-1}\tilde{\Gamma}_t\tilde{V}^{-1} \tag{1.9}$$

where $\tilde{\Gamma}_t$ can be described in one of the following depending on the error structure in the factor model:

$$\tilde{\Gamma}_t = \tilde{\sigma_e^2} \frac{1}{N} \sum_{i=1}^{N} \tilde{\lambda}_i \tilde{\lambda}_i' \tag{1.10a}$$

$$\tilde{\Gamma}_t = \frac{1}{N} \sum_{i=1}^{N} \tilde{e}_{it}^2 \tilde{\lambda}_i \tilde{\lambda}_i' \tag{1.10b}$$

$$\tilde{\Gamma}_t = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{\lambda}_i \tilde{\lambda}_j' \frac{1}{T} \sum_{t=1}^{T} \tilde{e}_{it} \tilde{e}_{jt} \tag{1.10c}$$

where $\tilde{e}_{it} = X_{it} - \tilde{\lambda}_i' \tilde{F}_t$ and $n/min[N,T] \to 0$. Above, the estimator in (1.10a) has the strongest assumptions: e_{it} is cross sectionally uncorrelated with e_{jt} and is homoscedastic, that is $E(e_{it}^2) = \sigma_e^2$ for all i and t (σ_e^2 can be consistently estimated by $\tilde{\sigma}_e^2 = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{e}_{it}^2$). (1.10b) weakens the latter assumption in (1.10a) by allowing heteroscedasticity in e_{it}. Finally, (1.10c) has the weakest assumptions on the correlation structure of the error terms by allowing both cross sectional correlation and heteroscedasticity[6]

b) Covariance matrix of estimated loadings: Note that in Theorem B2, the asymptotic covariance matrix of $\tilde{\lambda}_i$, $(\tilde{\Theta}_i)$, is constructed with the series $\tilde{F}_t \tilde{e}_{it}$. Since there is a natural order in this sequence, HAC estimator of Newey-West can be applied as follows

$$\tilde{\Theta}_i = D_{0,i} + \sum_{\nu=1}^{q} \left(1 - \frac{\nu}{q+1}\right) (D_{\nu i} + D_{\nu i}') \tag{1.11}$$

where $D_{\nu i} = (1/T) \sum_{t=\nu+1}^{T} \tilde{F}_t \tilde{e}_{it} \tilde{e}_{it-\nu} \tilde{F}_{t-\nu}'$, and $q \to \infty$ as $T \to \infty$ with $q/T^{1/4} \to 0$. It is important to note that while a HAC estimator based on $F_t e_{it}$ (i.e. true factors and idiosyncratic errors) estimate Φ_i (the middle term in $\tilde{\Theta}_i$), HAC estimator above directly estimates $\tilde{\Theta}_i$, since $\tilde{F}_t$ estimates $H' F_t$.

c) Covariance matrix of estimated common components: Bai (2003) derives a consistent estimator of the common component asymptotic variance in the same spirit with those of factors and loadings as follows:

$$\tilde{V}_{it} = \tilde{\lambda}_i' \left(\frac{\tilde{\Lambda}'\tilde{\Lambda}}{N}\right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \hat{e}_{it}^2 \tilde{\lambda}_i \tilde{\lambda}_i'\right) \left(\frac{\tilde{\Lambda}'\tilde{\Lambda}}{N}\right)^{-1} \tilde{\lambda}_i \tag{1.12}$$

[6]Note that consistency of (1.10c) is shown in Bai and Ng (2006). Due to its strong assumptions, consistency of (1.10c) requires covariance stationarity with $E(e_{it}e_{jt})$.

$$\tilde{W}_{it} = \tilde{F}_t' \tilde{\Theta}_i \tilde{F}_t \tag{1.13}$$

1.3 Inference with Estimated Factors

Consider the following forecast problem

$$y_{t+h} = \alpha_x' X_t + \beta' w_t + \epsilon_{t+h} \tag{1.14}$$

where y_t is a scalar variable to be predicted, $h \geq 0$ is the forecast horizon, X_t is a set of N predictor variables with the structure described in (1.3) and w_t is another (smaller) set of K observed predictors that do not assume a factor structure. Note that in the above forecast equation, the number of predictor series $(N+K)$ can be very large, even larger than the number of available time series observations (T). Let $A = (X : W)$ be the design matrix in (1.14). If $N + K > T$, the design matrix in (1.14) becomes rank deficient (hence singular). As a result, $G'G$ cannot be inverted and the regression coefficients cannot be estimated. On the other hand, if T is barely larger than $N + K$, small sample coefficient estimates becomes highly unstable and their consistency becomes very questionable, and using these coefficients in forecast equation leads to significantly large forecast errors. Additionally, even if T is sufficiently larger than $N + K$, having a better fit in-sample does not necessarily lead to a better out-of-sample result. In general, having a complex model with too many predictors leads to *overfitting* in the in-sample piece; that is the model starts to explain the random noise rather than the underlying relationship between the variables. In this sense, recent availability of new data sets for large number of predictor variables has become both a blessing and curse for researchers in macroeconomics and finance. A method of rising interest in macroeconomic forecast literature is to summarize information embedded in large data panels with common factors estimated by principal components, and then to augment an otherwise standard regression with the estimated factors. Utilizing this method, we can convert the forecast problem above into the following more parsimonious one:

$$y_{t+h} = \alpha' F_t + \beta' w_t + \epsilon_{t+h} \tag{1.15}$$

where F_t is the set of common factors described above. If F_t were observed, (1.15) would be regular linear forecasting model and standard techniques and inferences would apply. Let $g_t = (F_t', w_t')'$. Then the mean-squared optimal prediction of y_t (conditonal mean) is given by

$$y_{T+h|T} = E(y_{T+h}|g_T, g_{T-1}, ...) = \alpha' F_T + \beta' w_T = \delta' g_T$$

However, F_t is not observed and must be extracted from X. Similarly α and β are not known, and must be estimated from (1.15). Therefore, forecast equation (1.15) is not feasible. The feasible equation is obtained by replacing the unknown variables with their estimates as follows

$$\hat{y}_{T+h|T} = \hat{\alpha}' F_T + \hat{\beta}' w_T = \hat{\delta}' \hat{g}_T \tag{1.16}$$

where $\hat{g}_t = (\tilde{F}_t', w_t')'$. Note that in (1.16) $\tilde{F}$ affects $\hat{y}_{T+h|T}$ via two channels: directly via $\tilde{F}_T$, and indirectly via the estimated coefficients $\hat{\alpha}$ and $\hat{\beta}$. Therefore, in order to determine the statistical properties of $\hat{y}_{T+h|T}$, one needs to examine the properties of both the factors and coefficient estimators. Factor estimators and their properties are examined in Section 1.2. Below we briefly consider recent research on the statistical properties of feasible coefficient estimators and conditional mean.

Stock and Watson (2002a) showed that both coefficient estimates and the feasible forecast converge to their infeasible counterparts that would be obtained if the factors were known, and provide the rate of convergence in the following theorem:

Theorem SW2: If $N, T \longrightarrow \infty$

$$\hat{y}_{T+h|T} - y_{T+h|T} \xrightarrow{p} 0 \tag{1.17a}$$

$$\hat{\beta} - \beta \xrightarrow{p} 0 \tag{1.17b}$$

$$S_i\hat{\alpha} - \alpha \xrightarrow{p} 0 \quad (for \ \text{i=1,...,r}) \tag{1.17c}$$

Bai and Ng (2006) extends Stock and Watson (2002a) by determining the limiting distributions of forecast errors and least squares estimates obtained from factor augmented regressions. This work makes it possible to carry out hypothesis testing and generate confidence intervals for factor augmented coefficient estimates and feasible forecasts. Define $\hat{\delta} = (\hat{\alpha}', \hat{\beta}')'$ and $\delta = (\alpha' H^{-1\prime}, \hat{\beta}')'$.

Theorem BN1: If $\sqrt{T}/N \to 0$

$$\sqrt{T}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \Sigma_\delta) \tag{1.18}$$

where $\Sigma_\delta = E'^{-1}\Sigma_{gg}^{-1}\Sigma_{gg,\epsilon}\Sigma_{gg}^{-1}E^{-1}$ with $\Sigma_{gg} = plim(G'G/T)$, $\Sigma_{gg,\epsilon} = plim(Z'\epsilon\epsilon'Z/T)$, $E = diag(V^{-1}Q\Sigma_\Lambda, I)$ being block diagonal. A consistent estimator for Σ_δ denoted by $\hat{\Sigma}_\delta$, is given by

$$\hat{\Sigma}_\delta = \left(\frac{1}{T}\sum_{t=1}^{T-h}\hat{g}_t\hat{g}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T-h}\hat{\epsilon}_{t+h}^2\hat{g}_t\hat{g}_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T-h}\hat{g}_t\hat{g}_t'\right)^{-1} \tag{1.19}$$

Note that the covariance estimator in (1.19) is robust to heteroscedasticity. If homoscedasticity is assumed, we have $E(\epsilon_{t+h}^2) = \sigma_\epsilon^2$ for all t and the following covariance estimator is obtained

$$\hat{\Sigma}_\delta = \hat{\sigma}_\epsilon^2 \left(\frac{1}{T} \sum_{t=1}^{T-h} \hat{g}_t \hat{g}_t' \right)^{-1} \tag{1.20}$$

A consistent estimator for (1.20) is given by $\hat{\sigma}_\epsilon^2 = \sum_{t=1}^{T-h} \hat{\epsilon}_{t+h}^2$. Once the statistical properties of the coefficient estimators are obtained, Bai and Ng (2006) extends the result to the feasible conditional mean:

Theorem BN3: Let $\hat{y}_{T+h|T} = \hat{\delta}' \hat{g}_T$. Under the assumptions of Theorem BN1 and $\sqrt{N}/T \to 0$,

$$\frac{\hat{y}_{T+h|T} - y_{T+h|T}}{var(\hat{y}_{T+h|T})} \xrightarrow{d} N(0,1) \tag{1.21}$$

where $var(\hat{y}_{T+h|T}) = \frac{1}{T}\Sigma_\delta \hat{g}_T + \frac{1}{N}\hat{\alpha}' \Pi_T \hat{\alpha}$ and Π_T is the asymptotic variance of $\tilde{F}_T$ described in Theorem B1. $var(\hat{y}_{T+h|T})$ can be consistently estimated replacing Σ_δ and Π_T with their consistent estimators $\hat{\Sigma}_\delta$ and $\hat{\Pi}_T$ described above.

The forecasting error for (1.15) is given by

$$\hat{\epsilon}_{T+h} = \hat{y}_{T+h|T} - y_{T+h} = (\hat{y}_{T+h|T} - y_{T+h}) - \epsilon_{T+h}$$

Therefore, assuming that ϵ_t is normally distributed, $\hat{\epsilon}_{T+h}$ is also (approximately) normally distributed with variance

$$var(\hat{\epsilon}_{T+h}) = var(\hat{y}_{T+h|T} - y_{T+h}) = \sigma_\epsilon^2 + var(\hat{y}_{T+h|T})$$

Corollary BN1: Under the assumptions of Theorem BN3 and assuming that ϵ_t is normally distributed, then the forecasting error $\hat{\epsilon}_{T+h}$ is

$$\hat{\epsilon}_{T+h} \ \sim \ N(0, \sigma_\epsilon^2 + var(\hat{y}_{T+t|T})) \tag{1.22}$$

1.4 Missing Data

All large sample results obtained for factor models assume that the data set under consideration is balanced, i.e. it does not suffer from missing data. Minimization problem in (1.7) to estimate factors and loadings cannot be directly carried out when data set X has missing values.

Below we first discuss common missing data patters. Then, we give a brief account of missing data mechanisms. Finally, we go over the basic possible remedies to missing-data problem.

1.4.1 Missing Data Patterns

Missing data patterns describe which values are missing and which ones are observed. Below we briefly outline four common missingness patterns. Note that in tables below cells with NAs corresponds to unobserved values.

Univariate Nonresponse: This case is observed when the missingness is confined to a single variable. This is the first incomplete-data problem to receive systematic attention in statistics literature. Table 1.1 shows an example of univariate nonresponse.

X_1	X_2	X_3	X_4	X_5
NA				
NA				
NA				

Table 1.1: Univariate Nonresponse

Block Missingness: This pattern is also known as multivariate two patterns. As can be seen in Table 1.2, a group of variables are missing for exactly the same samples. Note that univariate nonresponse is a special case of block missingness.

X_1	X_2	X_3	X_4	X_5
NA	NA	NA		
NA	NA	NA		
NA	NA	NA		

Table 1.2: Block Missingness

Monotone Missingness: A common missing-data problem in longitudinal studies is attrition, where subjects drop out prior to the end of the study. Similarly, data on some subjects may start to

be collected later than the other subjects, or the data panel may be obtained by combining time series of different periods. In such cases we observe a reverse attrition. The pattern of attrition and reverse attrition are examples of monotone missing data. Table 1.3 shows a case of reverse attrition.

X_1	X_2	X_3	X_4	X_5
NA	NA	NA		
NA	NA			
NA				

Table 1.3: Monotone Missingness

General Missingness: In this case missingness does not gave a certain structure, and distributed to the data set randomly. Table 1.4

X_1	X_2	X_3	X_4	X_5
				NA
NA			NA	
NA				
		NA		
NA			NA	

Table 1.4: General Missingness

1.4.2 Missing-Data Mechanisms

Missing-data mechanisms are about the relation between missingness and the values of variables in the data matrix. In this respect, mechanisms that lead missing data concern how variables that are missing are connected to variables that are observed in the data set. These mechanisms determine the appropriateness of missing-data methods in different settings.

Let $X = (x_{it})$ be a data panel, and define a missing-data indicator matrix $M = (m_{it})$, such that $m_{it} = 1$ if x_{it} is missing and $m_{it} = 0$ if it is observed. Note that M defines the pattern of the missing data. The missing-data mechanism is characterized by the conditional distribution of M

given X. Statistically, we can express this characterization with $f(M|X, \phi)$, where ϕ denotes unknown parameters (if any).

If missingness does not depend on the values of X (missing or observed), we have

$$f(M|X, \phi) = f(M|\phi) \quad \text{for all } X,\ \phi$$

and the data are called missing completely at random (MCAR). Although MCAR states that missingness does not depend on the data values, it does not put any restrictions on the missingness pattern.

Let X^m and X^a denote missing and available (observed) components of X, respectively. If missingnes depends only on the observed variables of X, the data is said to be missing at random (MAR). MAR case can be expressed as follows:

$$f(M|X, \phi) = f(M|X^a, \phi) \quad \text{for all } X^m,\ \phi$$

Finally, if the distribution of M depends on the missing values of , the data are called not missing at random (NMAR):

$$f(M|X, \phi) = f(M|X^m, \phi) \quad \text{for all } X^a,\ \phi$$

A good example of NMAR case is the official questionnaires with income related questions: People with high income are more likely to omit income related questions considering possible tax consequences. In this case, the fact that a variable (income) is missing in the survey is explained by the level of the income itself. If the data are NMAR, the observed data set is not representative for the population it comes from.

MAR case can be found in many financial data panels. As an example, consider a data set containing certain firm characteristics (columns) such as market cap, momentum, leverage etc. across firms (rows): In general it is less likely to access financial statements for firms with lower market capitalization since they tend to be less professional and unregulated, and their data are generally stale. Therefore one experiences more missing values in firm characteristics of smaller firms. However the market capitalization variable is almost always observed regardless of the firm being an essential statistic for any firm that is traded in some exchange. In this case the fact that a variable (e.g. momemtum) is missing can be explained by a variable (market cap) that is observed.

Throughout this study, we assume that the missing data panel we work on is missing at random. The factor-based imputation method we discuss below becomes functional with the help of this assumption.

1.4.3 Remedies for Missing Data

In general, there are two possible remedies for missing data: Amputation(truncation) and imputation. Below we discuss each of them briefly.

Truncation

Truncation means deleting missing columns or rows of X to obtain a balanced data set. It basically takes two forms:

(a) Listwise Deletion: In this approach, each sample observation (row) with missing values are eliminated. This strategy, which is called complete-case analysis, is generally inappropriate since the researcher is in general interested in making inferences about the entire population, rather than the portion of the target population that could be observed.

(b) Variable Deletion: Each variable with missing values are eliminated. This case is not suggested in many settings since the resulting data set certainly does not represent the population that initial data set comes from.

In general, truncation methods result in artificially created balanced data sets that may not be representative, and and lead to inefficient estimators due to resulting observation loss. Under the assumption that the data set is missing at random, standard large sample result holds for listwise deleted data. On the other hand, in the case of variable deletion, large sample results are not directly comparable since the variable set after the deletion is smaller than that of the original data panel.

Imputation

Imputation means estimating missing values in a data set by using the available information in the data set, using outside information, or both. Since imputed values are estimates, they contain estimation error. Ignoring the estimation error embedded in imputed values lead to biased estimates, underestimated standard errors and invalid hypothesis test. For example Stock and Watson (2002a) and Banbura and Modugno (2010) recursively impute the missing values during factor estimation using the available observations in the data set, and do not take into consideration the effect of the estimation error on factor and loading estimates and their standard errors.

In section 1.5 below, we develop a factor-based single imputation method. There are three main reasons we choose this method: First, it works very well with large data sets that follow a factor

structure. Second, it is very *parsimonious* since it uses only k ($k << min(N,T)$) auxiliary factors which are estimated from balanced subpanels of the incomplete data set. Therefore, the number of predictive variables (factors) in the imputation method remains the same even if the dimension of the incomplete data set and the amount of missingness increase arbitrarily. Finally, factor-based imputation allow us to be able to apply the findings in the factor-augmented forecasting literature introduced by the seminal papers of Stock and Watson (2002a), Bai(2003) and Bai and Ng (2006).

1.5 Factor Models under Missing Data

In this section we introduce factor models for block-missing and monotone-missing data sets. Next, using the models for missing data, we develop a parsimonious factor-based imputation (FBI) method that consistently imputes the common component of missing values. Finally, we show how to obtain the ID factor and loading estimators.

1.5.1 Factor Models under Block-Missing Data

In this section we consider the analytics of factor model estimation under block missingness. Consider the $T \times N$ block-missing data matrix X in Table 1.5 where gray and white cells represent missing and available values, respectively. The data set is partitioned as $X = [Y\ Z]$ where $Y = [Y^{m\prime}\ Y^{a\prime}]'$ and $Z = [Z^{m\prime}\ Z^{a\prime}]'$[7]. Y^m and Y^a represent the missing and available parts of Y, respectively[8], while Z denotes variables which are completely observed[9]

Note that we have $N = N_m + N_a$ and $T = T_m + T_a$ where N_m and N_a are the numbers of missing and available variables in X, while T_m and T_a represent the numbers of missing and available time series observations for the variables in Y, respectively.

Assuming X has a factor structure, we aim to estimate ID factors under block-missingness and uncover their distributional properties. Since the data set is *unbalanced*, the method described in Bai (2003) is not directly applicable for estimation. In order to overcome this problem, we suggest to impute the missing data with a FBI method. To this end, first we describe the factor structure that X follows, and then introduce our consistent FBI method that is used to impute missing values.

---

[7] One can also partition X as $X = [X^{m'}\ X^{a'}]'$ where $X^m = [Y^m\ Z^m]$ and $X^a = [Y^a\ Z^a]$.

[8] Without loss of generality, we assume that the missing block lies on the northwest corner of X. Note that the order in which we place the series in the data matrix does not effect factor estimation. Therefore, given that all the missing variables are unobserved over the same period, we can always adjust the order of the series to generate a data matrix with a missing block on the northwest corner.

[9] Although Z is fully observed, we partition it to match the partition in Y. In calculations below this representation will be useful.

	N_m	N_a
T_m	Y^m	Z^m
T_a	Y^a	Z^a

Table 1.5: Block-missing data panel

We assume X has a r-factor structure, and can be expressed as in (1.1). Following the partition of X above, we can partition Λ, F and e as follows:

$$\Lambda = [\ \Lambda^{m\prime}\ \Lambda^{a\prime}\ ]'$$

$$F = [\ F^{m\prime}\ F^{a\prime}\ ]'$$

$$e = \left[\begin{array}{cc} e^y & e^x \end{array}\right] = \left[\begin{array}{cc} e^{y^m} & e^{x^m} \\ e^{y^a} & e^{x^a} \end{array}\right]$$

where Λ^m and Λ^a are $N_m \times r$ and $N_a \times r$, $F^{m\prime}$ and $F^{a\prime}$ are $T_m \times r$ and $T_a \times r$, e^y and e^x are $T \times N_m$ and $T \times N_a$ respectively. Using this notation, the factor model given in (1.1) can be expressed in vector form as

$$Y_t = \Lambda^m F_t + e_t^y \tag{1.23a}$$

$$Z_t = \Lambda^a F_t + e_t^x \tag{1.23b}$$

and in matrix form as

$$Y^m = F^m \Lambda^{m\prime} + e^{y^m} \tag{1.24a}$$

$$Y^a = F^a \Lambda^{m\prime} + e^{y^a} \tag{1.24b}$$

$$Z^m = F^m \Lambda^{a\prime} + e^{z^m} \tag{1.24c}$$

$$Z^a = F^a \Lambda^{a\prime} + e^{z^a} \tag{1.24d}$$

The Factor-Based Imputation Method for Block-Missing Data

In this section we propose a FBI algorithm for block-missing data sets, which imputes the missing values Y^m with the estimate of its common component[10]. Note that Y^m and Y^a have the same loadings, Λ^m. If we knew the true factors F, we could have obtained the estimate of Λ^m by regressing Y^a on F^a. Fortunately Bai and Ng (2006) and Stock and Watson (2002a) show that under certain assumptions, regressing Y^a on F^a yields the same common component estimate with regressing it on $\tilde{F}^a$ as $N, T \to \infty$. Since X is missing $\tilde{F}$ is infeasible; however another estimate of F that is obtained from the observed portion of X can be used as a proxy for $\tilde{F}$. In this respect, let $\tilde{F}_z$ denote factor estimate obtained from the partition Z. As long as $\tilde{F}_z$ satisfies the conditions under which Bai and Ng (2006) and Stock and Watson (2002a) obtain their results[11], it can be used to estimate[12] Λ^m.

Since Z is completely observed, $\tilde{F}_z$ is simply $\sqrt{T}$ times the eigenvectors of the matrix ZZ' corresponding to the first r eigenvalues of ZZ' in decreasing order. By construction, we have have $\tilde{F}_z'\tilde{F}_z/T = I$. Using $\tilde{F}_z$, the regression equation for the i^{th} variable (column) in Y^a is given by

$$\underline{Y}_i^a = \tilde{F}_z^a \lambda_i + \nu_i \quad , \quad i = 1, 2, ..., N_m \tag{1.25}$$

where $T_a \times 1$ vector $\underline{Y}_i^a$ is the observed part of the i^{th} series, $r \times 1$ vector λ_i is the unobserved loading for variable i, $T_a \times r$ matrix $\tilde{F}_z^a$ is the available portion of the factor estimate $\tilde{F}_z$, and $\mathrm{T}_a \times 1$ vector ν_i is the error term. Carrying out the time series regression in (2.1), we obtain

$$\hat{\lambda}_i = (\tilde{F}_z^{a\prime}\tilde{F}_z^a)^{-1}\tilde{F}_z^{a\prime}\underline{Y}_i^a \tag{1.26}$$

Using $\hat{\lambda}_i$ above, an estimate of $\underline{\hat{Y}}_{it}^m$ can be obtained as

$$\underline{\hat{Y}}_{it}^m = \hat{\lambda}_i'\tilde{F}_{zt}^a \tag{1.27}$$

---

[10]One might be tempted to investigate the properties of factor model estimates under an imputation procedure that consistently estimates Y^m itself (rather than the common component of Y^m). However, idiosyncratic component of Y^m is pure noise, hence it cannot be estimated.

[11]In the derivations below, we assume that these conditions are satisfied.

[12]At this point, it is important to establish an important difference between $\tilde{F}_z$ and $\tilde{F}$: In the complete-data factor model that we review in Section 1.2, the elements of the model (i.e F, Λ and e) and its assumptions are defined for $X = F\Lambda' + e$, and the large sample results for the factor model estimators obtained in the literature are obtained for this "whole" factor model. However, Z represents a portion of X, and the factor model it exhibits, $Z = F\Lambda^{a'} + e^z$, is "partial". In this sense, although the complete-data assumptions lead to a consistent and asymptotically normal factor estimator for X, they do not guarantee the same well-behaved factor estimator for Z. In the Chapter II, we will address this issue and show that under very weak additional assumptions $\tilde{F}_z$ is well-behaved.

Calculating (1.26) for all i, we obtain the estimate of the $r \times N_m$ matrix $\hat{\Lambda}^{m\prime} = [\hat{\lambda}_1,\ \hat{\lambda}_2, \ldots, \hat{\lambda}_{N_m}]$. Then the common component estimate of Y^m is obtained as

$$\hat{Y}^m = \tilde{F}_z^m \hat{\Lambda}^{m\prime} \tag{1.28}$$

Replacing Y^m in X with $\hat{Y}^m$, we obtain the imputed data matrix $\hat{X}$. Since $\hat{X}$ does not suffer from missingness, standard estimation procedures that we discuss in Section 1.2 apply to it.

1.5.2 Factor Models under Monotone-Missing Data

In this section we extend the discussion in the previous section to monotone missing data. Consider the following $T \times N$ data set $X = [Y^1\ \ Y^2\ \ Z]$:

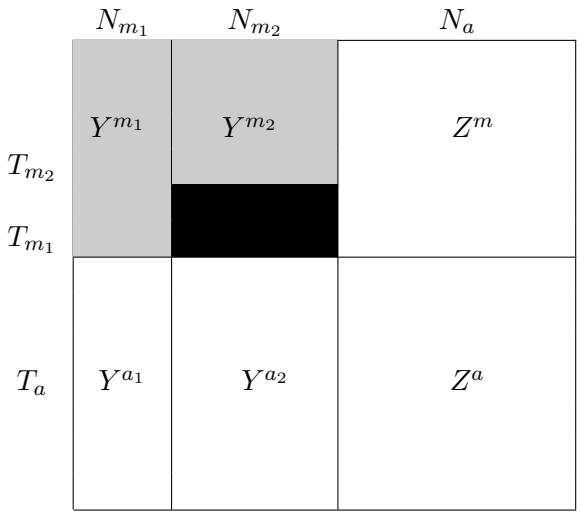

Table 1.6: 2-layer monotone-missing data panel

Unlike the block-missing data set in section 1.5, here we have two[13] sets of missing variables, i.e. *layers*[14]: Y^1 and Y^2. In Table 1.6, as in the case of block-missingness, gray regions denote the missing observations. First and second missing layers are given by Y^{m_1} and Y^{m_2}, respectively. As can be seen, while for the variables in the first layer first T_{m_1} observations are missing, missingness is confined to the first T_{m_2} observations for the second layer variables. There are N_{m_1} and N_{m_2} variables in the first and second layers, respectively. Alternatively, one can approach the missing data problem at hand

[13]To keep the exposition brief, we initially consider a monotone-missing data set with two missing layers. Below, we generalize our findings to an arbitrary number of missing layers

[14]In this sense, one can interpret block-missingness as a special type of monotone-missingness with one missing layer.

from the point of view of *sample observations* (i.e. rows) at each point of time rather than from that of the *variables* (i.e. columns). In this respect, for samples from time 1 through T_{m_2}, first $N_{m_1} + N_{m_2}$ observations are missing, and for samples from time $T_{m_2} + 1$ through T_{m_1}, first N_{m_1} observations are missing [15].

N_a denotes the number of available series, and T_a stands for the number of completely observed samples in X; therefore it follows that $N = N_{m_1} + N_{m_2} + N_a$ and $T = T_{m_1} + T_a$. We define the available parts of first and second layer of missing variables with Y^{a_1} and Y^{a_2}, respectively and denote the available part of all missing layers[16]. with $Y^a = [Y^{a_1}\ Y^{a_2}]$.

As in the previous section, variables in Z are completely observed, and Z is partitioned to match the partitions in the missing variables. The extension of the number of missing layers to an arbitrary number J is straightforward. To that end, below first we characterize a factor model with J-level missingness, and then discuss the FBI under monotone missingness. Suppose X exhibits J-layer monotone missingness, and has the static r-factor structure given in (1.1). Then we have[17]

$$\Lambda = [\ \Lambda^{m'_1}\ \Lambda^{m'_2}\ ...\ \Lambda^{m'_J}\ \Lambda^{a'}\ ]' \qquad , \qquad j = 1, 2, ..., J$$

$$e = [\ e^{y_1}\ e^{y_2}\ ...\ e^{y_J}\ e^{x}\ ]$$

where Λ^{m_j} is $N_{m_j} \times r$, and e^{y_j} is $T \times N_{m_j}$. For each e^{y_j}, we have e^{m_j} and e^{a_j} which are $T_{m_j} \times N_{m_j}$ and $T_a \times N_{m_j}$, respectively. Let $S_c = \sum_{j=1}^{c} N_{m_j}, c = 1, 2, ...J$. Then the number of completely observed series and samples are given by $N_a = N - S_J$ and $T_a = T - T_{m_1}$, respectively. Partitioning F for each missing layer is a bit tricky since each layer has overlapping factor partitions. In this respect, we define the $T_{m_j} \times r$ matrix F^{m_j} as the portion of F that stems from the cross-section observation $t = 1$ to $t = T_{m_j}$. Then, the missing layers of X can be expressed in vector form as the following:

$$Y_t^j = \Lambda^{m_j} F_t^j + e_t^{y_j} \qquad t = 1, 2, ..., T_{m_j}$$

---

[15] In the simulation experiments below, we will show that this second interpretation is more relevant in explaining the large sample properties of the ID factor estimators.

[16] Note that the observations shown in the black region of Table 1.6 are not defined in Y^{a_2} although they are observed. This definition of Y^{a_2} does not affect the large sample properties of the ID estimators and it yields much more tractable analytical results. On the other hand, it is obvious that leaving out the black region in estimation yields inefficient estimators in the small sample. Therefore one can add the black region to Y_1^a while estimating ID estimators in applications

[17] Note that the structure of Z is the same with the block-missing data; hence we do not rewrite it here to preserve space.

Accordingly, missing layers and available parts of Y are given in matrix form as follows:

$$Y^{m_j} = F^{m_j}(\Lambda^{m_j})' + e^{m_j} \tag{1.29a}$$

$$Y^{a_j} = F^{a}(\Lambda^{m_j})' + e^{a_j} \tag{1.29b}$$

Factor-Based Imputation Algorithm for Monotone-Missing Data

Following the FBI introduced above, we impute missing values by their common component estimate[18]. Let $\tilde{F}_z$ denote the auxiliary factor estimate obtained from Z as in the previous section. Since missing and available observations in the j^{th} layer share the same loadings, Λ^{m_j}, we can estimate the loadings of Y^{m_j} by regressing Y^{a_j} on $\tilde{F}$ as follows:

$$\underline{Y}_i^{a_j} = \tilde{F}_z^a \lambda_i^j + \nu_i^j \quad , \quad i = 1, 2, ..., N_{m_j} \quad , \quad j = 1, 2, ..., J \tag{1.30}$$

where $T_a \times r$ matrix $\tilde{F}_z^a$ is the available portion of the factor estimate $\tilde{F}_z$, $r \times 1$ vector λ_i^j and ν_i^j is the unobserved loading for the i^{th} variable (column) in the j^{th} missing layer and ν_i^j is the corresponding error term. Carrying out the time series regression in (2.9), we obtain

$$\hat{\lambda}_i^j = (\tilde{F}^{a\prime}\tilde{F}^a)^{-1}\tilde{F}^{a\prime}\underline{Y}_i^{a_j}$$

We can run the above regression for each series in each Y^{a_j}, and obtain the estimate of the $r \times N_{m_j}$ matrix $\hat{\Lambda}^{m_j'} = [\hat{\lambda}_1^j \;\; \hat{\lambda}_2^j \; \ldots, \hat{\lambda}_{N_{m_j}}^j]$. Let $\tilde{F}_z^{m_j}$ denote the part of $\tilde{F}_z$ between cross-section observations $t = 1$ and $t = T_{m_j}$. Then, a consistent estimate of the common component of Y^{m_j}, $F^{m_j}\Lambda^{m_j'}$, is obtained as

$$\hat{Y}^{m_j} = \tilde{F}^{m_j}\hat{\Lambda}^{m_j'}$$

Finally, replacing Y^{m_j} with $\hat{Y}^{m_j}$ for $j = 1, 2, ..., J$, we obtain the completed data matrix $\hat{X}$.

1.5.3 Estimation of ID Factors and Loadings

The imputed data matrix $\hat{X}$ is balanced, therefore the factor and loading estimates can be obtained following the standard complete-data procedure. Accordingly, if $N < T$, ID loading estimate is given by $\bar{\hat{\Lambda}} = \sqrt{N}Z$ where Z is the eigenvectors corresponding to the k largest eigenvectors of the $N \times N$ matrix $\hat{X}'\hat{X}$. $\bar{\hat{F}}$ is found by simply regressing $\hat{X}$ on $\bar{\hat{\Lambda}}$, as $\bar{\hat{F}} = \hat{X}\bar{\hat{\Lambda}}/N$.

[18]This section closely follows section 1.5.1; therefore some details are omitted in the presentation to avoid repetition.

Likewise, for $T < N$, ID factor estimate, $\tilde{\tilde{F}}$, is given by $\sqrt{T}$ times the eigenvectors of the matrix $\hat{X}\hat{X}'$ corresponding to the first k eigenvalues in decreasing order. Given $\tilde{\tilde{F}}$, ID loading estimate is obtained as $\tilde{\tilde{\Lambda}} = \hat{X}'\tilde{\tilde{F}}/T$.

Using the result in section 1.2.1, one can show that $\bar{\bar{F}}$ is a linear function of $\tilde{\tilde{F}}$, and both ID factor estimators estimate the same latent true factor space. In the rest of this paper in order to avoid confusion, we only consider the case $T < N$ and denote ID factor estimators with $\tilde{\tilde{F}}_{ID}$.

1.6 Monte Carlo Experiments

This section considers our main research questions, simulation design and experiments. For both block and monotone missingness, the questions and the experiments are identical. Below, we first focus on the research questions regarding the large sample properties of ID estimators. Next, we design a simulation experiment around those questions. Finally, we carry out an extensive simulation experiment to unveil the large sample properties of the ID estimators.

1.6.1 Research Questions

Standard large sample theory provided by large dimensional factor models literature does not consider the effect of imputed values in data set $\hat{X}$. In this respect, this paper seeks answers to the following questions:

i. Should the missingness rate be kept constant or should it be allowed to decrease as N and T approach infinity?

ii. Does $\tilde{\tilde{F}}_{ID}$ consistently estimate the factor space?

iii. Is $\tilde{\tilde{F}}_{ID}$ asymptotically normal?

iv. How does estimation error incurred during imputation affect the asymptotic variance of ID factor estimator?

v. How does missingness affect the convergence rates described in the standard large sample literature?

We try to answer the questions above with Monte Carlo experiments. To this end, we use the factor-based imputation procedure described above to estimate the missing values in X. Although

Monte Carlo experiments are generally used to study the small sample behavior of estimators, our current study uses them to approximate the answers to the questions above regarding large sample behaviors of factor estimators under missingness. Simulation experiments provide the benefit of a controlled experimental setting; with the help of the experiments below, we hope that the results of these experiments lead to new insights and directions for the future theoretical research.

Note that the first question above asks how to treat the relative growth rates of the number of missing and available values in large sample. The answer to this question is one of the most important contribution of the current study since it determines the way we approach the analytics of the limiting distribution of ID estimators. In this respect in the simulation experiments below, we first decide whether the missingness ratio should be kept constant or let decrease. Once the proper treatment of the missingness ratio is determined, we focus on the remaining questions.

1.6.2 Simulation Design

Below we design a Monte Carlo experiment to answer the questions raised above. First, we introduce the factor structure to generate data sets to be used in the simulation experiment. Next, we discuss the generation of block and monotone-missing data sets. Finally, we define various ID estimators that are distinguished from one another with respect to the characterization of the relative growth of missingness and the imputation procedure applied.

Data Generation

We consider the following factor structure for $i = 1, ..., N$ and $t = 1, ..., T$

$$\begin{aligned} X_{it} &= \lambda_i' F_t + e_{it} \\ e_{it} &\sim N(0,1) \\ \lambda_i &\sim N(0, diag(\delta/r)) \\ F_{t+1}^* &= \rho F_t^* + \eta_t \\ F_t &= \frac{F_t^* - \mu}{\sigma_F} \end{aligned} \tag{1.31}$$

where F_t and λ_i are $r \times 1$ factors at time t and loadings for series i, respectively. $diag(\delta/r)$ denotes a $r \times r$ diagonal matrix with on-diagonal elements equal to δ/r. Since λ follow a zero-mean normal distribution, it takes both negative and positive values (with equal rates) so that series X_i loads on F_t in both negative and positive directions. We fix $\rho = .7$ so that factor follows a persistent AR(1)

process. Note that μ_F and σ_F denote the $r \times 1$ mean and standard error vectors of F_t^*. Therefore, the normalized factors F_t have mean 0 and variance 1. δ determines the signal-to-noise (STN) ratio[19] of each series defined by

$$STN = \frac{r\sigma_\lambda^2}{\sigma_e^2} = \delta$$

where σ_e and σ_λ are the standard errors of e and λ_i, respectively. Throughout the simulations below, we let $\delta = 2$ so that common component explains 66% of the variation[20] in X.

We divide the series in X into left and right halves. In the two-factor model, we consider two different cases for the loadings: (a) all loadings are kept in the same way they are generated. We call this case the *complete loadings* case. (b) for series on the left half, second loading is set to zero, and for series on the right half, first loading is set to zero; that is, series on left load only on the first factor while the ones on right load only on the second one. We call this case the *half loadings* case. We consider the half loadings case in order to understand the behavior of ID factors when the factors that missing and observed variables load on are not entirely the same.

Missingness Generation

After creating the data set described above we generate block and monotone missingness, and obtain data sets similar to those in tables 1.5 and 1.6. In the case of block missing data set, we simply define the percentage of missing values by determining the percentage of missing time series and cross-section observations. On the other hand, for monotone missingness we need to define several missing blocks and this requires us to choose both the number of missing blocks multiple missingness levels of time series and cross-section observations for the the simulation experiment. The details of this selection process and model parameters are discussed thoroughly in section 1.6.3 below.

ID Estimators

After block and monotone-missing data sets are generated, missing values are imputed using three different imputation methods: (i) unconditional mean imputation, (ii) our FBI procedure and (iii) random imputation. In method (i), we impute the mean of each missing column[21], whereas in method

[19]This result is obtained since the variance of each factor is equal to 1, and factors, loadings and idiosyncratic errors are independent.

[20]Note that our simulation results below are seen to be robust to other reasonable choices of δ, as well.

[21]Note that in the simulations below we generate the data set X in such a way that each column of X has zero mean. In the mean-imputation case for the sake of simplicity we missing values with the population mean.

(iii) missing values are replaced with some random number[22]. In addition, we consider two scenarios for the relative size of the missing piece as data dimensions increase: (a) missingness expands *at the same rate* with the data set, (b) missingness expands *slower* than the data set. Under the above setting, we obtain four ID estimators:

i. Constant ID estimator: It uses method (ii) and box scenario (a), and is denoted by $\tilde{\tilde{F}}_{CID}$.

ii. Decreasing ID estimator: It uses method (ii) and box scenario (b), and is denoted by $\tilde{\tilde{F}}_{DID}$.

iii. Unconditional Mean ID estimator: It uses method (i) and box scenario (a)[23], and is denoted by $\tilde{\tilde{F}}_{UID}$.

iv. Random ID estimator: It uses method (iii) and box scenario (b), and is denoted by $\tilde{\tilde{F}}_{RID}$.

In addition to the estimators above, we consider the *infeasible* factor estimator $\tilde{F}$ which we study in detail in Section 1.2. In order to conform with the notation in this section we denote it with $\tilde{\tilde{F}}_{INF}$. Note that standard inferential theory applies $\tilde{\tilde{F}}_{INF}$ since it is obtained from the complete data X. In this sense, among other uses $\tilde{\tilde{F}}_{INF}$ serves as a check mechanism for the accuracy of our simulation experiment.

In Monte Carlo experiments, N and T are simultaneously increased, and missing data creation and factor model estimation processes are repeated 500 times for each (N, T) pair. For consistency-related questions in Section 1.5, following Stock and Watson (2002a) we summarize the results by trace R^2 of multivariate regression of $\tilde{\tilde{F}}$ on F given by

$$R^2_{\tilde{\tilde{F}},F} = \frac{\hat{E}||P_F\tilde{\tilde{F}}||}{\hat{E}||\tilde{\tilde{F}}||} = \frac{\hat{E}tr(\tilde{\tilde{F}}'P_F\tilde{\tilde{F}})}{\hat{E}tr(\tilde{\tilde{F}}'\tilde{\tilde{F}})} \tag{1.32}$$

where $||.||$ denotes the Frobenius norm[24], $P_F = F(F'F)^{-1}F'$, and $\hat{E}$ denotes the expectation estimated by averaging the relevant statistics over Monte Carlo repetitions. We use the measure in equation (1.32) for all factor estimators mentioned above. Note that under fairly weak assumptions Stock and Watson

[22]We replace each missing cell with a value drawn from $U[-100, -80]$, where $U[.]$ denotes the uniform distribution. Note that method (iii) serves as a check mechanism for whether the accuracy of the estimates obtained with other imputation methods is solely due to the increase in N and T.

[23]Note that we let the behavior of the missing block in UID case be the same with that of CID. Therefore, the only difference between UID and CID cases is that the former uses unconditional mean imputation while the latter uses conditional mean imputation.

[24]Frobenius norm of a matrix A is given by $||A|| = [tr(A'A)]^{1/2}$

(2002a) shows that $R^2_{\tilde{F},F} \xrightarrow{p} 1$ as $N, T \to \infty$. Therefore, we test the consistency of the ID estimators and their convergence rate with the behavior of their R^2 as N and T grow large.

Regarding the limiting distribution related questions in Section 1.6.1, for the INF estimator we use $\tilde{\Pi}_t$ in (1.9). By mimicking INF asymptotic variance estimator, for ID estimators we consider

$$\tilde{\tilde{\Pi}}_t = \hat{V}^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\tilde{\tilde{e}}_{it}^2\tilde{\tilde{\lambda}}_i\tilde{\tilde{\lambda}}_i'\right)\hat{V}^{-1} \tag{1.33}$$

where $\hat{V}$ is a diagonal matrix consisting of the first r eigenvalues of $(1/NT)\hat{X}\hat{X}'$ and $\tilde{\tilde{e}}_{it} = \hat{X}_{it} - \tilde{\tilde{\lambda}}_i'\tilde{\tilde{F}}_t$. Note that above $\tilde{\tilde{\Pi}}_t$ does not take into account the additional estimation error in $\hat{X}$ due to imputation, and is not necessarily consistent for Π_t. Therefore, although scaling the INF estimator by definition yields an asymptotically standard normal factor estimator, scaling the ID estimators may fail to do so. In the simulation experiments below we consider the effect of imputation error on ID estimators' asymptotic variances using the scaled asymptotic variance estimator given in (1.33).

1.6.3 Simulation Experiment

As mentioned above, we aim to determine the large sample properties of ID estimators. We increase data dimensions and missingness, and check the large sample properties of ID estimators during this process. To this end, we let N and T increase with the following rule:

$$N(j) = 50(1.1^j),\ T(j) = 25(1.1^j)$$

where $N(0) = 50$ and $T(0) = 25$ are the base dimensions, j is a *multiple* whose range changes across simulations (between[25] 10 and 40). Let $X(j)$ denote the $T(j) \times N(j)$ data panel. After determining the evolution of the complete data set, we consider the missing part and its evolution relative to the complete data set.

Block-Missing Data

Let $T_m(j)$ and $N_m(j)$ denote the number of missing time series samples (rows) and variables (columns) in $X(j)$, respectively. Using these, we can define the *missingness ratios* for time series and cross-section dimensions as $TR(j) = \frac{T_m(j)}{T(j)}$ and $NR(j) = \frac{N_m(j)}{N(j)}$, respectively. It follows that $MR(j) = TR(j)NR(j)$ is the total missingness ratio in $X(j)$. In scenario (a), we fix $TR(j) = NR(j) = \sqrt{\psi}$ (for some $\psi > 0$) which yields $MR(j) = \psi$. Therefore, as T and N increase T_m and N_m are increased proportionally so that MR remains constant at ψ.

[25]Note that $N(10) = 130$, $N(40) = 2262$, $T(10) = 65$, $T(40) = 1131$.

On the other hand, in scenario (b) we let T_m and N_m increase *slower* than T and N so that $TR(j)$ and $NR(j)$ decrease as j increases. We let $TR(j) = NR(j) = \sqrt{\psi}(0.95^j)$ which results in $MR(j) = \psi(0.95^j)$. Note that since $0.95 * 1.1 > 1$, although TR and NR decrease T_m and N_m increase. In other words, the number of missing values increases (in the limit there are infinitely many of them) while the ratio of missing values to observed ones decreases.

In order to test the normality of ID estimators, we follow a slightly different path: First, we fix $T = T(j)$, $N = N(j)$ for some high j. We then generate a data set and a missing block, impute the missing block with different imputation methods described above, and obtain the ID estimates. We repeat this process 500 times and sample from each ID estimate. Finally, we process these samples to test the asymptotic normality of ID estimators. Section 1.7.1 below give a thorough account of the entire process of normality tests.

Monotone-Missing Data

We consider a 2-layer missingness structure for the experiment regarding the monotone missingness. Missingness ratios of the first layer is denoted by $TR_s(j) = \frac{T_{m_s}(j)}{T(j)}$ and $NR_s(j) = \frac{N_{m_s}(j)}{N(j)}$ for $s = 1, 2$. The total missingness ratio for the 2-layer case is given by

$$MR(j) = \frac{TR_1(j)NR_1(j) + TR_2(j)NR_2(j)}{T(j)N(j)}$$

In the simulation experiments, we let the dimensions for the second layer be $TR_2(j) = (0.4)TR_1(j)$ and $NR_2(j) = (1.25)NR_1(j)$. Note that except for the existence of the additional missing layer, we keep the all the other features (scenarios, growth rates etc.) same with the block missingness case above.

1.7 Results

Below we discuss the results for the Monte Carlo experiments considered in the previous section. First, we address the results for block-missing data sets for which we consider different factor models in terms of the number of true factors and the structure of loadings. Next, we treat the results for monotone-missing data sets. Finally, in the discussion section we elaborate on the large sample properties of ID estimators and their role in statistical inference.

1.7.1 Block Missingness

In this section we investigate the consistency and asymptotic normality properties of ID estimators under block missingness. First, we study the large sample properties of the estimated factors and loadings in the original form and size they are estimated. Next, we partition the estimators into missing and available parts, and study their large sample properties as if they were separate estimators; we refer to them as *partitioned* estimators. For example for estimated CID factors $\tilde{\tilde{F}}_{CID}$, the missing partition of the factor estimator is given by $\tilde{\tilde{F}}^m_{CID}$, while the available partition is equal to $\tilde{\tilde{F}}^a_{CID}$. Below we show that asymptotics of missing and available parts of the ID factor estimators are not the same, a property which is nonexistent for INF estimator.

Large Sample Properties of ID Estimators

Consistency Figure 1.1 below shows the R^2 given in (1.32) for the INF, CID, DID, UID and RID factor estimators for $r = 1$ and $MR = 30\%$[26], while Figures 1.2 through 1.4 demonstrates the same results for INF and CID for $r = 1$ and $r = 2$, with $MR = 50\%$ and 80%.

We start with the 1-factor model: As can be seen in Figure 1.1, INF estimator's R^2 monotonically converges to 1 as suggested by Stock and Watson (2002a). This suggests that INF estimator consistently estimates the true factor space as N and T approach to infinity. Similarly CID estimator estimates the true factor space consistently. Therefore, we show that the factor-based imputation method we suggest leads to a consistent factor estimator under a constant missingness ratio.

Another important point in Figure 1.1 is that RID estimator is consistent for the true latent factor space. It starts with a very low R^2 value for small data dimensions, but as the size of the missing block shrinks relative to the size of the data panel, the importance of missing data and imputed values plummet. This finding suggests that the estimators obtained from data sets that are imputed with completely random values would be consistent for their population counterparts as long as MR goes to zero (even if the total number of missing observations goes to infinity). In this respect, we use RID estimator as a test case: it demonstrates that factor consistency achieved under decreasing MR is obtained solely due to the increase in T and N, and does not depend on the quality of the imputation procedure. Therefore, the idea of slower growth of missingness compared to data set dimensions is useless for inference purposes. In this respect, we can ignore the results for DID estimator as well since it also assumes decreasing MR. Accordingly, below we confine our attention only to the CID

[26]Results for $r = 2$ through $r = 10$ are very similar, and are not reported here.

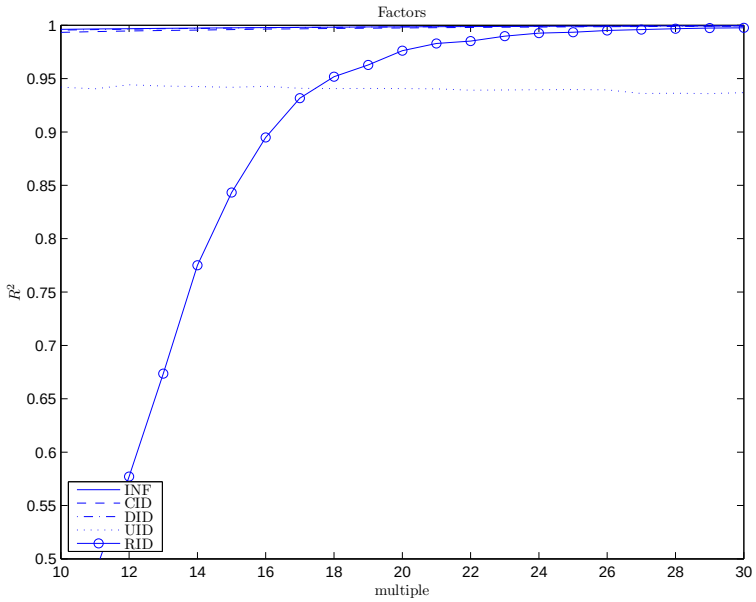

Figure 1.1: R^2 of all estimated factors with $r = 1$ (MR=30%)

estimator for which the missingness and data set dimensions grow at the same rate.

Note that UID estimator is inconsistent; its R^2 starts pretty high (around 0.94) but fails to converge to 1 for higher N and T values. In other words, imputing each missing value by its (true) unconditional mean does not deliver consistency in factor estimation when MR stays fixed. From this observation we can infer that if the missing values are imputed with random values when MR is fixed, the resulting factor estimation would be inconsistent, as well. In this sense, UID estimator serves as a test case: factor consistency achieved under constant MR does not only stem from the increase in T and N, but also results from the quality of the imputation method. This points out the fact that our FBI procedure, which delivers consistent CID factor estimates, sufficiently utilizes the dynamics of the factor structure and the missingness in the data set, and proves to be a useful imputation procedure.

Finally, Figure 1.2 considers an extreme case for INF and CID factor estimators with 80% missingness ratio. As can be seen in the figure, although R^2 of the CID estimator starts with a lower value compared to Figure 1.1, it nonetheless converges to 1 as data dimensions increase. Therefore, even with 80% missingness, CID estimator continues to be consistent.

Next, we consider the 2-factor model with complete and half loadings: As mentioned above, in the complete loadings case each series load on all the loadings while in half loadings case, left half of the series load only on one loading and the right half loads on the other. Figure 1.3 shows the R^2 of the INF and CID estimators for the complete loadings case. As can be seen, CID estimator is consistent

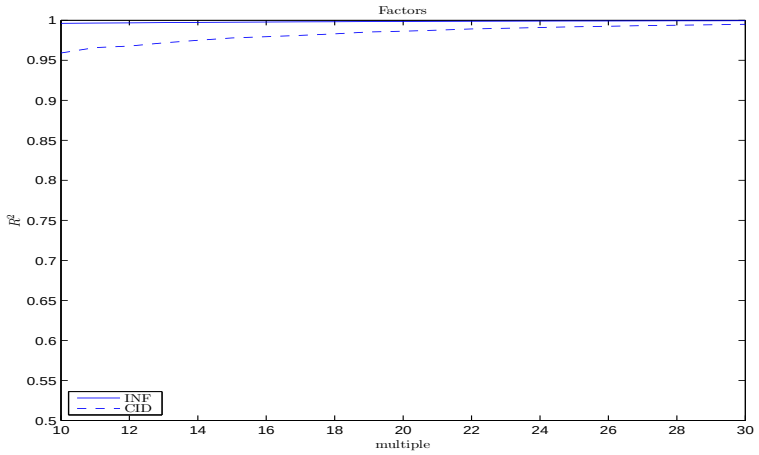

Figure 1.2: R^2 of INF and CID factor estimators with $r = 1$ (MR=80%)

again when the number of factors increased to two. Although CID estimator starts with a slightly smaller R^2 compared to Figure 1.2, it eventually converges to 1. Note that in both Figure 1.2 and 1.3 missingness ratio is 80%. Therefore controlling for the missingness ratio, that the CID estimator in the two-factor model converges slowly can be explained by the fact that with two factors the statistical error incurred in estimates of $\tilde{F}_z$ and (hence) $\hat{\lambda}$ are greater. Accordingly, these errors contaminate the common component imputation of Y^m for small N and T values.

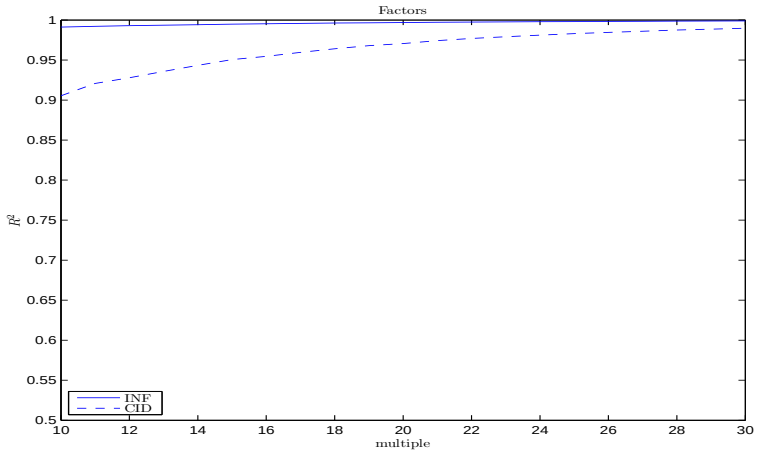

Figure 1.3: R^2 of INF and CID factor estimators with $r = 2$ (MR=80%)

Figure 1.4 reports R^2 results for the two-factor model with half loadings when the missingness ratio is 50%. By letting $MR = 50\%$ we guarantee that missing and available series load on different factors; therefore the initial factor estimate ($\tilde{F}_z$) used in imputation has no predictive value. As a

result CID estimator fails to converge and its R^2 levels around 0.64 even for very high N and T values. In addition, CID estimator is outperformed by the UID estimator (not shown in the graph). This observation suggests that if it is believed that available series contain little or no information about the missing ones, then the unconditional mean imputation can be preferred to conditional mean imputation.

Note that in a given incomplete data set, missing and observed variables can have different factor structures. If this "factor structure shift" is the case, then factors obtained from the complete portion of the data set X may not have the desired predictive value for the missing-block Y^m. One can expect higher R^2s as the factors that missing and observed series load on get closer to each other and vice versa[27]. In this sense, half loadings case is an extreme version of factor structure shift where missing and observed variables do not share any common factors.

Factor structure shift can be encountered in many different settings. For example, introduction of exchanges and collection of most time series in many emerging markets did not start until mid-80's, while many time series started to be collected much earlier in most developed economies. If time series from these two different country groups are combined (say to form a global factor model), the resulting data set would be unbalanced. In addition, emerging markets contain their own country and regional factors that cannot be estimated from the data of the developed economies. In this case if imputation is carried out by some conditional expectation method using available series, then the initial factors estimated from developed economies data would inaccurately impute the missing values in the emerging market data. As a result, the resulting factor estimates may not be consistent.

Another situation where factor structure shift phenomenon can be observed is the incorrect estimation of the number of factors. In real data applications, the number of factors is not known and has to be estimated from the data. Although there are consistent estimators for r (eg Bai and Ng, 2002), in small sample the estimate of r can be inaccurate, and this results in higher estimation error. In this case imputation equation would suffer from either omitted variable bias (if $r > k$) or irrelevant variable noise (if $r < k$), and ID estimators would fail to be consistent.

Asymptotic Normality Figure 1.5 below demonstrates the limiting distributions of standardized (with the inverse square root of asymptotic variance estimate) INF, CID, DID and RID[28] factor estimators for $r = 1$, $N(30)$, $T(30)$ and $MR = 30\%$. As can be seen in the figure, INF estimator

[27]One should also take into account the factors that are present only in observed variables. Inclusion of these factors would contaminate the imputation and lead to lower R^2.

[28]Note that the result for UID is very similar to that of CID. In order preserve space we do not present it here.

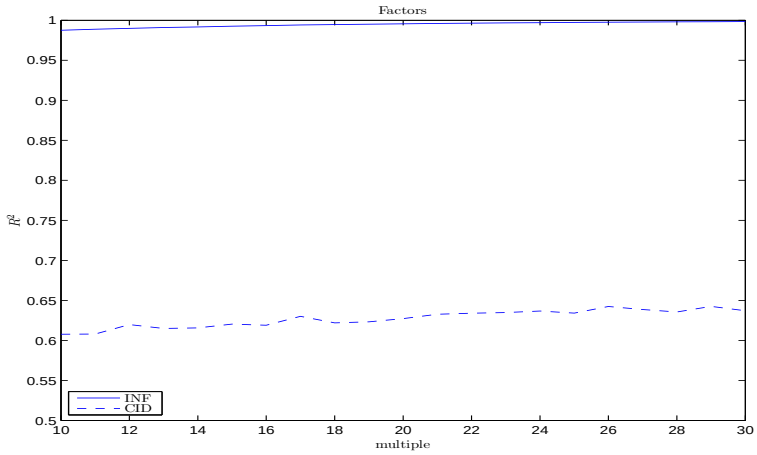

Figure 1.4: R^2 of INF and CID factor estimators with $r = 2$ and half-loadings (MR=50%)

is standard normal even for small N and T as suggested by Bai (2003). In addition, DID and RID estimators look close to standard normal, CID is certainly not. Although visually it is hard to tell whether CID is normal, Table 1.7 below, which reports the first four moments of estimators and their p-values for Jarque-Bera (JB) normality test, shows that CID estimator is normal (up to 14 % significance level). However, its asymptotic variance is much higher than 1.

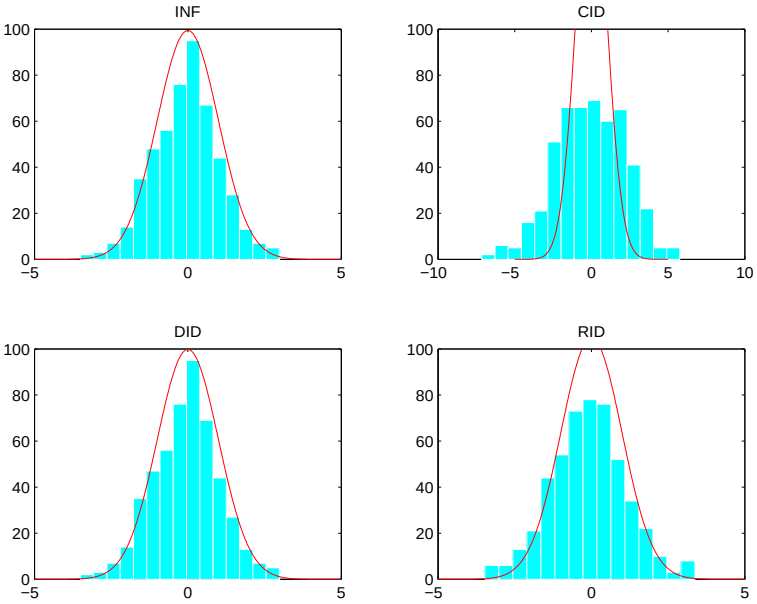

Figure 1.5: Histogram of all estimated factors for $N(30)$ and $T(30)$ with 30% missingness.

The histograms in Figure 1.5 (and other figures below) are obtained as follows: For a given

	INF	CID	DID	RID
$mean$	-0.06	-0.11	-0.06	-0.11
$variance$	1.11	7.80	1.16	1.83
$skewness$	-0.05	-0.02	-0.05	0.05
$kurtosis$	3.39	2.80	3.40	3.40
$JB\ pvalue$	0.19	0.14	0.17	0.18

Table 1.7: Moments and Normality for Figure 1.5

$(T(j), N(j))$ pair we generate 500 data sets, and estimate factors and loadings. Let $t = [T/2]$ where $[x]$ denotes the largest integer smaller than or equal to x. For ID estimators and the INF estimator for each data set $d = 1, 2..., 500$, we calculate the following scaled estimators, respectively:

$$\tilde{\tilde{f}}_t = \tilde{\tilde{\Pi}}_t^{-1/2}\sqrt{N}(\tilde{\tilde{F}}_t - H'F_t)$$

$$\tilde{f}_t = \tilde{\Pi}_t^{-1/2}\sqrt{N}(\tilde{F}_t - H'F_t)$$

where $\tilde{\tilde{\Pi}}_t$ and $\tilde{\Pi}_t$ are the estimated asymptotic variances for ID and INF estimators discussed in Section 1.6. The histograms are generated by using the samples $f_t(d)$. For INF estimators, this statistic is asymptotically standard normal. In order to check the normality of the estimators visually, we overlay the density of N(0,1) on each histogram. Note that the area under the histogram bars is *not* equal to 1. In order to obtain comparable results, we multiply the values of the normal density with the total area of the histogram bars.

Figure 1.5 shows that scaled asymptotic distribution of CID estimator is not standard normal. Therefore, the asymptotic variance result shown in Bai (2003) seizes to be applicable when factor estimates are obtained from imputed data panels. The intuition behind this result is clear: Common component imputation (although consistent) contains estimation error which leads to an increase in CID factor estimator's asymptotic variance. Table 1.7 verifies that DID and RID estimators are normal and their asymptotic variances are much closer to 1 than that of CID. The fact that RID behaves like a standard normal variable asymptotically is very interesting. It indicates that even if the number of irrelevant imputed values in the data set increase (they become infinite in the limit), as their relative size shrinks they seize to affect the limiting distribution of the factor estimator. In this sense, the RID estimator serves as a test case: It shows that in general any decreasing-block ID estimator is useless for statistical inference since they behave the same way with the complete-data estimator irrespective of the values used in imputation. Accordingly, we can ignore the results for the DID estimator, as well. Therefore, below we restrict our attention only to INF and CID estimators.

Table 1.8 reports the moments, normality test and χ^2 test results for results CID factor estimator for different missingness ratios and multiples. As can be seen in the first column, with 10% missingness CID estimator is normal even for a reasonably small multiple (j=10). However, the second column of the table shows that when missingness is increased to 30% at the same N and T values, it seizes to be normal and its asymptotic variance increases significantly. Finally, we see in the third column of the table that as data dimensions are increased to $N(35)$ and $T(35)$, CID estimator becomes normal again. These results show that the CID estimator is asymptotically normal. However, it takes larger T and N values to demonstrate the normality of the CID estimator for higher missingness ratios.

Note that the last row of Table 1.8 reports the p-values of the χ^2 test for the asymptotic variances of CID factor estimator. Our null hypothesis is $H0 : \sigma^2 = 1$, where σ^2 denotes the variance of CID estimator. The alternative hypothesis is $H1 : \sigma^2 > 1$. As can be seen in the table, null hypothesis is always rejected for CID at 5% level. Therefore together with the normality test results above, we may conclude that for large N and T standardized CID factor estimator is normally distributed with a variance greater than 1.

	$MR = 10\%,\ j = 10$	$MR = 30\%,\ j = 10$	$MR = 30\%,\ j = 35$
mean	0.02	0.05	0.01
variance	1.14	6.87	4.53
skewness	0.10	-0.21	0.18
kurtosis	3.18	3.53	2.74
JB pvalue	0.49	0.02	0.24
χ^2 *pvalue*	0.04	0	0

Table 1.8: Moments and Normality for missingness ratio 10% and 30% with multiples 10 and 35

Finally, we investigate the relation between the level of missingness and the asymptotic variance of the CID estimator and summarize the results in Figure 1.6, which shows that the asymptotic variance of CID exponentially increases with missingness. The intuition behind this result can be explained with the following example: Let a and m denote the available and missing values in a data panel. By normalizing the number observations in the data panel to 1, one can write $a + m = 1$. Then the proportion of number of missing to observed values is expressed by $m/a = \frac{m}{1-m}$. Table 1.9 below illustrates the values of m/a as a function of m. Since in the imputation procedure observed values are used to estimate the missing ones, one can argue that the estimation error of the imputation follows a pattern very similar to that of m/a[29].

[29]This illustration is very simplistic in the sense that it does not address any statistical properties in the data (e.g. the

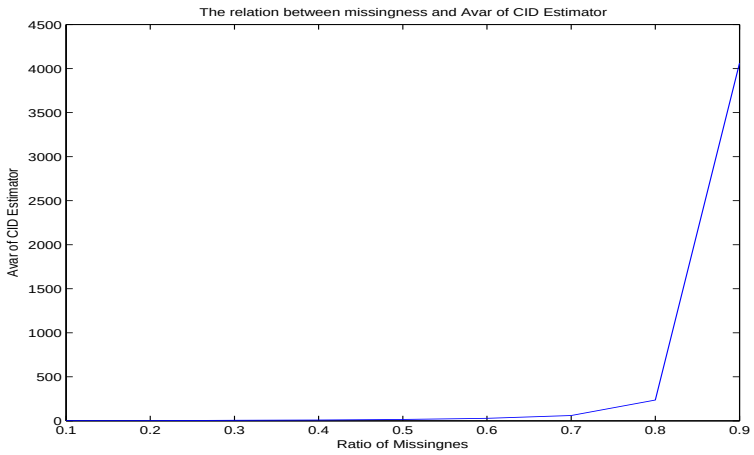

Figure 1.6: Asymptotic variance of CID factor estimator vs. MR ($N(20)$ and $T(20)$).

m	.1	.2	.3	.4	.5	.6	.7	.8	.9
$\frac{m}{1-m}$	0.11	0.25	0.43	0.67	1.00	1.50	2.33	4	9

Table 1.9: The non-linear change of m/a with m

Partitioned Factors and Loadings

Above we discuss consistency and asymptotic normality of the ID estimators. We show that the CID estimator is consistent (unless there is a factor shift) and asymptotically normal with an asymptotic variance higher than the one predicted by the standard large sample theory. The CID estimator above is obtained from a data set whose northwest corner is imputed. As we show in section 1.5.1 above, missingness divides the data set into four regions, and partition factors as $F = [F^{m'}\ F^{a'}]'$. In section 1.7.1 we treat these different portions as *one*, and obtain the large sample results for the *whole* factor estimators. Therefore it is natural to ask whether the imputation affects missing and available parts of factor and loading estimators differently. To that end in this section we treat missing and available parts of estimated factors as separate estimators and investigate their large sample properties. Below we denote missing and available parts of the CID and INF factor estimators with CIDM, CIDA, INFM and INFA, respectively[30].

level of dependence across variables, missing data mechanism etc.), but it still provides an intuition for the observation that asymptotic variance of CID estimator increases non-linearly with missingness.

[30]In this section we show the results for the CID factors. Results for CID loadings are very similar and are not presented to save space

Consistency Figure 1.7 shows the R^2 of INFM, INFA, CIDM and CIDA factor estimators obtained from a two-factor model with 60% missingness ratio. R^2s of all the estimators monotonically converge to 1 as N and T increase. Therefore all estimators consistently estimates their respective factor spaces. We have the following observations:

i. INFM and INFA converge at the same rate.

ii. Both INF estimators converge faster than CID estimators

iii. CIDA converges faster than CIDM

As can be seen, different portions of INF estimator have the same consistency properties, and both have higher convergence rates than their respective CID estimators. Additionally, the fact that CIDA converges faster than CIDM is an indication that missingness affects CIDM more severely that it does CIDA.

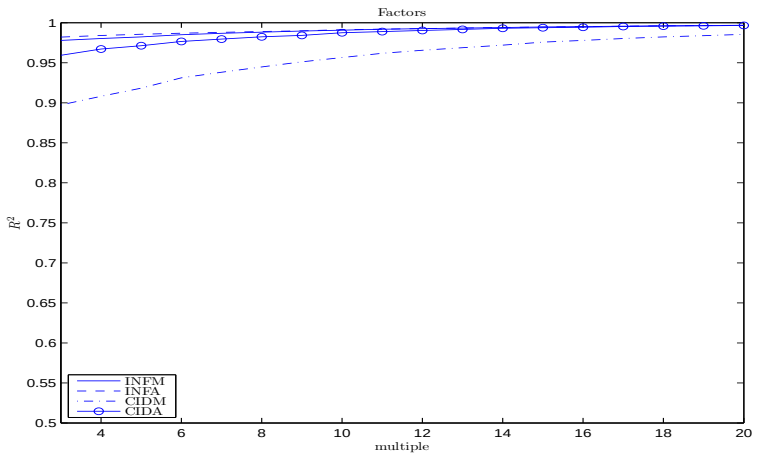

Figure 1.7: R^2 of CIDM and CIDA factor estimators with $r = 2$ and $MR = 60\%$

We also find that the results listed above becomes stronger when missingness ratio is increased. Therefore one can conclude that the level of missingness is an element in determining the convergence rates of both CIDM and CIDA estimators. Additionally, one can infer that missingness ratio affects the convergence rate of CIDM more than that of CIDA[31]

[31] Additionally, we find that the above results gets stronger with the number of true factors. This observation can be explained as follows: As the number of factors increase, errors in factor estimation and imputation propagate. These additional noise leads to the same effect with increased missingness ratio.

Asymptotic Normality Figure 1.8 demonstrates the limiting distributions of standardized INFM, INFA, CIDM and CIDM estimators. The figure yields the following observations:

i. INFM and INFA are asymptotically standard normal

ii. CIDA looks asymptotically standard normal

iii. CIDM seems normal (but not standard normal)

The figure reveals an important feature about the factor estimators obtained under missingness: Although different parts of the true factors have the same asymptotic distribution, their CID counterparts have different distributions. Therefore, even if the true factors are stationary, their CID estimators are non-stationary. On the other hand, both portions of the INF estimator have the same limiting distribution; hence INF estimators preserve the stationarity property of the true factors.

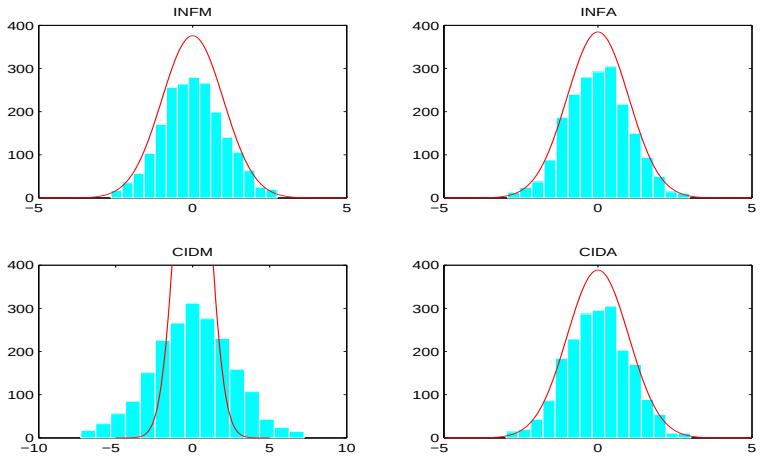

Figure 1.8: Histogram of INFM, INFA, CIDM and CIDA factor estimators with $r = 2$ and $MR = 35\%$

Above we say that both CIDM and CIDA factor estimators look normal. It is hard to verify this claim with histograms alone. To that end, Table 1.10 reports the first four moments, Jarque-Bera and variance test p-values for CIDM and CIDA estimators for 15% and 35% missingness, and with multiples 20 and 30. Note that in χ^2 test, null hypothesis is $H0 : \sigma^2 = 1$, and the alternative hypothesis is $H1 : \sigma^2 > 1$. As can be seen in first two columns of the table, for a mild missingness ratio (15%) both CIDM and CIDA are normal, and asymptotic variance of CIDA is not significantly different from 1. When missingness ratio is increased to 35% keeping N and T at the same level, CIDM seizes to be normal, and CIDA's variance differs from unity. Finally, increasing the multiple to

30 at the same missingness ratio results in (columns 5 and 6) a normal CIDM and a standard normal CIDA. Therefore this table shows that both CIDM and CIDA are normal; however it takes higher T and N values to establish when missingness ratio is high. Note that although the table assures us that missing and observed parts of CID estimator are normal, it does not clarify whether CIDA is standard normal like INFA.

	$MR = 15\%,\ j = 20$		$MR = 35\%,\ j = 20$		$MR = 35\%,\ j = 30$	
	$CIDM$	$CIDA$	$CIDM$	$CIDA$	$CIDM$	$CIDA$
mean	-0.03	0.00	-0.08	-0.02	-0.03	-0.00
variance	2.94	1.05	7.73	1.12	6.77	1.04
skewness	-0.01	0.03	0.09	0.02	-0.09	0.01
kurtosis	3.13	2.94	3.31	2.98	3.20	3.07
JB *pvalue*	0.50	0.50	0.01	0.50	0.06	0.50
χ^2 *pvalue*	0	0.21	0	0.03	0	0.26

Table 1.10: Moments and Normality for CIDM and CIDA

Figures 1.9 and 1.10 show the asymptotic variance of CIDM and CIDA factor estimators for multiples of 10, 20 and 30 with 60% missingness ratio. As can be seen in the figures, asymptotic variance of CIDM and CIDA differs significantly; missingness affects the asymptotic variance of CIDM much more severely than it does that of CIDA. In addition, asymptotic variance of both estimators becomes smaller as the data dimensions get larger. Finally, the asymptotic variance of CIDA stays very close to unity for a wide range of missingness ratios, but it eventually increases above it. However it is still ambiguous whether CIDA factor's asymptotic variance would be above 1 for very high N and T values (that we cannot achieve with a simulation study). Therefore, although the evidence is clear in the case of CIDM, we can not clearly state whether the asymptotic variance of CIDA is significantly different from 1. To that end, in our second paper we try to address these issues by deriving the large sample properties of CIDM and CIDA analytically.

To sum up, in this section we partition the CID estimators into their missing and available components, and investigate their large sample properties. We show that both CIDM and CIDA factor estimators are consistent and asymptotically normal. This is one of the central findings of this study; it indicates that when data is imputed with FBI the desirable distributional properties suggested by the standard large sample theory continue to hold in. In other words, FBI preserves the consistency and asymptotic normality properties of partitioned estimators.

Additionally, we find the intuitive result that convergence rates and the asymptotic variances of missing and available parts are determined by the amount of missingness a given partition is exposed

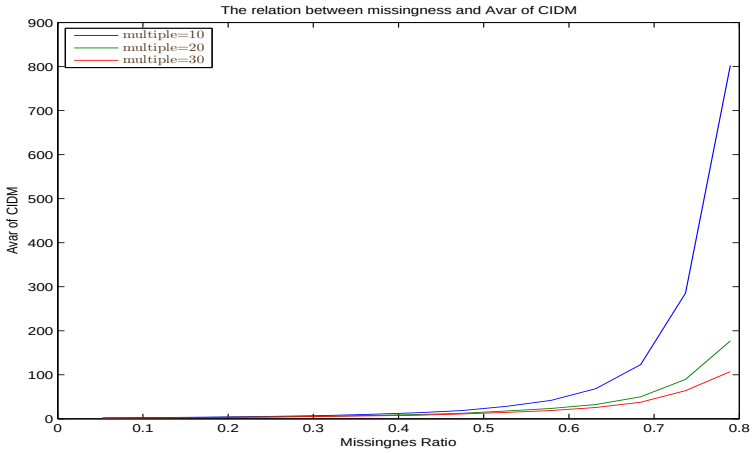

Figure 1.9: Asymptotic variance of CIDM factor estimator vs MR for multiples 10, 20 and 30

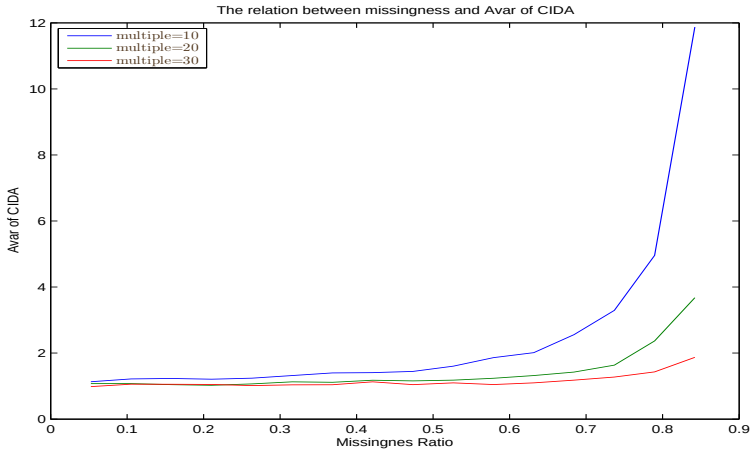

Figure 1.10: Asymptotic variance of CIDA factor estimator vs MR for multiples 10, 20 and 30

to. In this respect, since CIDM is exposed to mode missingness than CIDA, its convergence rate is lower and asymptotic variance is higher.

The findings above point out that the results in Section 1.7.1 are actually a combination of the results for missing and available partitions; thus results regarding the large sample distribution of *whole* CID estimators reflect some weighted average of CIDM and CIDA. In this sense, in order to unveil the true nature of the CID estimators and to describe their asymptotic behaviors, one needs to consider the partitioned estimators rather than the whole ones. Therefore, for the simulation results regarding monotone-missing data sets below we consider only the partitioned estimators.

1.7.2 Monotone Missingness

This section investigates the large sample behavior of partitioned CID factor estimators under 2-layer monotone missingness. The two layers of missingness defines three partitions for the CID factor: $CIDM_1$ which is the part of CID from the cross-section observation $t = 1$ to $t = T_{m_1}$, $CIDM_2$ which is the part of CID from $t = T_{m_2}$ to $t = T_{m_1}$, and $CIDA$ which is the part of CID from $t = T_{m_1} + 1$ to $t = T$. Following Section 1.7.1 above, we consider consistency and normality properties of the partitioned factor estimators, and attempt to determine the relation between their asymptotic variance and the ratio of missingness.

Consistency Figure 1.11 shows the R^2 for $CIDM_1$, $CIDM_2$ and $CIDA$ factor estimators with $r = 1$ and 60% missingness. As can be seen on the figure, $CIDA$ converges to its latent true counterpart the fastest, which is followed by $CIDM_1$, and $CIDM_2$ converges the slowest. On the other hand, as in the case of block-missing data, each partitioned INF estimator has the same convergence rate (not presented on the figure). The intuition behind this result can be explained as follows: For a given cross section of $CIDM_2$, $CIDM_1$ and $CIDA$, the ratio of missingness of the corresponding cross-section of $\hat{X}$ is $(N_{m_1} + N_{m_2})/N$, N_{m_1}/N and 0, respectively. A higher missingness ratio leads to a larger the imputation error which in turn results in larger estimation error in the corresponding partition of the estimated factor. Finally, a larger estimation error in factor estimation leads to slower convergence.

We also compare the convergence rate of $CIDA$ to that of INF estimator. It is observed that $CIDA$ estimator converges to its true counterpart slower than INF estimator. This observation indicates that missingness affects the convergence rate of $CIDA$ estimator even though X has no missing values corresponding to the cross-sections of $CIDA$.

Asymptotic Normality Figure 1.12 demonstrates the limiting distributions of $CIDM_1$, $CIDM_2$ and $CIDA$ factor estimators with $r = 1$ and 60% missingness. As can be seen, all three estimators look normal. In addition, $CIDA$ looks standard normal whereas $CIDM_2$ does not; one can argue both ways for $CIDM_1$.

In order to asses the distributional properties of the $CIDM_1$ and $CIDM_2$ estimators better, we calculate their major moments, and Jarque-Bera and χ^2 variance test p-values for different missingness levels and data dimensions shown in Table 1.11. The null and alternative hypotheses for Jarque-Bera and the χ^2 tests are $H_0 : \sigma^2 = 1$ and $H_1 : \sigma^2 > 1$, respectively, where σ^2 is the variance of the factor estimator under consideration. χ^2 test indicates that neither estimator is standard normal even

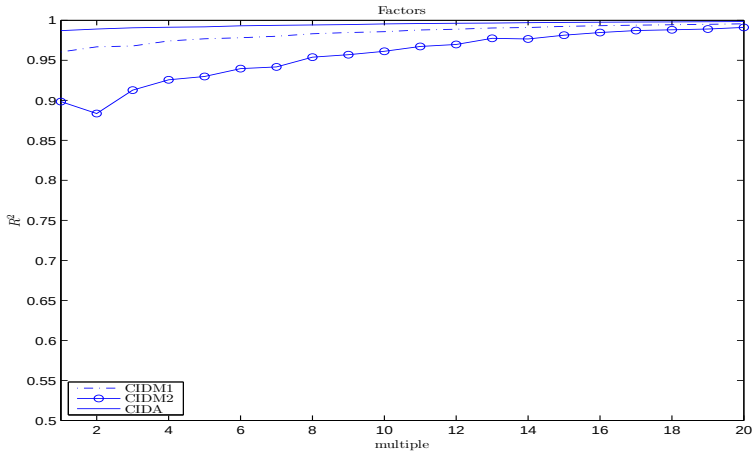

Figure 1.11: R^2 of $CIDM_1$, $CIDM_2$ and $CIDA$ factor estimators with $r = 2$ and $MR = 60\%$

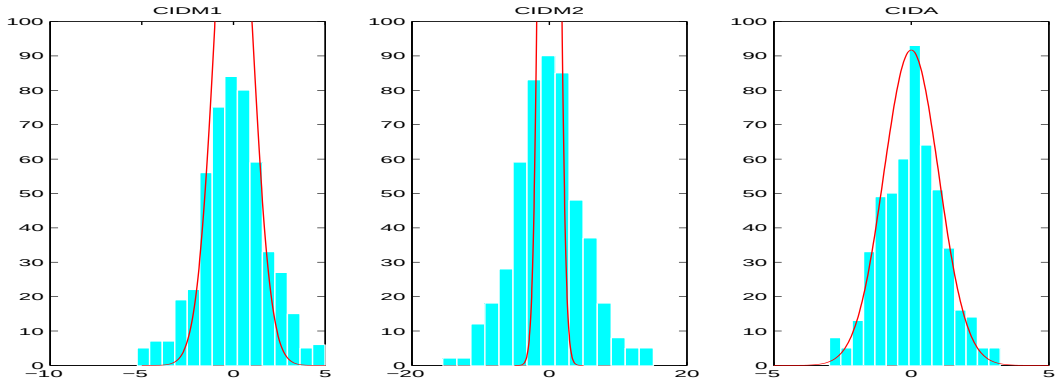

Figure 1.12: Histogram of $CIDM_1$, $CIDM_2$ and $CIDA$ factor estimators with $r = 1$ and $MR = 60\%$

for small missingness levels. For 10% missingness both estimators are normal even for a multiple of $j = 10$, but $CIDM_2$ seizes to be normal when the missingness is increased to 30%. However, when the multiple is increased the $j = 30$ keeping the missingness at 30%, $CIDM_2$ becomes normal again. These results show that both estimators are normal, and for higher levels of missingness one needs to increase the data dimensions further to obtain normality. Note that $CIDA$ factor estimator is very close to standard normal for all the cases considered (not shown in the table). As in the case of block missingness, this raises the question whether the partition of the factor estimator corresponding to the available part of X is asymptotically standard normal as suggested by the standard large sample theory. In our second paper, we will answer this question by calculating the asymptotic variance of $CIDA$ analytically.

We also consider the relation between missingness ratio and the asymptotic variance for $CIDM_1$,

	$MR = 10\%$, $j = 10$		$MR = 30\%$, $j = 10$		$MR = 30\%$, $j = 30$	
	$CIDM_1$	$CIDM_2$	$CIDM_1$	$CIDM_2$	$CIDM_1$	$CIDM_2$
mean	-0.02	0.11	0.15	0.07	0.01	-0.14
variance	1.77	2.68	2.49	9.80	2.14	7.31
skewness	-0.03	0.02	-0.07	0.01	0.08	-0.07
kurtosis	3.39	3.20	3.12	3.52	3.00	2.68
JB pvalue	0.20	0.50	0.50	0	0.50	0.24
χ^2 *pvalue*	0	0	0	0	0	0

Table 1.11: Moments and Normality for $CIDM_1$ and $CIDM_2$

$CIDM_2$ and $CIDA$ factor estimators. Our findings are very similar to those of the block missingness case: Asymptotic variance increases with missingness for all three estimators. In addition, keeping the missingness level constant one obtains a smaller variance as N and T are increased. Furthermore given the levels of missingness and N and T, the asymptotic variance of $CIDM_2$ is higher than that of $CIDM_1$, which in turn is higher than the variance of $CIDA$. These results indicate that missingness affect different partitions of the CID estimators differently. Basically, the higher the cross-section missingness of a partition, the larger its asymptotic variance.

One may be tempted to generalize these results to the case of k-layer monotone missingness, where k is an arbitrary number. In this case we obtain $k + 1$ partitions for the CID estimator, namely $CIDM_1, ..., CIDM_k$ and $CIDA$. Then one would expect $CIDM_k$ to have the lowest convergence rate, and highest asymptotic variance (and higher than 1) for any missingness ratio since it has the highest cross-section missingness. By the same token, $CIDA$ would have highest convergence rate, and lowest asymptotic variance. In Chapter II we derive analytical expressions for CIDA and CID_j, $j = 1, .., k$ and show how the level of exposure to missingness determines their distributional properties.

1.8 Conclusion

Many economic and financial institutions, including Federal Reserve, credit rating agencies, investment banks and hedge funds study the empirical properties of the factor models. The methods developed and the results obtained in this study can have important implications since applications of factor models in recent empirical macroeconomics and finance literature are rapidly rising.

There are many factor model applications in recent literature that extracts factor and loading estimates from imputed data sets; for example Stock and Watson (2002a,b), Bernanke and Boivin (2003), Bernanke et al. (2005) impute missing data sets with the EM algorithm. However, these studies ignore the effect of imputation error on their estimates and forecast which leads to biased

results, underestimated standard errors and invalid hypothesis tests. Our study offers an insight to the behavior of the factor estimators that are obtained from the imputed data sets.

This chapter develops an intuitive and parsimonious factor-based imputation method (FBI) aimed for data sets that suffer from monotone missingness and assumes a factor structure. We show that the large sample theory for factor models under missingness should be confined to the CID estimators (obtained under FBI) since other ID estimators for which missingness increases slower than the data panel yield asymptotics that cannot be used for inference purposes.

In addition, we demonstrate that CID estimators are consistent and asymptotically normal unless there is a factor structure shift. We also extend this result to partitioned factor estimators CIDA and $CIDM_j$. These findings indicate that FBI preserves the desirable distributional properties of consistency and asymptotic normality found under standard complete data framework.

Finally, we find the intuitive result that the distributional behavior of a given partitioned factor estimator is determined by that partition's exposure to missingness; as the missingness exposure increases the convergence rate decreases and the asymptotic variance increases.

Note that ignoring estimation error impairs inference by leading to underestimated standard errors and low-power hypothesis tests. The fact that estimation error in imputation is ignored in many academic and applied work is an important shortcoming. In this respect, this chapter provides a new perspective to the study of factor model estimators under missingness. In the next chapter, we continue our quest with more attention to internal dynamics of FBI and the distributional properties of the imputed values.

2

FBI ALGORITHM UNDER MONOTONE MISSINGNESS: THEORETICAL RESULTS

2.1 Introduction

In the previous chapter we studied the large sample properties of the estimated factors under monotone missingness. We developed an intuitive and parsimonious FBI algorithm, and showed that it preserves the desirable distributional properties implied by standard complete-data procedures for the factor model estimators extracted from imputed data. We also showed that the asymptotic properties of the CID factor partition estimators are determined by the missingness level a given partition is exposed to. In this respect, we found that CID partition estimators that are exposed to missingness more converge slower and have a higher asymptotic variance.

This chapter continues the study of large dimensional factor models under missing data focusing on the workings and main statistical properties of FBI algorithm, and the distributional properties of the imputed values obtained under FBI. First, we delve into the details of FBI under block and monotone missingness cases, and express imputation error in terms of assumed factor structure and estimated factor model variables.

Next, we derive the statistical properties of the auxiliary (interim) factor and loading estimators that constitute the backbone of the FBI algorithm and show that they are both consistent and asymptotically normally distributed. By the help of these findings, we derive the asymptotic distribution of the imputation error under FBI, and show that it is asymptotically normally distributed.

Finally, we express the CID factor estimators in terms of the observed and imputed components of the completed data set $\hat{X}$. We show different sources of variation for partitioned estimators, and how these differences can explain the observed convergence rate and asymptotic variance differences in Chapter I. We also show that when missingness vanishes (i.e. as MR=0), CID estimators become equal to (standard complete-data) INF estimators. This result verifies our derivations for CID factor estimators and shows that CID estimators can be seen as a proper generalization of the INF estimators.

The organization of the paper is follows: Section 2.2 discusses the structure of the factor models under block and monotone missingness in more detail than Chapter I. Section 2.3 derives asymptotic properties of auxiliary factor and loading estimators, and Section 2.4 obtains the large sample distri-

bution of the imputation error under FBI. Section 2.5 shows the link between CID factor partitions and observed and imputed data blocks. Section 2.6 considers the limitations of the current study and suggests directions for future research. Proofs for propositions, theorems and corollaries that we present in the text are provided in the appendices.

2.2 Analytics of FBI under Missingness

In Chapter I we introduced the basics of the FBI procedure for block and monotone missing data sets to show how they are applied in the simulation experiment. In this section we go deeper into the workings of the algorithm, discuss how missing values are estimated in detail, and express imputed data values in terms of the algorithm variables and parameters. The statistical properties of the main components of the FBI algorithm and the imputed values studied below will be derived in the following section.

2.2.1 Block Missing Data

We assume that the incomplete data matrix $X = [Y\ Z]$ given in Table 2.1 admits a factor structure with r factors[1].

	N_m	N_a
T_m	Y^m	Z^m
T_a	Y^a	Z^a

Table 2.1: Block-missing data panel

As we discussed in Chapter I, since X has missing values, factor model estimates cannot be directly obtained. In order estimate them one should first convert the unbalanced data set at hand to a balanced one. To this end, we show how to utilize the FBI procedure for block-missing data sets in detail. As we

[1]The discussion in this section is a continuation of the FBI procedure under block-missingness considered in our first paper.

mentioned in the previous chapter, the FBI algorithm imputes the missing values Y^m with the estimate of its common component, $\tilde{F}^m\hat{\Lambda}^{m\prime}$. Note that Y^m and Y^a have the same loadings, Λ^m. If we knew the true factors F, we could have obtained the OLS estimate of Λ^m by regressing Y^a on F^a. Fortunately, Bai and Ng (2006) and Stock and Watson (2002a) show that under certain assumptions regressing Y^a on F^a yields the same results (in terms of common component estimation) with regressing it on $\tilde{F}^a$ as $N, T \to \infty$. However, since X has missing values, $\tilde{F}$ is infeasible. In this case, another estimate of F that is obtained from the observed portion of X can be used as a proxy for $\tilde{F}$. In this respect, let $\tilde{F}_z$ denote factor estimate obtained from the (complete) partition Z of X. As long as $\tilde{F}_z$ satisfies the conditions under which Bai and Ng (2006) and Stock and Watson (2002a) obtain their results, it can be used to estimate Λ^m. In the next section we show that $\tilde{F}_z$ is consistent and asymptotically normal, and it can be used as a proxy for F for imputation purposes.

Since Z is completely observed, $\tilde{F}_z$ is simply $\sqrt{T}$ times the eigenvectors of the matrix ZZ' corresponding to the first r eigenvalues of ZZ' in decreasing order. By construction, we have have $\tilde{F}_z'\tilde{F}_z/T = I$.

$$\underline{Y}_i^a = \tilde{F}_z^a\lambda_i + \nu_i \quad , \quad i = 1, 2, ..., N_m \tag{2.1}$$

where $T_a \times 1$ vector $\underline{Y}_i^a$ is the observed part of the i^{th} series (column), $r \times 1$ vector λ_i is the unobserved loading for variable i, and $\mathrm{T}_a \times 1$ vector ν_i is the error term. Carrying out the time series regression in (2.1), we obtain

$$\hat{\lambda_i} = (\tilde{F}_z^{a\prime}\tilde{F}_z^a)^{-1}\tilde{F}_z^{a\prime}\underline{Y}_i^a \tag{2.2}$$

Letting $\tilde{K} = (\tilde{F}_z^{a\prime}\tilde{F}_z^a)^{-1}\tilde{F}_z^{a\prime}$ we have

$$\hat{\lambda_i} = \tilde{K}\underline{Y}_i^a$$

We can run the above regression for each series in Y^a, and obtain the estimate of the $r \times N_m$ matrix $\hat{\Lambda}^{m\prime} = [\hat{\lambda_1}, \hat{\lambda_2}, \ldots \hat{\lambda}_{N_m}]$. We can express $\hat{\Lambda}^m$ simply as

$$\begin{aligned}
\hat{\Lambda}^{m\prime} &= [\hat{\lambda}_1\ \hat{\lambda}_2 \ldots \hat{\lambda}_{N_m}] \\
&= [\tilde{K}\underline{Y}_1^a\ \ \tilde{K}\underline{Y}_2^a \ldots \tilde{K}\underline{Y}_{N_m}^a] \\
&= \tilde{K}[\underline{Y}_1^a\ \ \underline{Y}_2^a \ldots \underline{Y}_{N_m}^a] \\
&= \tilde{K}Y^a
\end{aligned} \tag{2.3}$$

Using (2.3), the prediction of Y^m is obtained as

$$\begin{aligned}\hat{Y}^m &= \tilde{F}_z^m \hat{\Lambda}^{m\prime} \\ &= \tilde{F}_z^m (\tilde{F}_z^{a\prime} \tilde{F}_z^a)^{-1} \tilde{F}_z^{a\prime} Y^a \\ &= \tilde{P} Y^a \end{aligned} \tag{2.4}$$

where $\tilde{P} = \tilde{F}_z^m (\tilde{F}_z^{a\prime} \tilde{F}_z^a)^{-1} \tilde{F}_z^{a\prime}$ is a $T_m \times T_a$ projection matrix. Replacing Y^m with $\hat{Y}^m$ we obtain the completed data matrix $\hat{X}$ as follows:

$$\hat{X} = \begin{bmatrix} \hat{Y}^m & Z^m \\ Y^a & Z^a \end{bmatrix} = \begin{bmatrix} \tilde{P} Y^a & Z^m \\ Y^a & Z^a \end{bmatrix}$$

An alternative way to see characterize $\hat{X}$ is by introducing the $T \times N$ difference matrix, $\hat{D}$:

$$\hat{D} = \begin{bmatrix} \hat{Y}^m - Y^m & 0 \\ 0 & 0 \end{bmatrix}$$

Using $\hat{D}$ we can link X and $\hat{X}$ as follows:

$$\begin{aligned}\hat{X} &= \begin{bmatrix} Y^m & Z^m \\ Y^a & Z^a \end{bmatrix} + \begin{bmatrix} \hat{Y}^m - Y^m & 0 \\ 0 & 0 \end{bmatrix} \\ &= X + \hat{D} \end{aligned} \tag{2.5}$$

Using (2.5) and the definition of $\hat{D}$, we can write

$$\hat{X}_t = \begin{cases} X_t + \hat{D}_t & \text{if } t \in M \\ X_t & \text{if } t \in A \end{cases} \tag{2.6}$$

The (i,t) element of $\hat{Y}^m$ and $\hat{D}$ are given as follows:

$$\hat{Y}_{it} = \tilde{F}_{zt}' \hat{\lambda}_i \quad , \quad \text{for} \quad i,t \in M$$

$$\hat{D}_{it} = \begin{cases} \hat{Y}_{it} - Y_{it} = \hat{d}_{it} - e_{it} & , \quad \text{for} \quad i,t \in M \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

where $\hat{d}_{it} = \tilde{F}_{zt}' \hat{\lambda}_i - F_t' \lambda_i$.

2.2.2 Monotone Missing Data

Consider a data set[2] $X = [Y\ Z]$ with J missing layers where $Y = [Y^1\ .\ .\ .\ Y^J]$, $Y^j = [Y^{a_j\prime}\ Y^{m_j\prime}]\prime$, $j = 1, ..., J$. The missing portion of Y is given by $Y^m = \{Y^{m_1}, ..., Y^{m_J}\}$, where each component is $T_{m_j} \times N_{m_j}$, and J available layers $Y^a = \{Y^{a_1}, ..., Y^{a_J}\}$, each of which is $T_a \times N_{m_j}$ where $T = T_{m_1} + T_a$ and $N = N_{m_1} + ... + N_{m_j} + N_a$ for $j = 1, ..., J$. Suppose X has J-layer monotone missingness, and exhibits the factor structure described in Section 2.2. Following the partition of X, we can partition Λ and e as follows:

$$\Lambda = [\ \Lambda^{m_1'}\ \Lambda^{m_2'}\ ...\ \Lambda^{m_J'}\ \Lambda^{a'}\ ]'$$

$$e = [\ e^{y_1}\ e^{y_2}\ ...\ e^{y_J}\ e^{z}\ ]$$

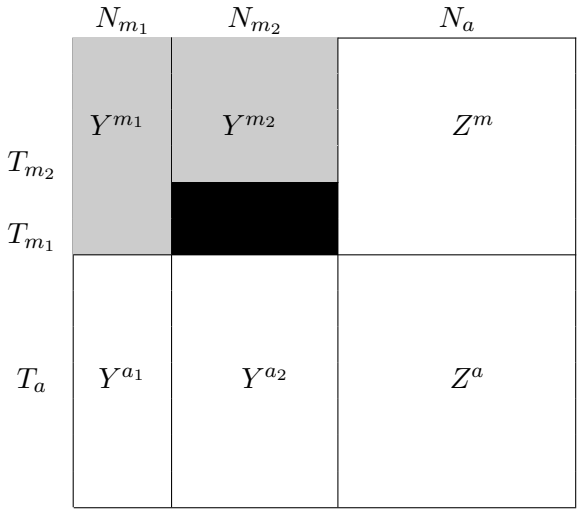

Table 2.2: 2-layer monotone-missing data panel

where Λ^{m_j} is $N_{m_j} \times r$, and e^{y_j} is $T \times N_{m_j}$. Let $S_c = \sum_{j=1}^{c} N_{m_j}, c = 1, 2, ...J$. Then the number of completely observed cross-section and time series observations are given by $N_a = N - S_J$. Partitioning F for each missing layer is a bit tricky when there is more than one missing layer since each layer has overlapping factor partitions. In this respect, we define the $T_{m_j} \times r$ matrix F^{m_j} as the portion of F that runs from observation $t = 1$ to $t = T_{m_j}$ for $j = 1, 2..., J$. Let $t = 1, ..., T_{m_j}$ and $i = 1, ..., N_{m_j}$ be abbreviated as $t \in M_j$ and $i \in M_j$, respectively. Similarly let $t \in A_j$ denote *available* indices[3] such

[2]For more introductory material on monotone missingness, please refer to section 1.5.2 in Chapter I

[3]Below we denote the index of completely observed samples by $A_1 = A$.

that $t = T_{m_j} + 1, ..., T$. The missing and available layers of X can be expressed in vector form as the following:

$$Y_t = \Lambda^{m_j} F_t + e_t^{y^j} \quad , \quad t \in M_j$$

$$Y_t = \Lambda^{m_j} F_t + e_t^{y^j} \quad , \quad t \in A_j$$

$$Z_t = \Lambda^{a} F_t + e_t^{z} \quad , \quad t = 1, ..., T$$

Accordingly, available and missing parts of Y and Z are given in matrix form as follows:

$$Y^{m_j} = F^{m_j}(\Lambda^{m_j})' + e^{y^{m_j}} \tag{2.8a}$$

$$Y^{a_j} = F^{a}(\Lambda^{m_j})' + e^{y^{a_j}} \tag{2.8b}$$

$$Z^{m} = F^{m_1}(\Lambda^{a})' + e^{z^{m}} \tag{2.8c}$$

$$Z^{a} = F^{a}(\Lambda^{a})' + e^{z^{a}} \tag{2.8d}$$

Following the exposition in section 2.2.1, we impute each missing block Y^{m_j} with the estimate of its common component, $\tilde{F}_z^{m_j}(\hat{\Lambda}^{m_j})'$. Consider estimation of the loadings for the j^{th} missing block:

$$\underline{Y}_i^{a_j} = \tilde{F}_z^{a} \lambda_i^j + \nu_i^j \quad , \quad i = 1, 2, ..., N_{m_j} \quad , \quad j = 1, 2, ...J \tag{2.9}$$

Carrying out the time series regression in (2.9), we obtain

$$\hat{\lambda}_i^{m_j} = (\tilde{F}_z^{a\prime} \tilde{F}_z^{a})^{-1} \tilde{F}_z^{a\prime} \underline{Y}_i^{a_j} \tag{2.10}$$

$$= \tilde{K} \underline{Y}_i^{a_j} \tag{2.11}$$

where $\tilde{K} = (\tilde{F}_z^{a\prime} \tilde{F}_z^{a})^{-1} \tilde{F}_z^{a\prime}$. We can run the above regression for each series in Y^{a_j}, and obtain the estimate of the $r \times N_{m_j}$ matrix $\hat{\Lambda}^{m_j'} = [\hat{\lambda}_1^{m_j}, \hat{\lambda}_2^{m_j}, \ldots \hat{\lambda}_{N_{m_j}}^{m_j}]$. Then one can obtain $\hat{\Lambda}^{m_j'}$ simply as

$$\begin{aligned}
\hat{\Lambda}^{m_j'} &= [\hat{\lambda}_1^{j}, \hat{\lambda}_2^{j}, \ldots, \hat{\lambda}_{N_{m_j}}^{j}] \\
&= [\tilde{K} \underline{Y}_1^{a_j}, \tilde{K} \underline{Y}_2^{a_j}, \ldots, \tilde{K} \underline{Y}_{N_{m_j}}^{a_j}] \\
&= \tilde{K} [\underline{Y}_1^{a_j}, \underline{Y}_2^{a_j}, \ldots, \underline{Y}_{N_{m_j}}^{a_j}] \\
&= \tilde{K} Y^{a_j}
\end{aligned} \tag{2.12}$$

Let $\tilde{F}_z^{m_j}$ denote the portion of $\tilde{F}_z$ between $t = 1$ and $t = T_{m_j}$. Then, using (2.12), the prediction of Y^{m_j} is obtained as follows:

$$\begin{aligned}\hat{Y}^{m_j} &= \tilde{F}_z^{m_j}\hat{\Lambda}^{m_j'} \\ &= \tilde{F}_z^{m_j}(\tilde{F}_z^{a\prime}\tilde{F}_z^{a})^{-1}\tilde{F}_z^{a\prime}Y^{a_j} \\ &= \tilde{P}_j Y^{a_j} \end{aligned} \tag{2.13}$$

where $\tilde{P}_j = \tilde{F}_z^{m_j}(\tilde{F}_z^{a\prime}\tilde{F}_z^{a})^{-1}\tilde{F}_z^{a\prime}$ is a $T_{m_j} \times T_a$ projection matrix. Finally, replacing Y^{m_j} with $\hat{Y}^{m_j}$ for $j = 1, 2, ...J$, we obtain the completed data matrix $\hat{X}$.

Let $\hat{D}^j$ be a $T \times N$ matrix of zeros everywhere except the original position of Y^{m_j} in X, which is replaced with $\hat{Y}^{m_j} - Y^{m_j}$, for $j = 1, 2, ..., J$. Using the *imputation error matrices* $\hat{D}^j$, $\hat{X}$ can be expressed as follows:

$$\hat{X} = X + \sum_{j=1}^{J} \hat{D}^j \tag{2.14}$$

Let U_c denote the cross-section samples in $(T_{m_{c+1}}, T_{m_c}]$ with $c = 1, 2, ..., J$ and $T_{m_J+1} = 0$. Then, we can express the t^{th} sample observation of $\hat{X}$

$$\hat{X}_t = \begin{cases} X_t + \sum_{j=1}^{c} \hat{D}_t^j & \text{if } t \in U_c \quad , \quad c = 1, 2, ..., J \\ X_t & \text{if } t \in A \end{cases} \tag{2.15}$$

Equations (2.14) and (2.15) connect the original and imputed data sets, and will be very useful in expressing CID partitioned factor estimator in terms of different sub-blocks of $\hat{X}$ below. Note that for $t \in U_c$ the number of cross-section missing values is S_c, and the cross-section missingness ratio (CSMR) is S_c/N. As can be seen above, as we move up the missing layers, CSMR increases. In section 2.5 we show that CSMR is one of the main determinants of the factor partition behavior in large sample.

2.3 Statistical Properties of Auxiliary Factors and Loadings in FBI

This section discusses the dynamics and statistical properties of auxiliary factors ($\tilde{F}_z$) and loadings ($\hat{\Lambda}$) in the FBI algorithm. Note that although $\tilde{F}_z$ and $\tilde{F}$ look very close to each other in terms of their derivation and underlying data, their statistical properties may be different. Basically, the complete-data factor model, its elements (i.e F, Λ and e) and its assumptions (assumptions A-H) are defined

for $X = F\Lambda^{'} + e$ in Chapter I, and the large sample results for the factor model estimators obtained in the literature are obtained for this "whole" factor model. However, Z represents the available portion of X, and the factor model it exhibits (i.e. $Z = F\Lambda^{a'} + e^z$) is "partial". In this sense, although assumptions A-H lead to a consistent and asymptotically normal factor estimate for $\tilde{F}$, they do not necessarily guarantee the same large sample properties for $\tilde{F}_z$.

In order to fully exploit the standard complete-data inferential theory in FBI, below we establish the consistency and asymptotic normality of $\tilde{F}_z$ with very weak additional assumptions. For completeness, we present the additional assumptions with the same letters including a "z" subscript to stress their correspondence with their original counterparts in Chapter I[4].

Additional Assumptions: Below M denotes the *missing* indices (i.e. $t = 1, ..., T_m$ and $i = 1, ..., N_m$), and A denotes the *available* indices (i.e. $t = T_m + 1, ..., T$ and $i = N_m + 1, ..., N$).

ASSUMPTION $\mathbf{A}_z$: F_t is covariance stationary so that $T_a^{-1} \sum_{t \in A} F_t F_{t'} \xrightarrow{p} \Sigma_F$ for any available value index A.

ASSUMPTION $\mathbf{B}_z$: $\left|\left|\Lambda^{a'}\Lambda^a/N_a - \Sigma_{\Lambda^a}\right|\right| \to 0$ for some $r \times r$ positive definite matrix Σ_{Λ^a}.

ASSUMPTION $\mathbf{C}_z$: For all N_a and T we have:

2. $E(e_s^{z'} e_t^z / N_a) = E(N_a^{-1} \sum_{i \in A} e_{is} e_{it}) = \gamma_{N_a}(s,t)$, and

$$T^{-1} \sum_{s=1}^{T} \sum_{t=1}^{T} |\gamma_{N_a}(s,t)| \leq M$$

5. For every (t,s), $E|N_a^{-1/2} \sum_{i \in A} [e_{is} e_{it} - E(e_{is} e_{it})]|^4 \leq M$

ASSUMPTION $\mathbf{E}_z$: For all T and N_a, and for every $t \leq T$ and every $i \leq N_a$:

1. $\sum_{s=1}^{T} |\gamma_{N_a}(s,t)| \leq M$

ASSUMPTION $\mathbf{F}_z$: For all N_a and T we have:

3. for each t, as $N_a \to \infty$,

$$\frac{1}{\sqrt{N_a}} \sum_{i \in A} \lambda_i e_{it} \xrightarrow{d} N(0, \Gamma_{zt})$$

where $\Gamma_{zt} = lim_{N_a \to \infty} (1/N_a) \sum_{i \in A} \sum_{j \in A} \lambda_i \lambda_j^{'} E(e_{is} e_{it})$

[4]Note that the assumptions that we do not extend are either not used or already suitable for the asymptotics of $\tilde{F}_z$

4. for each i , as $T_a \to \infty$,

$$\frac{1}{\sqrt{T_a}} \sum_{t \in A} F_t e_{it} \xrightarrow{d} N(0, \Phi_{zi})$$

where $\Phi_{zi} = lim_{T_a \to \infty}(1/T_a) \sum_{s \in A} \sum_{t \in A} E(F_t F_s^{'} e_{is} e_{it})$

ASSUMPTION G_z: The eigenvalues of the $r \times r$ matrix $(\Sigma_{\Lambda^a} \Sigma_F)$ are distinct.

Given the additional assumptions above (and assumptions A-H in Chapter I), we obtain the following results that characterize the large sample behavior of F_z:

Proposition 1. *As $N_a, T \to \infty$ we have,*

$$\delta_{N_a T}^2 \left(T^{-1} \sum_{t=1}^{T} \| \tilde{F}_{zt} - H_z^{'} F_t \|^2 \right) = O_p(1)$$

where $\delta_{N_a T} = min(\sqrt{N_a}, \sqrt{T})$, $H_z = (\Lambda^{a'} \Lambda^a / N_a)(F^{'} \tilde{F}_z / T) \tilde{V}_z^{-1}$ and $\tilde{V}_z$ is a diagonal matrix consisting of the first r eigenvalues of $(1/TN_a) ZZ^{'}$ in decreasing order.

Proposition 1 closely resembles Theorem B1 mentioned in section 1.2.2 of Chapter I, and shows that the time average of the squared deviations between the estimated factors and the true factors vanish as $N_a, T \to \infty$. The convergence rate is $min(N_a, T)$, hence it depends on the panel structure. We give a short proof of Proposition 1 in Appendix 2.2.

The following proposition shows that $\tilde{F}_z$ is $min(T, \sqrt{N_a})$-consistent and asymptotically normal.

Proposition 2. *As $N_a, T \to \infty$ we have,*

(i) *If $\sqrt{N_a}/T \to 0$, then for each t,*

$$\begin{aligned} \sqrt{N_a}(\tilde{F}_{zt} - H_z^{'} F_t) &= \tilde{V}_z^{-1} \left(\frac{\tilde{F}_z^{'} F}{T} \right) \frac{1}{\sqrt{N_a}} \sum_{i \in A} \lambda_i e_{it} + o_p(1) \\ &\xrightarrow{d} N(0, V_z^{-1} Q_z \Gamma_{zt} Q_z^{'} V_z^{-1}) \end{aligned}$$

where V_z is a diagonal matrix consisting of the first r eigenvalues of $\Sigma_{\Lambda^a}^{1/2} \Sigma_F \Sigma_{\Lambda^a}^{1/2}$ in decreasing order, $Q_z = V_z^{1/2} \Upsilon_z^{'} \Sigma_{\Lambda^a}^{-1/2}$, and Υ_z is the eigenvector matrix corresponding to V_z such that $\Upsilon_z^{'} \Upsilon_z = I$.

(ii) *If* $\sqrt{N_a}/T \geq \tau > 0$*, then*

$$T(\tilde{F}_{zt} - H_z^{'} F_t) = O_p(1)$$

A brief proof of Proposition 2 can be found in Appendix 2.3.

Note that by (2.1) the auxiliary loading estimator $\hat{\lambda}_i$ is a function of $\tilde{F}_z$. Therefore, given the asymptotics of $\tilde{F}_z$, one can derive the large sample distribution of $\hat{\lambda}_i$. The proposition below derives the asymptotic distribution of $\hat{\lambda}_i$ utilizing Propositions 1 and 2, and shows that $\hat{\lambda}_i$ is $\sqrt{T_a}$-consistent and asymptotically normal.

Proposition 3. *If* $\sqrt{T_a}/N \to 0$*, then*

$$\sqrt{T_a}(\hat{\lambda}_i - H_z^{-1}\lambda_i) \xrightarrow{d} N(0, Q_z^{-1'}\Phi_{zi}Q_z^{-1}) \quad , \quad for \;\; i \in M$$

where H_z*,* Q_z *and* Φ_{zi} *are defined in Assumption* F_z4 *and Proposition 1 above.*

Note that the limiting distribution of $\hat{\lambda}_i$ is very similar to that of $\tilde{\lambda}_i$ which is the infeasible PC estimator obtained from X and is given as $\sqrt{N}(\tilde{F}_t - H'F_t) \xrightarrow{d} N(0, (Q')^{-1}\Phi_i Q^{-1})$ for $\sqrt{T}/N \to 0$. The reason behind this similarity is as follows: Both loading estimators $\hat{\lambda}_i$ and $\tilde{\lambda}_i$ are calculated by simply regressing $\underline{X}_i$ and $\underline{Y}_i^a$ on $\tilde{F}$ and $\tilde{F}_z^a$ that are obtained as scaled eigenvector matrices of X and Z, respectively. As a result, both loading estimators have a similar expression in terms of the series and factor estimators involved which leads to similar asymptotics. Proposition 1 can be shown directly by using Theorem 1 in Bai and Ng (2006), we give a detailed proof of Proposition 3 in Appendix 2.4.

A straightforward generalization of Proposition 3 to the monotone missing case is given as follows:

Corollary 2.1 If $\sqrt{T_a}/N \to 0$, then for $i, t \in M_j$ and $j = 1, .., J$ we have

$$\sqrt{T_a}(\hat{\lambda}_i^{m_j} - H_z^{-1}\lambda_i^{m_j}) \xrightarrow{d} N(0, Q_z^{-1'}\Phi_{zi}Q_z^{-1})$$

2.4 Large Sample Distribution of Imputation Error

This section obtains the asymptotic distribution of the imputation error $\hat{d}_{it} = \tilde{F}_{zt}'\hat{\lambda}_i - F_t'\lambda_i$ introduced in (2.7). Note that we refer to $\hat{d}_{it}$, rather than $\hat{D}_{it} = \hat{Y}_{it} - Y_{it} = \hat{d}_{it} - e_{it}$, as the imputation error since it represents the part of the factor model that we can systematically explain. Once the behavior of $\hat{d}_{it}$ is known, it is elementary[5] to derive the distribution of $\hat{D}_{it}$, for which we provide the small sample approximate distribution in two corollaries below.

[5]Note that the derivation of $\hat{D}_{it}$ below assumes that X does not exhibit any factor structure shift that guarantees that $\hat{d}_{it}$ and $e_i t$ are orthogonal.

Theorem 2.1 below shows that FBI procedure consistently estimates the common component of missing values, and the imputed values and imputation error are asymptotically normal.

Theorem 2.1 For $i, t \in M$ we have

$$\left(\frac{1}{N_a}G_{zit} + \frac{1}{T_a}W_{zit}\right)^{-1/2} (\tilde{F}_{zt}^{'}\hat{\lambda}_i - F_t^{'}\lambda_i) \xrightarrow{d} N(0,1)$$

where $G_{zit} = \lambda_i^{'}\Sigma_{\Lambda^a}^{-1}\Gamma_{zt}\Sigma_{\Lambda^a}^{-1}\lambda_i$, $W_{zit} = F_t^{'}\Sigma_F^{-1}\Phi_{zi}\Sigma_F^{-1}F_t$, and Γ_{zt} and Φ_{zi} are defined above. Both Γ_{zt} and Φ_{zi} can be replaced by their consistent estimators.

Theorem 2.1 does not restrict the the relation between N_a and T_a to achieve asymptotic normality for the common component estimator. The convergence rate above is $\delta_{N_aT_a} = min(\sqrt{N_a}, \sqrt{T_a})$ which can be seen rewriting the theorem as follows:

$$\frac{\delta_{N_aT_a}(\tilde{F}_{zt}^{'}\hat{\lambda}_i - F_t^{'}\lambda_i)}{\left(\frac{\delta_{N_aT_a}^2}{N_a}G_{zit} + \frac{\delta_{N_aT_a}^2}{T_a}W_{zit}\right)^{1/2}} \xrightarrow{d} N(0,1)$$

Note that Theorem 2.1 has two special cases; we provide the small sample approximate distribution of $\hat{D}_{it}$ along with these cases below:

Corollary 2.2 If $N_a/T_a \to 0$, then

$$\sqrt{N_a}(\hat{Y}_{it} - \lambda_i^{'}F_t) \xrightarrow{d} N(0, G_{zit})$$

Additionally, if there is no factor structure shift in X, then we have

$$\hat{D}_{it} = \hat{Y}_{it} - \lambda_i^{'}F_t - e_{it} \overset{a}{\sim} N(0, G_{zit}/N_a + \sigma_{e_{it}}^2)$$

where $\sigma_{e_{it}}^2$ is the variance of e_{it}.

Corollary 2.3 If $T_a/N_a \to 0$, then

$$\sqrt{T_a}(\hat{Y}_{it} - \lambda_i^{'}F_t) \xrightarrow{d} N(0, W_{zit})$$

Additionally, if there is no factor structure shift in X, then we have

$$\hat{D}_{it} = \hat{Y}_{it} - \lambda_i^{'}F_t - e_{it} \overset{a}{\sim} N(0, W_{zit}/T_a + \sigma_{e_{it}}^2)$$

We end this section with a corollary that is a straightforward generalization of Theorem 2.1 to the case monotone missingness:

Corollary 2.4 For $i, t \in M_j$ and $j = 1, .., J$ we have

$$\left(\frac{1}{N_a}G_{zit} + \frac{1}{T_a}W_{zit}\right)^{-1/2} (\tilde{F}'_{zt}\hat{\lambda}_i^{m_j} - F'_t\lambda_i) \xrightarrow{d} N(0,1)$$

Obviously, Corollary 2.4 can be extended along the lines of Corollary 2.2 and 2.3 to account for the idiosyncratic variation in the missing value imputation error. In order to preserve space we do not present neither these corollaries nor the proofs for the corollaries above. The intuition behind the proof of the corollaries 2.1-2.4 is as follows: The initial factor estimator $\tilde{F}_z$ is identical in block and monotone missingness cases. Additionally, each layer j can be considered as a missing block in itself. Bringing these observations together, we can simply extend the results we found for the loading and common component estimators to each layer of the monotone missing data panel.

2.5 Analytics of Partitioned CID Factor Estimators

In this section we express partitioned CID factor estimators in terms of their underlying components. This decomposition

i. allows us to understand the internal structure of CID estimators

ii. shows the relation between CID partitions and the sub-blocks of $\hat{X}$

iii. implies that CID factor estimators are a proper generalization of the standard complete data factor estimators

iv. establishes a starting point for future research that would focus on deriving the large sample distribution of CID estimators

Let $\hat{V}$ denote a diagonal matrix whose diagonal elements are the first r eigenvectors of the $T \times T$ matrix $\hat{X}\hat{X}'/NT$ in decreasing order. Using the definition of $\tilde{\tilde{F}}$ above, we can write

$$(NT)^{-1}\hat{X}\hat{X}'\frac{\tilde{\tilde{F}}}{\sqrt{T}} = \frac{\tilde{\tilde{F}}}{\sqrt{T}}\hat{V}$$

Rearranging the above equation we find

$$\tilde{\tilde{F}} = (NT)^{-1}\hat{X}\hat{X}'\tilde{\tilde{F}}\hat{V}^{-1}$$

Taking the t^{th} row from the above equation and taking its transpose yields

$$\tilde{\tilde{F}}_t = (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'\hat{X}\hat{X}_t \tag{2.16}$$

Below, we first consider the case of block missingness, and dissect CID factors into their various components and then extend our findings to monotone missingness.

2.5.1 Block Missingness

Substituting (2.5) and (2.6) in (2.16), we obtain

$$\begin{aligned}\tilde{\tilde{F}}_t &= (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'(X+\hat{D})(X_t+\hat{D}_t) \\ &= \begin{cases}(NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'(\ XX_t + \hat{D}X_t + X\hat{D}_t + \hat{D}\hat{D}_t\ ) & \text{if } t \in M \\ (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'(\ XX_t + \hat{D}X_t\ ) & \text{if } t \in A\end{cases}\end{aligned} \tag{2.17}$$

Comparing (2.17) with the corresponding complete-data case in Bai (2003), we see that the imputation error $\hat{D}$ leads to some extra terms. As can be seen $\hat{D}$ affects both $\tilde{\tilde{F}}^m$ and $\tilde{\tilde{F}}^a$, although the missingness is confined to X^m. Note that $\tilde{\tilde{F}}^m$ has two more extra terms compared to $\tilde{\tilde{F}}^a$.

Substituting (1.2) and (1.3) in the first term in parenthesis in (2.17) we obtain

$$\begin{aligned}(NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'XX_t &= (NT)^{-1}\hat{V}^{-1}\left(\tilde{\tilde{F}}'(\ F\Lambda' + e\ )(\ \Lambda F_t + e_t\ )\right) \\ &= (NT)^{-1}\hat{V}^{-1}\left(\tilde{\tilde{F}}'F\Lambda'\Lambda F_t + \tilde{\tilde{F}}'F\Lambda' e_t + \tilde{\tilde{F}}'e\Lambda F_t + \tilde{\tilde{F}}'ee_t\right)\end{aligned}$$

Letting $\tilde{\tilde{H}} = (NT)^{-1}(\Lambda'\Lambda)(\mathrm{F}'\tilde{\tilde{F}})\hat{V}^{-1}$, then we can write

$$(NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'XX_t - \tilde{\tilde{H}}'F_t = (NT)^{-1}\hat{V}^{-1}\left(\tilde{\tilde{F}}'F\Lambda' e_t + \tilde{\tilde{F}}'e\Lambda F_t + \tilde{\tilde{F}}'ee_t\right) \tag{2.18}$$

Then substituting (2.18) in (2.17) we obtain

$$\tilde{\tilde{F}}_t - \tilde{\tilde{H}}'F_t = \begin{cases}(NT)^{-1}\hat{V}^{-1}\left((\tilde{\tilde{F}}'F\Lambda' e_t + \tilde{\tilde{F}}'e\Lambda F_t + \tilde{\tilde{F}}'ee_t) + \tilde{\tilde{F}}'DX_t + \tilde{\tilde{F}}'XD_t + \tilde{\tilde{F}}'DD_t\right) & \text{if } t \in M \\ (NT)^{-1}\hat{V}^{-1}\left((\tilde{\tilde{F}}'F\Lambda' e_t + \tilde{\tilde{F}}'e\Lambda F_t + \tilde{\tilde{F}}'ee_t) + \tilde{\tilde{F}}'DX_t\right) & \text{if } t \in A\end{cases}$$

Letting

$$I = (NT)^{-1}\hat{V}^{-1}(\tilde{\tilde{F}}'F\Lambda'e_t + \tilde{\tilde{F}}'e\Lambda F_t + \tilde{\tilde{F}}'ee_t) \tag{2.19a}$$

$$II = (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'DX_t \tag{2.19b}$$

$$III = (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'XD_t \tag{2.19c}$$

$$IV = (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'DD_t \tag{2.19d}$$

we can rewrite the above identity as

$$\tilde{\tilde{F}}_t - \tilde{\tilde{H}}'F_t = \begin{cases} I + II + III + IV & \text{if } t \in M \\ I + II & \text{if } t \in A \end{cases} \tag{2.20}$$

(2.20) is the main identity we will consider to study the consistency and asymptotic normality properties of the estimated CID factors under block missingness.

Note that if there is no missingness in the system, then D and D_t become both 0, and $\tilde{\tilde{F}}_t$ and become equal to $\tilde{F}_t$ and $\tilde{H}$, respectively. As a result the remaining equation would be

$$\begin{aligned} \tilde{F}_t - H'F_t &= (NT)^{-1}XX'\tilde{F}'\tilde{V}^{-1} - H'F_t \\ &= (NT)^{-1}\tilde{V}^{-1}\left(\tilde{F}'F\Lambda'e_t + \tilde{F}'e\Lambda F_t + \tilde{F}'ee_t\right) \end{aligned}$$

which is where Bai (2003) starts with in deriving the asymptotic distribution of estimated factors under standard large dimensional factor model setting[6].

[6]We discuss the derivation of the asymptotics of $\tilde{F}_t$ using the above setting in Appendix 2.1.

2.5.2 Monotone Missingness

Substituting (2.14) and (2.15) in (2.16), for $c = 1, 2, ..., J$ we obtain

$$\begin{aligned}\tilde{\tilde{F}}_t &= (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'\left(X + \sum_{j=1}^{J}\hat{D}^j\right)\left(X_t + \sum_{j=1}^{c}\hat{D}_t^j\right) \\ &= \begin{cases} (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'(\ XX_t + X\sum_{j=1}^{c}\hat{D}_t^j + \sum_{j=1}^{J}\hat{D}^j X_t + \sum_{j=1}^{J}\hat{D}^j\sum_{j=1}^{c}\hat{D}_t^j & \text{if } t \in M_c \\ (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'(\ XX_t + \sum_{j=1}^{J}\hat{D}^j X_t\ ) & \text{if } t \in A \end{cases}\end{aligned} \tag{2.21}$$

As in the case of block missingness above, in (2.21) the imputation error $\sum_{j=1}^{J}\hat{D}^j$ leads to some extra terms compared to the standard case considered in Bai (2003). As can be seen $\sum_{j=1}^{J}\hat{D}^j$ affects both missing and available partitions of F although cross-section samples corresponding to $\tilde{\tilde{F}}^a$ are completely observed. A similar *layer spillover* across partitions is also observed among the missing partitions of $\tilde{\tilde{F}}$ thorough the term $\sum_{j=1}^{J}\hat{D}^j$. This term contains imputation errors from *all* of the missing layers. Therefore although for $t \in U_c$ missingness is confined to the first S_c elements of $\hat{X}_t$, missingness from the rest of the data set also affects the asymptotic distribution of $\tilde{\tilde{F}}_t$. In addition as we move up in X (i.e. as c increases) CSMR increases, and it follows from (2.21) that the number of terms in $\tilde{\tilde{F}}$ increases. Therefore, higher CSMR in X_t results in a more complicated expression for $\tilde{\tilde{F}}$, and (intuitively) leads to a higher asymptotic variance.

Using (2.18) and letting

$$I = (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'\left(F\Lambda' e_t + \tilde{\tilde{F}}' e\Lambda F_t + \tilde{\tilde{F}}' e e_t\right) \tag{2.22a}$$

$$II = (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}' X\sum_{j=1}^{c}\hat{D}_t^j \tag{2.22b}$$

$$III = (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'\sum_{j=1}^{J}\hat{D}^j X_t \tag{2.22c}$$

$$IV = (NT)^{-1}\hat{V}^{-1}\tilde{\tilde{F}}'\sum_{j=1}^{J}\hat{D}^j\sum_{j=1}^{c}\hat{D}_t^j \tag{2.22d}$$

We can rewrite the identity (2.21) as

$$\tilde{\hat{F}}_t - \tilde{\hat{H}}' F_t = \begin{cases} I + II + III + IV & \text{if } t \in M_c \ , \ \ c = 1, 2, ..., J \\ I + II & \text{if } t \in A \end{cases} \tag{2.23}$$

where we utilized the expressions $X = F\Lambda' + e$ and $X_t = \Lambda F_t + e_t$.

Note that when missingness is removed we obtain $\hat{D}^j = 0$ (thus $\hat{D}_t^j = 0$) which leads to $\hat{X} = X$, $\tilde{\hat{F}} = \tilde{F}$ and $\tilde{\hat{H}} = H = (\frac{\Lambda'\Lambda}{N})(\frac{F'\tilde{F}}{T})\tilde{V}^{-1}$. In addition, the observed samples index set A becomes the set of all the observations since there is no missing value. As a result, equation (2.23) becomes

$$\tilde{F}_t - H'F_t = (NT)^{-1}\tilde{V}^{-1}\left(\tilde{F}'F\Lambda' e_t + \tilde{F}' e \Lambda F_t + \tilde{F}' e e_t\right)$$

which is the same form that we obtain in the previous section. The above results shows that the setting we develop above is a proper generalization of the standard asymptotic theory to the case in which X has monotone missingness.

2.6 Limitations and Further Research

The study of the factor models under missingness can be extended in several dimensions: First, our FBI algorithm is a single imputation procedure; as an alternative, one can extended it with multiple imputation. Although this would create more complex (possibly intractable) analytics, it has the potential to yield more efficient estimators .

Second, our FBI procedure yields consistent and asymptotically normal CID estimators, but does not study the efficiency properties of this particular method. Further research can consider finding the most efficient imputation method from among the class of consistent imputation methods.

Finally, in factor estimation process we treat the number of factors to be estimated as a known constant. In practice the number of factors is not known and needs to be estimated from the data. Estimation error incurred in the estimation of number of factors would inflate the variance of factor model estimators further. To this end one can estimate r with the usual AIC or BIC in equation (1.7), or the alternative criterion developed by Bai and Ng (2002) that consistently estimates the number of factors, and consider the effect of this additional estimation error on the limiting distribution of factor model estimators.

2.7 Appendix 2.1: Limiting Distribution of Estimated Factors under Complete Data

This section presents a sketch of the results for limiting distribution of estimated factors considered in Bai (2003). Theorem 1 in Bai (2003) is as follows:

Theorem [7]**:** Under the assumptions in Bai (2003) if $\sqrt{N}/T \to 0$, then for each t

$$\sqrt{N}(\tilde{F}_t - H'F_t) \xrightarrow{d} N(0, V^{-1}Q\Gamma_t Q'V^{-1})$$

where $V = diag(\nu_1, \nu_2, ...\nu_r)$, $\nu_1 > \nu_2 > ...\nu_r$ are the eigenvalues of $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}$ (Σ_Λ and Σ_F are the covariance matrices of Λ and F, respectively.), Γ_t is the asymptotic covariance matrix of sequence $\{\lambda_i e_{it}\}$, and Q is equal to $plim_{T,N\to\infty}\frac{\tilde{F}'F}{T}$.

In order to prove the theorem above, Bai (2003) utilizes the following identity:

$$\tilde{F}_t - H'F_t = \tilde{V}^{-1}\left(\frac{1}{T}\sum_{s=1}^{T}\tilde{F}_s\gamma_N(s,t) + \frac{1}{T}\sum_{s=1}^{T}\tilde{F}_s\zeta_{st} + \frac{1}{T}\sum_{s=1}^{T}\tilde{F}_s\eta_{st} + \frac{1}{T}\sum_{s=1}^{T}\tilde{F}_s\xi_{st}\right) \tag{2.24}$$

where

$$\gamma_N(s,t) = E(e_s'e_t/N) \tag{2.25a}$$

$$\zeta_{st} = e_s'e_t/N - \gamma_N(s,t) \tag{2.25b}$$

$$\eta_{st} = F_s'\Lambda'e_t/N \tag{2.25c}$$

$$\xi_{st} = F_t'\Lambda'e_s/N \tag{2.25d}$$

Below we will first show how the identity (2.24) is obtained. Then we will give a sketch of how the above theorem is proven. Let $\tilde{V}$ be the $k \times k$ diagonal matrix consisting of the first k eigenvalues of XX'/TN, arranged in a decreasing order. When $T < N$, the factor estimates $\tilde{F}$ is $\sqrt{T}$ times the eigenvectors corresponding to the k largest eigenvalues of $T \times T$ matrix XX'. Using this we can write

$$(NT)^{-1}XX'\frac{\tilde{F}}{\sqrt{T}} = \frac{\tilde{F}}{\sqrt{T}}\tilde{V}$$

Rearranging the above equation we obtain

$$\tilde{F} = (NT)^{-1}XX'\tilde{F}'\tilde{V}^{-1}$$

Taking the t^{th} row from the above equation and taking its transpose we obtain

$$\tilde{F}_t = (NT)^{-1}\tilde{V}^{-1}\tilde{F}'XX_t \tag{2.26}$$

[7] We refer to this theorem as Theorem B1 in Chapter I section 1.2.2

Substituting $X = F\Lambda' + e$ and $X_t = \Lambda F_t + e_t$ in (2.26) we obtain

$$\tilde{F}_t = (NT)^{-1}\tilde{V}^{-1}\left(\tilde{F}'F\Lambda'\Lambda F_t + \tilde{F}'F\Lambda' e_t + \tilde{F}'e\Lambda F_t + \tilde{F}'ee_t\right) \tag{2.27}$$

Now let $H = (\frac{\Lambda'\Lambda}{N})(\frac{F'\tilde{F}}{T})\tilde{V}^{-1}$, then we have

$$\begin{aligned}
\tilde{F}_t - H'F_t &= (NT)^{-1}\tilde{V}^{-1}\left(\tilde{F}'F\Lambda' e_t + \tilde{F}'e\Lambda F_t + \tilde{F}'ee_t\right) \\
&= \tilde{V}^{-1}\left(\frac{1}{T}\sum_{s=1}^{T}\tilde{F}_s\frac{F_s'\Lambda' e_t}{N} + \frac{1}{T}\sum_{s=1}^{T}\tilde{F}_s\frac{e_s'\Lambda F_t}{N} + \frac{1}{T}\sum_{s=1}^{T}\tilde{F}_s\frac{e_s' e_t}{N}\right)
\end{aligned}$$

Rearranging the expression above using (2.25a) through (2.25d), we obtain (2.24). Bai (2003) shows the consistency of $\tilde{F}_t$ for the factor space spanned by F_t by showing that the right hand side of the identity (2.24) converges in probability to zero. Note that Bai (2003) does not estimate the true factors; it estimates the factor space spanned by the true factors. This point is captured by introducing the scaling matrix H in (2.24).

Proof of the Theorem:

After establishing the identity (2.24), Bai (2003) shows the following results that are used in the proof of Theorem 1 (below $\delta_{NT}^2 = min(N,T)$):

Lemma 2.1.1:

$$\delta_{NT}^2\left(\frac{1}{T}\sum_{t=1}^{T}\|\tilde{F}_t - H'F_t\|^2\right) = O_p(1)$$

Lemma 2.1.2:

(a) $\frac{1}{T}\sum_{s=1}^{T}\tilde{F}_s\gamma_N(s,t) = O_p(\frac{1}{\sqrt{T}\delta_{NT}})$

(b) $\frac{1}{T}\sum_{s=1}^{T}\tilde{F}_s\zeta_{st} = O_p(\frac{1}{\sqrt{N}\delta_{NT}})$

(c) $\frac{1}{T}\sum_{s=1}^{T}\tilde{F}_s\eta_{st} = O_p(\frac{1}{\sqrt{N}})$

(d) $\frac{1}{T}\sum_{s=1}^{T}\tilde{F}_s\xi_{st} = O_p(\frac{1}{\sqrt{N}\delta_{NT}})$

Lemma 2.1.3:

$$(a)\ T^{-1}\tilde{F}'\left(\frac{1}{TN}XX'\right)\tilde{F} = \tilde{V} \xrightarrow{p} V$$

$$(b)\ \frac{\tilde{F}'F}{T}\left(\frac{\Lambda'\Lambda}{N}\right)\frac{F'\tilde{F}}{T} \xrightarrow{p} V$$

Proposition 2.1.1:

$$plim_{T,N\to\infty}\frac{\tilde{F}'F}{T} = Q$$

The matrix Q is given by $Q = V^{1/2}\Upsilon'\Sigma_{\Lambda}^{-1/2}$, V, Σ_{Λ} and Σ_F are as given in Theorem 1, and Υ is the eigenvector matrix of $\Sigma_{\Lambda}^{1/2}\Sigma_F\Sigma_{\Lambda}^{1/2}$ (Σ_{Λ} such that $\Upsilon'\Upsilon = I$).

Now, given Lemma 2.1.1 and 2.1.2, we have

$$\tilde{F}_t - H'F_t = O_p(\frac{1}{\sqrt{T}\delta_{NT}}) + O_p(\frac{1}{\sqrt{N}\delta_{NT}}) + O_p(\frac{1}{\sqrt{N}}) + O_p(\frac{1}{\sqrt{N}\delta_{NT}}) \tag{2.28}$$

The limiting distribution is determined by the third term on the right hand side of (2.28) since it is the dominant term. Using the definition of η_{st}

$$\sqrt{N}(\tilde{F}_t - H'F_t) = \tilde{V}^{-1}\frac{1}{T}\sum_{s=1}^{T}(\tilde{F}_s F_s')\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i e_{it} + o_p(1) \tag{2.29}$$

By initial assumptions of Bai (2003), we have $(1/\sqrt{N})\sum_{i=1}^{N}\lambda_i e_{it} \xrightarrow{d} N(0,\Gamma_t)$. Then, applying Lemma 2.1.3 and the Proposition to (2.29), we obtain $\sqrt{N}(\tilde{F}_t - H'F_t) \xrightarrow{d} N(0, V^{-1}Q\Gamma_t Q'V^{-1})$ as stated.

2.8 Appendix 2.2: Proof of Proposition 2.1

Lemma 2.2.1 As $N_a, T \to \infty$, we have

(i) $T^{-1}\tilde{F}_z'\left(\frac{1}{N_a T}ZZ'\right)\tilde{F}_z = \tilde{V}_z \xrightarrow{p} V_z$

(ii) $(\tilde{F}_z'F/T)(\Lambda^{a\prime}\Lambda^a/N_a)(F'\tilde{F}_z/T) \xrightarrow{p} V_z$

where V_z is the diagonal matrix consisting of the eigenvalues of $\Sigma_{\Lambda^a}\Sigma_F$ in decreasing order.

Proof: Consider the identity

$$(1/N_aT)ZZ'\tilde{F}_z = \tilde{F}_z\tilde{V}_z \tag{2.30}$$

Multiplying both sides of (2.30) with $\tilde{F}_z'$ and using $\tilde{F}_z'\tilde{F}_z/T = I$ we obtain

$$T^{-1}\tilde{F}_z'\left(\frac{1}{N_aT}ZZ'\right)\tilde{F}_z = \tilde{V}_z$$

The rest is implicitly proved in Stock and Watson (1999).

Lemma 2.2.2 For all N_a and T, we have some $M \leq \infty$ such that,

(i) $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}\gamma_{N_a}(s,t)^2 \leq M$

(ii) $E\left(T^{-1}\sum_{t=1}^{T}\|N_a^{-1/2}\sum_{i\in A}e_{it}\lambda_i\|^2\right) \leq M$

(iii) $E\left(T^{-2}\sum_{t=1}^{T}\sum_{s=1}^{T}(N_a^{-1}\sum_{i\in A}Z_{it}Z_{is})^2\right) \leq M$

(iv) $E\left\|(N_aT)^{-1/2}\sum_{i\in A}\sum_{s=1}^{T}e_{it}\lambda_i\right\| \leq M$

where $\gamma_{N_a}(s,t) = \sum_{i\in A}e_{is}e_{is}$.

Proof: The results above are shown in Bai and Ng (2002) pp.212-213 for N, e and Λ. Their extension to N_a, e^z and Λ^a is easily achieved under our additional assumptions, and omitted here to preserve space.

Proof of Proposition 2.2: Following the proof of Theorem 1 of Bai (2003) given in Appendix 1 above, we can obtain the following identity for the factor estimate $\tilde{F}_z$ extracted from portion Z of X.

$$\tilde{F_{zt}} - H_z' F_t = \tilde{V}_z^{-1}\left(\frac{1}{T}\sum_{s=1}^{T}\tilde{F}_{zs}\gamma_{N_a}(s,t) + \frac{1}{T}\sum_{s=1}^{T}\tilde{F}_{zs}\zeta_{st}^z + \frac{1}{T}\sum_{s=1}^{T}\tilde{F}_{zs}\eta_{st}^z + \frac{1}{T}\sum_{s=1}^{T}\tilde{F}_{zs}\xi_{st}^z\right) \tag{2.31}$$

where

$$\gamma_{N_a}(s,t) = E\left(N_a^{-1}\sum_{i\in A} e_{is}e_{it}\right) \tag{2.32a}$$

$$\zeta_{st}^a = N_a^{-1}\sum_{i\in A} e_{is}e_{it} - \gamma_{N_a}(s,t) \tag{2.32b}$$

$$\eta_{st}^a = F_s'\Lambda^{a\prime}e_t/N_a \tag{2.32c}$$

$$\xi_{st}^a = F_t'\Lambda^{a\prime}e_s/N_a \tag{2.32d}$$

$$\tag{2.32e}$$

Above $\tilde{V}_z$ denotes a diagonal matrix consisting of the first r eigenvalues of $(1/N_aT)ZZ^{'}$ and H_z is given by $H_z = (\Lambda^{a'}\Lambda^{a'}/N_a)(F^{'}\tilde{F}_z/T)\tilde{V}_z^{-1}$. By Lemma A.1 $\tilde{V}_z$ converges to V_z which is a positive definite matrix; therefore $\tilde{V}_z = O_p(1)$. Note that $H_z = O_p(1)$ is since $\|H_z\| \leq \|\Lambda^{a'}\Lambda^{a'}/N_a\|\|F^{'}F/T\|\|\tilde{F}_z'\tilde{F}_z/T\|\|\tilde{V}_z^{-1}\|$ and each of the matrix norms on the right hand side is bounded by assumptions A, B_z and Lemma 2.2.1. Now using the inequality $(x+y+z+w)^2 \leq 4(x^2+y^2+z^2+w^2)$ we can write $\|\tilde{F}_{zt} - H_z'F_t\|^2 \leq 4(a_t+b_t+c_t+d_t)$, where $a_t = T^{-2}\left\|\sum_{s=1}^{T}\tilde{F}_{zs}\gamma_{N_a}(s,t)\right\|^2$, $b_t = T^{-2}\left\|\sum_{s=1}^{T}\tilde{F}_{zs}\zeta_{st}^z\right\|^2$, $c_t = T^{-2}\left\|\sum_{s=1}^{T}\tilde{F}_{zs}\eta_{st}^z\right\|^2$, and $d_t = T^{-2}\left\|\sum_{s=1}^{T}\tilde{F}_{zs}\xi_{st}^z\right\|^2$. Therefore we can write

$$T^{-1}\sum_{t=1}^{T}\|\tilde{F}_{zt} - H_z'F_t\|^2 \leq T^{-1}\sum_{t=1}^{T}(a_t+b_t+c_t+d_t) \tag{2.33}$$

Consider a_t above: Given $\|\sum_{s=1}^{T}\tilde{F}_{zs}\gamma_{N_a}(s,t)\|^2 \leq (\sum_{s=1}^{T}\|\tilde{F}_{zs}\|^2)(\sum_{s=1}^{T}\gamma_{N_a}(s,t)^2)$,

$$\begin{aligned} T^{-1}\sum_{t=1}^{T}a_t &\leq T^{-1}\left(T^{-1}(\sum_{s=1}^{T}\|\tilde{F}_{zs}\|^2\right)T^{-1}\left(\sum_{t=1}^{T}\sum_{s=1}^{T}\gamma_{N_a}(s,t)^2\right) \\ &= T^{-1}O_p(1)O_p(1) \\ &= O_p(T^{-1}) \end{aligned}$$

by Lemma 2.2.2 (ii). Similarly, following the procedure in Bai and Ng (2002) we can show that $b_t = O_p(T/N_a)$, $c_t = O_p(N_a^{-1})$ and $d_t = O_p(N_a^{-1})$. Substituting these results in (2.33) we obtain

$$T^{-1}\sum_{t=1}^{T}\|F_{zt} - H_z'F_t\|^2 \leq T^{-1}\sum_{t=1}^{T}(a_t + b_t + c_t + d_t) = O_p(T^{-1}) + O_p(N_a^{-1})$$

This completes the proof.

2.9 Appendix 2.3: Proof of Proposition 2.2

To prove the Proposition 2.2, we need the following lemmas[8]

Lemma 2.3.1 As $N_a, T \to \infty$, we have

(i) $T^{-1}\sum_{s=1}^{T}\tilde{F}_{zs}\gamma_{N_a}(s,t) = O_p\left(\frac{1}{\sqrt{T}\delta_{N_aT}}\right)$

(ii) $T^{-1}\sum_{s=1}^{T}\tilde{F}_{zs}\zeta_{st}^{z} = O_p\left(\frac{1}{\sqrt{N_a}\delta_{N_aT}}\right)$

(iii) $T^{-1}\sum_{s=1}^{T}\tilde{F}_{zs}\eta_{st}^{z} = O_p\left(\frac{1}{\sqrt{N_a}}\right)$

(iv) $T^{-1}\sum_{s=1}^{T}\tilde{F}_{zs}\xi_{st}^{z} = O_p\left(\frac{1}{\sqrt{N_a}\delta_{N_aT}}\right)$

Proof: Consider part (i)[9]. By adding and subtracting $H_z^{'}F_s$ we obtain

$$\begin{aligned} T^{-1}\sum_{s=1}^{T}\tilde{F}_{zs}\gamma_{N_a}(s,t) &= T^{-1}\sum_{s=1}^{T}(\tilde{F}_{zs} - H_z^{'}F_s + H_z^{'}F_s)\gamma_{N_a}(s,t) \\ &= T^{-1}\sum_{s=1}^{T}(\tilde{F}_{zs} - H_z^{'}F_s)\gamma_{N_a}(s,t) + H_z^{'}T^{-1}\sum_{s=1}^{T}F_s\gamma_{N_a}(s,t) \end{aligned} \tag{2.34}$$

Since $E\left|\sum_{s=1}^{T}F_s\gamma_{N_a}(s,t)\right| \leq (max_s E\|F_s\|)\sum_{s=1}^{T}|\gamma_{N_a}(s,t)| \leq M^{1+1/4}$ by assumption A and E_z1, we have $(1/T)\sum_{s=1}^{T}\tilde{F}_{zs}\gamma_{N_a}(s,t) = O_p(1/T)$. Now consider the first term in (2.34):

$$\begin{aligned} \left|T^{-1}\sum_{s=1}^{T}(\tilde{F}_{zs} - H_z^{'}F_s)\gamma_{N_a}(s,t)\right| &\leq \left(T^{-1}\sum_{s=1}^{T}\|\tilde{F}_{zs} - H_z^{'}F_s\|^2\right)^{1/2}\frac{1}{\sqrt{T}}\left(\sum_{s=1}^{T}|\gamma_{N_a}(s,t)|^2\right)^{1/2} \\ &= O_p\left(\frac{1}{\delta_{N_aT}}\right)\frac{1}{\sqrt{T}}O_p(1) \\ &= O_p\left(\frac{1}{\sqrt{T}\delta_{N_aT}}\right) \end{aligned} \tag{2.35}$$

[8]The proof of Proposition 2 follows the proof of Theorem 1 in Bai (2003) very closely.

[9]The proof of Lemma 2.3.1 closely follows that of Lemma 2.2.2 in Bai (2003)

by Proposition 2.1 and assumption E_z1. Parts (ii)-(iv) can be proven similarly referring to the proof of Theorem 1 in Bai (2003) and using our additional assumptions.

Lemma 2.3.2 As $N_a, T \to \infty$, we have

$$\frac{\tilde{F}_z' F}{T} \xrightarrow{p} Q_z$$

where $Q_z = \mathrm{V}_z^{1/2} \Upsilon_z' \Sigma_{\Lambda^a}^{-1/2}$, and Υ_z is the eigenvector matrix corresponding to V_z such that $\Upsilon_z' \Upsilon_z = I$.

Proof: Multiplying both sides of (2.30) with $T^{-1}(\Lambda^{a'}\Lambda^a/N_a)^{1/2}F'$, we obtain[10]

$$\left(\frac{\Lambda^{a'}\Lambda^a}{N_a}\right)^{1/2} T^{-1} F' \frac{ZZ'}{TN_a} \tilde{F}_z = \left(\frac{\Lambda^{a'}\Lambda^a}{N_a}\right)^{1/2} \frac{F'\tilde{F}_z}{T} \tilde{V}_z$$

Substituting ZZ' with $Z = F\Lambda^a + e^z$, we obtain

$$\left(\frac{\Lambda^{a'}\Lambda^a}{N_a}\right)^{1/2} \left(\frac{F'F}{T}\right) \left(\frac{\Lambda^{a'}\Lambda^a}{N_a}\right) \left(\frac{F'\tilde{F}_z}{T}\right) + \tilde{d}_z = \left(\frac{\Lambda^{a'}\Lambda^a}{N_a}\right)^{1/2} \frac{F'\tilde{F}_z}{T} \tilde{V}_z \tag{2.36}$$

where

$$\tilde{d} = \left(\frac{\Lambda^{a'}\Lambda^a}{N_a}\right)^{1/2} \left[\left(\frac{F'F}{T}\right) \Lambda^{a'} e^{z'} \tilde{F}_z/(TN_a) + \frac{1}{TN_a} F' e \Lambda^a F' \tilde{F}_z/T + \frac{1}{TN_a} F' e^z e^{z'} \tilde{F}_z/T \right] = o_p(1)$$

by Lemma 2.3.1. Let

$$B_z = \left(\frac{\Lambda^{a'}\Lambda^a}{N_a}\right)^{1/2} \left(\frac{F'F}{T}\right) \left(\frac{\Lambda^{a'}\Lambda^a}{N_a}\right)^{1/2}$$

and

$$R_z = \left(\frac{\Lambda^{a'}\Lambda^a}{N_a}\right)^{1/2} \left(\frac{F'\tilde{F}_z}{T}\right)$$

Then we can write (2.36) as

$$(B_z + \tilde{d}_z R_z^{-1}) R_z = R_z \tilde{V}_z$$

Letting $\tilde{\Upsilon}_z = R_z \tilde{V}_z^{\star -1/2}$, where $\tilde{V}_z^{\star}$ is the diagonal elements of $R_z' R_z$, we obtain

$$(B_z + \tilde{d}_z R_z^{-1}) \tilde{\Upsilon}_z = \tilde{\Upsilon}_z \tilde{V}_z \tag{2.37}$$

Equation (2.37) shows that $\tilde{\Upsilon}_z$ and $\tilde{V}_z$ the the eigenvector and the diagonal eigenvalue matrices of $B_z + \tilde{d}_z R_z^{-1}$. Since $\tilde{d}_z = o_p(1)$, $B_z + \tilde{d}_z R_z^{-1} \xrightarrow{p} \Sigma_{\Lambda^a}^{1/2} \Sigma_F \Sigma_{\Lambda^a}^{1/2}$ by assumptions A and B_z. Using

[10]The proof of Lemma B.1 closely follows the proof of Proposition 1 in Bai (2003).

our assumption that eigenvalues of $\Sigma_{\Lambda^a}^{1/2}\Sigma_F\Sigma_{\Lambda^a}^{1/2}$, and the eigenvector perturbation theory $\tilde{\Upsilon}_z \xrightarrow{p} \Upsilon_z$ where Υ_z is the eigenvector matrix of $\Sigma_{\Lambda^a}^{1/2}\Sigma_F\Sigma_{\Lambda^a}^{1/2}$. Similarly, $\tilde{V}_z^\star \xrightarrow{p} V_z$. Then using the arguments in Bai (2003) p.162, we can show that

$$\frac{F'\tilde{F}_z}{T} = \left(\frac{\Lambda^{a\prime}\Lambda^a}{N_a}\right)^{-1/2} \tilde{\Upsilon}_z(\tilde{V}_z^\star)^{1/2} \xrightarrow{p} \Sigma_{\Lambda^a}^{-1/2}\Upsilon_z V_z^{1/2}$$

Proof of Proposition 2.2:

Case 1: $\sqrt{N_a}/T \to 0$. By (2.31) and Lemma 2.3.1, we can write

$$\tilde{F}_{zt} - H_z'F_t = O_p\left(\frac{1}{\sqrt{T}\delta_{N_aT}}\right) + O_p\left(\frac{1}{\sqrt{N_a}\delta_{N_aT}}\right) + O_p\left(\frac{1}{\sqrt{N_a}}\right) + O_p\left(\frac{1}{\sqrt{N_a}\delta_{N_aT}}\right)$$

The limiting distribution is determined by the third term above. Therefore we can, express the above equation as

$$\begin{aligned}\sqrt{N_a}(\tilde{F}_{zt} - H_z'F_t) &= \tilde{V}_z^{-1}\left(\frac{\tilde{F}_z'F}{T}\right)\frac{1}{\sqrt{N_a}}\sum_{i\in A}\lambda_i e_{it} + o_p(1) \\ &\xrightarrow{d} N(0, V_z^{-1}Q_z\Gamma_{zt}Q_z'V_z^{-1}) \end{aligned} \tag{2.38}$$

where we use assumption F_z3, Lemma 2.2.1 and Lemma 2.3.2.

Case 2: $\sqrt{N_a}/T \geq \tau > 0$. In this case first and third terms are dominant, and we obtain

$$T(\tilde{F}_t - H'F_t) = O_p(1) + O_p(T/\sqrt{N_a}) = O_p(1)$$

via limsup$(T/\sqrt{N_a}) \leq 1/\tau < \infty$.

This completes the proof.

2.10 Appendix 2.4: Proof of Proposition 2.3

In order to prove the proposition we need the following lemmas.

Lemma 2.4.1 As $N_a, T \to \infty$, we have

$$H_z \xrightarrow{p} Q_z^{-1} = \left(V_z^{1/2} \tilde{\Upsilon}'_z \Sigma_{\Lambda^a}^{-1/2} \right)^{-1}$$

Proof: By Lemma 2.2.1 and 2.3.2 and assumption B_z, we have

$$\begin{aligned} H_z &= (\Lambda^{a\prime}\Lambda^a/N_a)(\mathrm{F}'\tilde{F}_z/T)\tilde{V}_z^{-1} \\ &\xrightarrow{p} \Sigma_{\Lambda^a} Q_z' V_z^{-1} \\ &= \Sigma_{\Lambda^a}(V_z^{1/2}\tilde{\Upsilon}'_z\Sigma_{\Lambda^a}^{-1/2})' V_z^{-1} \\ &= \Sigma_{\Lambda^a}\Sigma_{\Lambda^a}^{-1/2}\tilde{\Upsilon}_z V_z^{1/2} V_z^{-1} \\ &= \Sigma_{\Lambda^a}^{1/2}\Upsilon_z V_z^{-1/2} \\ &= \left(V_z^{1/2}\tilde{\Upsilon}'_z\Sigma_{\Lambda^a}^{-1/2} \right)^{-1} \\ &= Q_z^{-1} \end{aligned} \tag{2.39}$$

Lemma 2.4.2 Under our assumptions, for $i \in A$ we have
(i) $T_a^{-1}(\tilde{F}_z^a - F^a H_z)'\underline{e}_i = O_p(\delta_{N_a T}^{-2})$
(ii) $T_a^{-1}\tilde{F}_z^{a\prime}(F^a H_z - \tilde{F}_z^a) = O_p(\delta_{N_a T}^{-2})$

Proof: Proof of (i) and (ii) above follow from Bai(2003) Lemma 2.3.1 and 2.3.2, our Proposition 2.1 and 2.2, and assumption A_z. In order to preserve space, the details are omitted.

Proof of Proposition 2.3 Given $\underline{Y}_i = \tilde{F}_z^a\lambda_i + \nu_i$ for $i \in M$, $t \in A$ the OLS estimate of λ_i is calculated as

$$\hat{\lambda}_i = (\tilde{F}_z^{a\prime}\tilde{F}_z^a)^{-1}\tilde{F}_z^{a\prime}\underline{Y}_i \tag{2.40}$$

Note that $\underline{Y}_i$ can be expressed as

$$\begin{aligned} \underline{Y}_i &= F^a\lambda_i + \underline{e}_i \\ &= \tilde{F}_z^a H_z^{-1}\lambda_i + \underline{e}_i + (F^a H_z - \tilde{F}_z^a)H_z^{-1}\lambda_i \end{aligned} \tag{2.41}$$

Substituting (2.41) in (2.40) we obtain

$$\begin{aligned}
\hat{\lambda}_i &= (\tilde{F}_z^{a'}\tilde{F}_z^a)^{-1}\tilde{F}_z^{a'}\tilde{F}_z^a H_z^{-1}\lambda_i + (\tilde{F}_z^{a'}\tilde{F}_z^a)^{-1}\tilde{F}_z^{a'}\underline{e}_i + (\tilde{F}_z^{a'}\tilde{F}_z^a)^{-1}\tilde{F}_z^{a'}(F^a H_z - \tilde{F}_z^a)H_z^{-1}\lambda_i \\
&= H_z^{-1}\lambda_i + (\tilde{F}_z^{a'}\tilde{F}^a)^{-1}\tilde{F}_z^{a'}\underline{e}_i + (\tilde{F}_z^{a'}\tilde{F}_z^a)^{-1}\tilde{F}_z^{a'}(F^a H_z - \tilde{F}_z^a)H_z^{-1}\lambda_i \\
\Rightarrow \sqrt{T_a}(\hat{\lambda}_i - H_z^{-1}\lambda_i) &= \left(\frac{\tilde{F}_z^{a'}\tilde{F}_z^a}{T_a}\right)^{-1}\frac{1}{\sqrt{T_a}}\tilde{F}_z^{a'}\underline{e}_i + \left(\frac{\tilde{F}_z^{a'}\tilde{F}_z^a}{T_a}\right)^{-1}\left[\frac{1}{\sqrt{T_a}}\tilde{F}_z^{a'}(F^a H_z - \tilde{F}_z^a)\right]H_z^{-1}\lambda_i \\
&= \left(\frac{\tilde{F}_z^{a'}\tilde{F}_z^a}{T_a}\right)^{-1}\frac{1}{\sqrt{T_a}}\tilde{F}_z^{a'}\underline{e}_i + o_p(1) \qquad (2.42)
\end{aligned}$$

We have the above result since the term $\left(\frac{\tilde{F}_z^{a'}\tilde{F}_z^a}{T_a}\right)^{-1}\left[\frac{1}{\sqrt{T_a}}\tilde{F}_z^{a'}(F^a H_z - \tilde{F}_z^a)\right]H_z^{-1}\lambda_i$ is $O_p\left(\frac{T_a^{1/2}}{\delta^2_{N_a T}}\right)$ by boundedness of H_z,λ_i and $\tilde{F}_z^{a'}\tilde{F}_z^a/T_a$ and Lemma 2.4.2. Note that this term is $o_p(1)$ when $\sqrt{T_a}/N \to \infty$.

Now consider $\frac{1}{\sqrt{T_a}}\tilde{F}^{a'}\underline{e}_i$:

$$\begin{aligned}
\frac{1}{\sqrt{T_a}}\tilde{F}_z^{a'}\underline{e}_i &= \frac{1}{\sqrt{T_a}}\sum_{t\in A}\tilde{F}_{zt}e_{it} \\
&= \frac{1}{\sqrt{T_a}}\sum_{t\in A}(\tilde{F}_{zt} - H_z^{'}F_t)e_{it} + H_z^{'}\frac{1}{\sqrt{T_a}}\sum_{t\in A}F_t e_{it} \\
&= H_z^{'}\frac{1}{\sqrt{T_a}}\sum_{t\in A}F_t e_{it} + o_p(1) \qquad (2.43)
\end{aligned}$$

We have the above result since the term $\frac{1}{\sqrt{T_a}}\sum_{t\in A}(\tilde{F}_{zt} - H_z^{'}F_t)e_{it}$ is $O_p\left(\frac{T_a^{1/2}}{\delta^2_{N_a T}}\right)$ by Lemma C.1, which is $o_p(1)$ when $\sqrt{T_a}/N \to \infty$. Substituting (2.43) in (2.42), and using $\frac{\tilde{F}_z^{a'}\tilde{F}_z^a}{T_a} \xrightarrow{p} I$, Lemma 2.4.1 and assumption F_z4 we obtain

$$\sqrt{T_a}(\hat{\lambda}_i - H_z^{-1}\lambda_i) \xrightarrow{d} N(0, Q_z^{-1'}\Phi_{zi}Q_z^{-1}) \quad , \quad i = 1, ..., N_m$$

This completes the proof.

2.11 Appendix 2.5: Proof of Theorem 2.1

Let $\hat{C}_{it}$ and C_{it} denote $\tilde{F}_t^{'}\hat{\lambda}_i$ and $F_t^{'}\lambda_i$, respectively[11].Then we have

$$\begin{aligned}
\hat{C}_{it} - C_{it} &= \tilde{F}_{zt}^{'}\hat{\lambda}_i - F_t^{'}\lambda_i \\
&= (\tilde{F}_{zt} - H_z^{'}F_t)^{'}H_z^{-1}\lambda_i + \tilde{F}_{zt}^{'}(\hat{\lambda}_i - H_z^{-1}\lambda_i) \\
&= (\tilde{F}_{zt} - H_z^{'}F_t)^{'}H_z^{-1}\lambda_i + F_t^{'}H_z(\hat{\lambda}_i - H_z^{-1}\lambda_i) + (\tilde{F}_{zt} - H_z^{'}F_t)^{'}(\hat{\lambda}_i - H_z^{-1}\lambda_i) \\
&= \lambda_i^{'}H_z^{-1'}(\tilde{F}_{zt} - H_z^{'}F_t) + F_t^{'}H_z(\hat{\lambda}_i - H_z^{-1}\lambda_i) + O_p\left(\frac{1}{\delta_{N_aT}\delta_{NT_a}}\right) \qquad (2.44)
\end{aligned}$$

From the proof of Proposition 2.2, we have

$$\delta_{N_aT_a}(\tilde{F}_{zt} - H_z^{'}F_t) = \frac{\delta_{N_aT_a}}{\sqrt{N_a}}\tilde{V}^{-1}\left(\frac{\tilde{F}_z^{'}F}{T}\right)\frac{1}{\sqrt{N_a}}\sum_{k\in A}\lambda_k e_{kt} + O_p(1/\delta_{N_aT_a}) \qquad (2.45)$$

From the proof of Proposition 3, we have

$$\delta_{N_aT_a}(\hat{\lambda}_i - H_z^{-1}\lambda_i) = \frac{\delta_{N_aT_a}}{\sqrt{T_a}}H_z^{'}\frac{1}{\sqrt{T_a}}\sum_{s\in A}F_s e_{is} + O_p(1/\delta_{N_aT_a}) \qquad (2.46)$$

Substituting (2.45) and (2.46) in (2.44) we obtain

$$\begin{aligned}
\delta_{N_aT_a}(\hat{C}_{it} - C_{it}) &= \frac{\delta_{N_aT_a}}{\sqrt{N_a}}\lambda_i^{'}H_z^{-1'}\tilde{V}^{-1}\left(\frac{\tilde{F}^{'}F}{T}\right)\frac{1}{\sqrt{N_a}}\sum_{k\in A}\lambda_k e_{kt} \\
&\quad + \frac{\delta_{N_aT_a}}{\sqrt{T_a}}F_t^{'}H_zH_z^{'}\frac{1}{\sqrt{T_a}}\sum_{s\in A}F_s e_{is} + O_p(1/\delta_{N_aT_a})
\end{aligned}$$

By the definition of H_z, we have $H_z^{-1'}\tilde{V}_z^{-1}(\tilde{F}_z^{'}F/T) = (\Lambda^{a'}\Lambda^a/N_a)^{-1}$. Moreover it can be shown that $H_zH_z^{'} = (F^{'}F/T)^{-1} + O_p(1/\delta_{N_aT}^2)$. Then, the above equation becomes

$$\begin{aligned}
\delta_{N_aT_a}(\hat{C}_{it} - C_{it}) &= \frac{\delta_{N_aT_a}}{\sqrt{N_a}}\lambda_i\left(\frac{\Lambda^{a'}\Lambda^a}{N_a}\right)^{-1}\frac{1}{\sqrt{N}}\sum_{k\in A}\lambda_k e_{kt} \\
&\quad + \frac{\delta_{N_aT_a}}{\sqrt{T_a}}F_t^{'}\left(\frac{F^{'}F}{T}\right)^{-1}\frac{1}{\sqrt{T_a}}\sum_{s\in A}F_s e_{is} + O_p(1/\delta_{N_aT_a})
\end{aligned}$$

The result directly follows by applying the arguments used in Bai (2003) p.167. This completes the theorem.

---

[11]Note that the proof of Theorem 2.1 in this study follows the the proof of Theorem 3 in Bai (2003) (which we refer to as Theorem B3 in section 1.2.2 of Chapter I) very closely which shows the limiting distribution of $\tilde{F}_t^{'}\tilde{\lambda}_i - F_t^{'}\lambda_i$.

3

CORRELATION STRUCTURE UNDER FACTOR MODEL AND SMALL SAMPLE PERFORMANCE OF FACTOR-BASED IMPUTATION ALGORITHM

3.1 Introduction

This chapter serves two purposes: (1) characterization of correlation structure under factor model, (2) how the correlation structure affects the relative performance of FBI and Expectation- Maximization algorithms in small sample. To this end, below we first establish the relation between factor and correlation structures. Then, we show that there is a negative relation between the number of factors and the average absolute correlation (denoted by $\mu_{|\rho|}$) among the series that admit a factor representation. This is an expected result: The rise in the number of factors leads to an increase in potential sources of variation for each series. As a result, the series tend to move together less and the absolute correlation among them decreases. This is a key finding for imputation purposes since all imputation methods exploit the correlation structure of the data set in one way or another. For example, in the case of FBI the auxiliary factors ($\tilde{F}_z$) that are employed in estimating the loadings and conditional mean of missing values are derived by utilizing the covariance matrix of the available series. Similarly, the Expectation-Maximization (EM) algorithm[1] exploits the correlation structure while it generates the imputed values as the conditional mean expectation of missing values given the available ones.

With the above finding in mind, we conjecture that decreasing comovement (measured by $\mu_{|\rho|}$) among series in a data set should decrease the performance of EM algorithm more than the FBI algorithm. The intuition behind this idea is as follows: FBI uses the limited (average) information hidden in the correlation structure more efficiently since it employs a few *high quality* orthogonal variables (principal components) that accumulate and distill most of the available variation in the system as the conditioning variable set. On the other hand, EM's conditioning set is composed of a large number of diluted and *low quality* regressors (observed variables). Additionally, the conditioning variables employed by the EM algorithm are not orthogonal to each other which further weakens use.

In order to test this conjecture, we carry out a simulation experiment where we compare the small

[1]Throughout this chapter (unless otherwise noted) "EM" refers to the EM algorithm for multivariate Gaussian data as considered in Schneider (2001).

sample performance of FBI and EM for different r levels (namely, $r = \{2, 5, 10, 25\}$) under different R^2, T and N values. We show that as the number of factors in factor model increases, performance of both FBI and EM decreases in terms of every measure we consider. This is an expected finding: EM performs worse because as r increases the relation between missing and observed values weaken, while FBI's performance falls due to the increasing burden of estimating a higher number of factors accurately. However, we also show that the relative performance of FBI gets better compared to EM. In other words, FBI is more robust to the adverse affect of higher r levels on the correlation structure than the EM algorithm. This result is a very important finding for implementation purposes: Consider the multi factor risk models which are commonly employed in finance world (by hedge funds, mutual funds, dealers, investments banks etc) to explain and control different sources of portfolio risk and return. Many of these models apply to large security universes and utilize well over 25 factors (e.g. Bloomberg Equity Fundamental Risk Models use more than 25 factors). Unless the entire systematic variance is accumulated in a few factors[2], one should expect a weak correlation structure among the securities (i.e. low $\mu_{|\rho|}$) in a large and well-diversified security universe[3]. Our result implies that under this factor model setting even under the implausible assumption that the (financial) series have multivariate normal distribution, EM algorithm underperforms FBI that does not require any distributional assumption.

Organization of this chapter is as follows: Section 3.2 studies the correlation structure of a factor model and analyzes the behavior of average absolute correlation under equal and unequal factor variance assumptions. Section 3.3 develops a maximum likelihood based method to determine the number of factors in a factor model, and discusses the strong and weak features of it. Section 3.4 carries out a simulation experiment to compare the small sample performance of factor-based imputation algorithms and the EM algorithm under different r levels. We discuss the limitations of the current study and the possible directions to improve and extend them in Section 3.5. We present the concluding remarks in Section 3.6. Finally, Appendices A and B present the proofs and derivations omitted in the text.

---

[2]Note that in crisis periods (e.g. during 2007-2008 financial crisis) market, size, value, profitability and liquidity factors explain a very sizable portion of the overall variance for any given security. As a result, pairwise security correlations increase significantly, and all securities tend to move in tandem.

[3]Note that we assume a reasonably large universe so that one can find a large number of factors that can significantly explain the systematic behavior of the universe.

3.2 Correlation Structure of a Factor Model

This section starts with establishing the relation between factor and correlation structures. We show that correlation (ρ) between any two series from the same factor structure can be broken into two parts: (1) δ, which is a function of factors and loadings only, and (2) θ, which contains the information on percentage of explained systematic variation of the series. Then, we discuss the estimation of δ from data and suggest two easy-to-apply estimation approaches. Next, we consider alternative factor variance distributions, and derive the probability density function of correlation and δ between any two series under the assumption of equal factor variances. Finally, we characterize the behavior of correlation and δ under unequal factor variances via numerical methods.

3.2.1 Basics: The Relation between Factor and Correlation Structures

Consider

$$X_{it} = \lambda_i' F_t + \epsilon_{it} \tag{3.1}$$

where λ_i and F_t are $r \times 1$, and X_{it} and ϵ_{it} are scalars. We assume that F_{qt}, ϵ_{it} and ϵ_{jt} are orthogonal[4] for all i, t and $q = 1, 2, ..., r$. Under the orthogonality assumption, given the loadings the variance of X_{it} is given by[5]

$$\sigma_i^2 = \sum_{q=1}^{r} \lambda_{iq}^2 \sigma_{F_q}^2 + \sigma_{\epsilon i}^2 \tag{3.2}$$

Let R_i^2 denote the percentage of variation explained by the systematic component of X_{it}[6] in factor representation (3.1)

$$R_i^2 = \frac{\sum_{q=1}^{r} \lambda_{iq}^2 \sigma_{F_q}^2}{\sum_{q=1}^{r} \lambda_{iq}^2 \sigma_{F_q}^2 + \sigma_{\epsilon i}^2}$$

Rearranging the above equation we can express σ_i^2 as

$$\sigma_{\epsilon i}^2 = \frac{(1 - R_i^2)}{R_i^2} \sum_{q=1}^{r} \lambda_{iq}^2 \sigma_{F_q}^2$$

Substituting $\sigma_{\epsilon i}^2$ in (3.2), we obtain

---

[4]Note that this is impossible to achieve in a simulation setting when $N > T$

[5]For notational simplicity we suppress the time dependency, t, for the rest of this chapter.

[6]Note that if we consider (3.1) as a regression equation, R_i^2 corresponds to the coefficient of determination.

$$\begin{aligned}\sigma_i^2 &= \sum_{q=1}^{r} \lambda_{iq}^2 \sigma_{F_q}^2 + \frac{(1-R_i^2)}{R_i^2} \sum_{q=1}^{r} \lambda_{iq}^2 \sigma_{F_q}^2 \\ &= \frac{1}{R_i^2} \sum_{q=1}^{r} \lambda_{iq}^2 \sigma_{F_q}^2 \end{aligned} \tag{3.3}$$

Now consider the correlation between series i and j[7]

$$\rho_{ij} = \frac{cov(\lambda_i' F_t + \epsilon_{it}, \lambda_j' F_t + \epsilon_{jt})}{\sqrt{\sigma_i^2}\sqrt{\sigma_j^2}}$$

Substituting (3.3) in the above equation, for $i \neq j$ we obtain

$$\begin{aligned}\rho_{ij} &= \frac{\sum_{q=1}^{r} \lambda_{iq}\lambda_{jq}\sigma_{F_q}^2}{\sqrt{\frac{1}{R_i^2}\sum_{q=1}^{r} \lambda_{iq}^2 \sigma_{F_q}^2}\sqrt{\frac{1}{R_j^2}\sum_{q=1}^{r} \lambda_{jq}^2 \sigma_{F_q}^2}} \\ &= \delta_{ij}\theta_{ij} \end{aligned} \tag{3.4}$$

where

$$\delta_{ij} = \frac{\sum_{q=1}^{r} \lambda_{iq}\lambda_{jq}\sigma_{F_q}^2}{\sqrt{\sum_{q=1}^{r} \lambda_{iq}^2 \sigma_{F_q}^2}\sqrt{\sum_{q=1}^{r} \lambda_{jq}^2 \sigma_{F_q}^2}} \tag{3.5}$$

and

$$\theta_{ij} = \sqrt{R_i^2 R_j^2} \tag{3.6}$$

Therefore, each correlation pair of X is proportional to θ_{ij}. In other words, other things remaining the same, changing θ_{ij} changes each correlation coefficient proportionally. Another important observation is that scaling all the factor variances with the same constant leave the correlation and delta structure intact. In Appendix A we show that the same is true for loading variances, as well. Therefore, the only way to affect the structure of correlation (apart from via θ) is to change the distribution of factor and/or loading variances across factors or loadings. Sections 3.2.4 and 3.2.5, we discuss these observations in detail.

Note that by Cauchy-Schwarz inequality we have that $|\delta_{ij}| \leq 1$ where equality is achieved only when

[7]For notational simplicity we suppress t in ρ_{ij}

i. $r = 1$:

$$\begin{aligned} |\delta_{ij}| &= \frac{|\lambda_i \lambda_j|}{\sqrt{\lambda_i^2}\sqrt{\lambda_j^2}} \\ &= \frac{|\lambda_i||\lambda_j|}{|\lambda_i||\lambda_j|} \\ &= 1 \end{aligned} \tag{3.7}$$

ii. $\lambda_{iq} = \lambda_{jq}$ for all q:

$$\begin{aligned} |\delta_{ij}| &= \frac{\sum_{q=1}^{r} |\lambda_{iq}\lambda_{iq}|\sigma_{F_q}^2}{\sqrt{\sum_{q=1}^{r} \lambda_{iq}^2\sigma_{F_q}^2}\sqrt{\sum_{q=1}^{r} \lambda_{iq}^2\sigma_{F_q}^2}} \\ &= \frac{\sum_{q=1}^{r} \lambda_{iq}^2\sigma_{F_q}^2}{\sqrt{\left(\sum_{q=1}^{r} \lambda_{iq}^2\sigma_{F_q}^2\right)^2}} \\ &= 1 \end{aligned} \tag{3.8}$$

Note that for imputation purposes the sign of the correlation is not important in determining the strength of the statistical association between two series. For example $\rho_{ij} = -0.6$ and $\rho_{ij} = 0.6$ indicates the same degree of association between series i and j. Since θ_{ij} is always non-negative, the same is true for δ_{ij}, as well. Therefore, for simplicity and uniformity of the degree of association for the rest of this study we focus on the behaviors of $|\rho_{ij}|$ and $|\delta_{ij}|$.

Below we first discuss estimation of $|\delta_{ij}|$ given X and r (or an estimate of r). Then, we attempt to characterize the statistical behavior of $|\delta_{ij}|$ under the assumption that $\lambda_{iq} \sim iid\ N(0, \sigma_\lambda^2)$ under two scenarios for the factor variances: First, we consider the case where all the factor variances are equal. We derive the density function of $|\delta_{ij}|$ analytically and uncover how it behaves as r changes. Next, we study the unequal variance case and attempt to unveil the relation between factor variance distribution and the absolute correlation structure.

3.2.2 Sample estimate of $|\delta_{ij}|$

For $T \times N$ data panel $X = (X_{it})$, the $N \times N$ correlation matrix estimate $\hat{\rho}_X$ contains $n = N(N-1)/2$ unique cross-correlation pairs $\{\hat{\rho}_{ij}\}$ and deltas $\{\hat{\delta}_{ij}\}$ with $i, j = 1, ..., N$ and $i > j$. Assuming that an

estimate of r is available[8], one can compute the PCA estimates $\tilde{\Lambda} = (\tilde{\lambda}_{iq})$ and $\tilde{F} = (\tilde{F}_{tq})$ via PCA. Given these estimates, $|\hat{\delta}_{ij}|$ can be calculated in two ways:

a) Using loading and factor estimates: We first calculate sample factor variance estimates

$$\hat{\sigma}_q^2 = (1/(T-1)) \sum_{t=1}^{T} (\tilde{F}_{tq} - \bar{F}_q)^2$$

where $\bar{F}_q = (1/T) \sum_{t=1}^{T} \tilde{F}_{tq}$. Applying (3.5) we obtain $|\hat{\delta}_{ij}|$ as

$$|\hat{\delta}_{ij}| = \frac{\sum_{q=1}^{r} |\tilde{\lambda}_{iq}\tilde{\lambda}_{jq}| \hat{\sigma}_{F_q}^2}{\sqrt{\sum_{q=1}^{r} \tilde{\lambda}_{iq}^2 \hat{\sigma}_{F_q}^2} \sqrt{\sum_{q=1}^{r} \tilde{\lambda}_{jq}^2 \hat{\sigma}_{F_q}^2}} \tag{3.9}$$

b) Using correlation matrix and coefficient of determination Given X and $\tilde{F}$ we first carry out the time series regression

$$X_i = \lambda_i \tilde{F} + \nu_i$$

for each series and calculate coefficient of determination R_i^2 for i^{th} series[9]. Next, we calculate the sample correlation matrix of X, $\hat{\rho}_X$. Finally, using (3.4) and (3.6), $|\hat{\delta}_{ij}|$ can be estimated as

$$|\hat{\delta}_{ij}| = \frac{|\hat{\rho}_{ij}|}{\sqrt{R_i^2 R_j^2}} \tag{3.10}$$

3.2.3 Notation

For the rest of this study we will use the following notation: Let $\delta_{abs} = \{|\delta_{ij}|\}_{i=1,j>i}^{i=n}$ be the $n \times 1$ vector of unique $|\delta_{ij}|$. The population mean of δ_{abs} is given by $\mu_{|\delta|} = E(|\delta_{ij}|)$. Similarly, let $\hat{\delta}_{abs} = \{|\hat{\delta}_{ij}|\}_{i=1,j>i}^{i=n}$ be the $n \times 1$ vector of unique $|\hat{\delta}_{ij}|$, and $\bar{\delta}_{abs} = (1/n) \sum_{m=1}^{n} \hat{\delta}_{abs,m}$ denote the sample mean of $\hat{\delta}_{abs}$. We extend the same logic to correlation, as well: Let $\rho_{abs} = \{|\rho_{ij}|\}_{i=1,j>i}^{i=n}$ be the $n \times 1$ vector of unique $|\rho_{ij}|$, and $\mu_{|\rho|} = E(|\rho_{ij}|)$ denote the population mean of $|\rho_{ij}|$. Finally, sample analogs of ρ_{abs} and $\mu_{|\rho|}$ are denoted by $\hat{\rho}_{abs}$ and $\bar{\rho}_{abs}$, respectively.

3.2.4 Case 1: Equal Factor Variances

Assume that all the factor variances are equal in (3.4), that is, $\sigma_{F_q}^2 = \sigma^2$ for all q. Then we have

$$|\delta_{ij}| = \frac{|\sum_{q=1}^{r} \lambda_{iq}\lambda_{jq}|}{\sqrt{\sum_{q=1}^{r} \lambda_{iq}^2} \sqrt{\sum_{q=1}^{r} \lambda_{jq}^2}} \tag{3.11}$$

---

[8] One can apply the procedure outlined in Bai and Ng (2002) to estimate the number of factors.

[9] Note that loading estimate that would be obtained from this regression is the same with $\tilde{\Lambda}$.

In this case, each $|\delta_{ij}|$ is completely determined by λ_{ij} (3.11), and its entire behavior can be characterized analytically.

Theorem 1. *Suppose* $|\delta_{ij}|$*is given by* (3.11) *above and* $\lambda_{kq} \sim iid\ N(0, \sigma^2_\lambda)$*, for all* $k = 1, 2, ..., r$*. Then for* $r > 1$*,* $i, j = 1, ..., N$ *and* $i \neq j$*,* $|\delta_{ij}|$ *has the following probability density function*

$$f_{|\delta|}(x; r) = \begin{cases} \frac{2(1-x^2)^{(r-3)/2}}{B((r-1)/2, 1/2)} & for\ x \in [0, 1] \\ 0 & otherwise \end{cases} \tag{3.12}$$

where $B((r-1)/2, 1/2)$ *denotes beta function with arguments* $(r-1)/2$ *and* $1/2$*.*

We provide a proof for Theorem 1 in Appendix A.

Figure 3.1 depicts the probability density of $|\delta_{ij}|$ for different r values under the assumption of equal factor variances. As r increases the probability mass of $|\delta_{ij}|$ moves to the left and accumulates mostly around smaller $|\delta|$ values. As a result, one should expect average $|\delta|$ to be smaller as the number of factors increase even if the explanatory power of the factors as a whole (i.e. R^2) remains the same for each series.

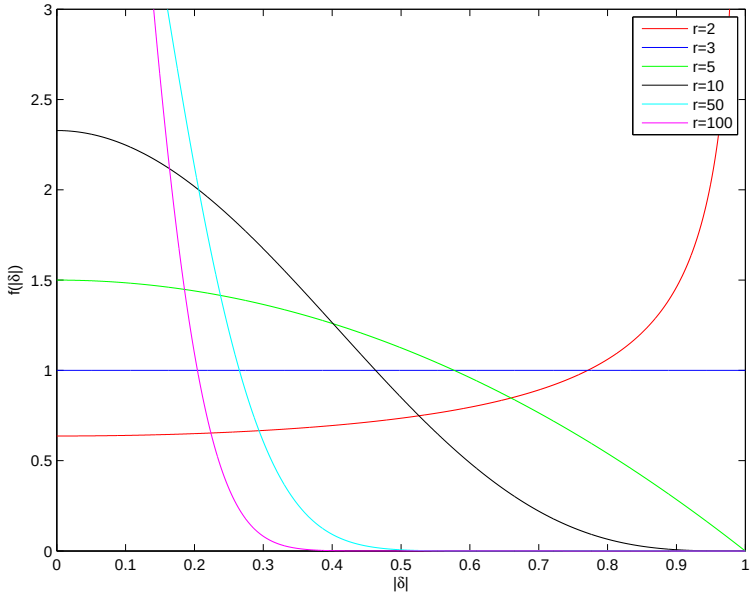

Figure 3.1: Probability Density of $|\delta_{ij}|$ for different r values

We can characterize the distribution of $|\rho_{ij}|$ using the distribution of $|\delta_{ij}|$ and (3.4) as follows:

Corollary 1. *Given* $|\rho_{ij}| = |\delta_{ij}|\theta_{ij}$ *and* $f_{|\delta|}(x;r)$*, under the assumptions of Theorem 1, the conditional probability density function of* $|\rho_{ij}|$ *for* $i, j = 1, ..., N$*,* $i \neq j$ *given* $\theta_{ij} = \theta$*,* $0 \leq \theta \leq 1$ *is as follows*

$$f_{|\rho|}(x;\theta,r) = \begin{cases} \frac{2(\theta^2 - x^2)^{(r-3)/2}}{\theta^{r-2}B((r-1)/2,1/2)} & \text{for } x \in [0,\theta] \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

Appendix A also provides a proof for Corollary 1. In order to preserve space we do not depict the density function of ρ_{ij} for different r and θ values. Basically, for a given θ_{ij} level the density of ρ_{ij} looks like that of δ_{ij}. That is, as r increases the probability density of ρ_{ij} moves toward left and bulk of the distribution accumulates around smaller values. Increasing (decreasing) θ_{ij} expands (contracts) the range of ρ_{ij} as it is evident in the support of $f_{|\rho|}(x;\theta,r)$.

3.2.5 Case 2: Unequal Factor Variances

In this part we study the behavior of $\bar{\delta}_{abs}$ under the assumption of different factor variances. Unlike the previous section, we characterize the behavior of $\bar{\delta}_{abs}$ numerically. We mainly consider two cases: (1) Non-Dominant (General) Factors case, where we let factor variances in a factor model be determined and differentiated from one another via a single half-life parameter, (2) Dominant Factor case, in which one factor explains significantly larger share of overall variation while the remaining factors share the remaining variation equally among themselves. Below, we first show the relation between r and $|\bar{\delta}_{abs}|$ under different half-life parameter choices. Then, for each half-life we find equivalent dominant factor models utilizing Herfindahl concentration index and show the relation between dominant factor variance and $\bar{\delta}_{abs}$. We conclude that similar to the equal-variance case above, as the number of factors increase $\bar{\delta}_{abs}$ (hence, $\bar{\rho}_{abs}$) decreases. However, the rate of decrease in $\bar{\delta}_{abs}$ gets slower as the inequality among factor variances increase. Therefore other things remaining the same, in factor models with unequal factor variances, one should expect a smaller degree of association between the number of factors and $\bar{\rho}_{abs}$.

Non-Dominant (General) Factors

In a given factor structure, magnitude factor variances can be sorted in a decreasing (or non-increasing) way. A parsimonious way to model the factor variance distribution is to use exponential decay.

Consider the following variance sequence

$$\sigma_q^2(h) = \sigma_b^2 \left(\frac{1}{2}\right)^{(q-1)/h} \quad , \quad q = 1, ..., r \tag{3.14}$$

where σ_0^2 is an arbitrary base variance and h is the half-life of a decay process. (3.14) implies a factor variance structure where factor variances decay at a rate determined by the half-life h. For larger values of h the decay is slower and vice versa. For example, for $h = 1$ each factor variance is half of the factor variance preceding it, i.e. $\sigma_q^2(1) = (1/2)\sigma_{q-1}^2(1)$.

The top panel in Figure 3.2 shows the distribution of variances under the general case for different h values when $r = 10$.

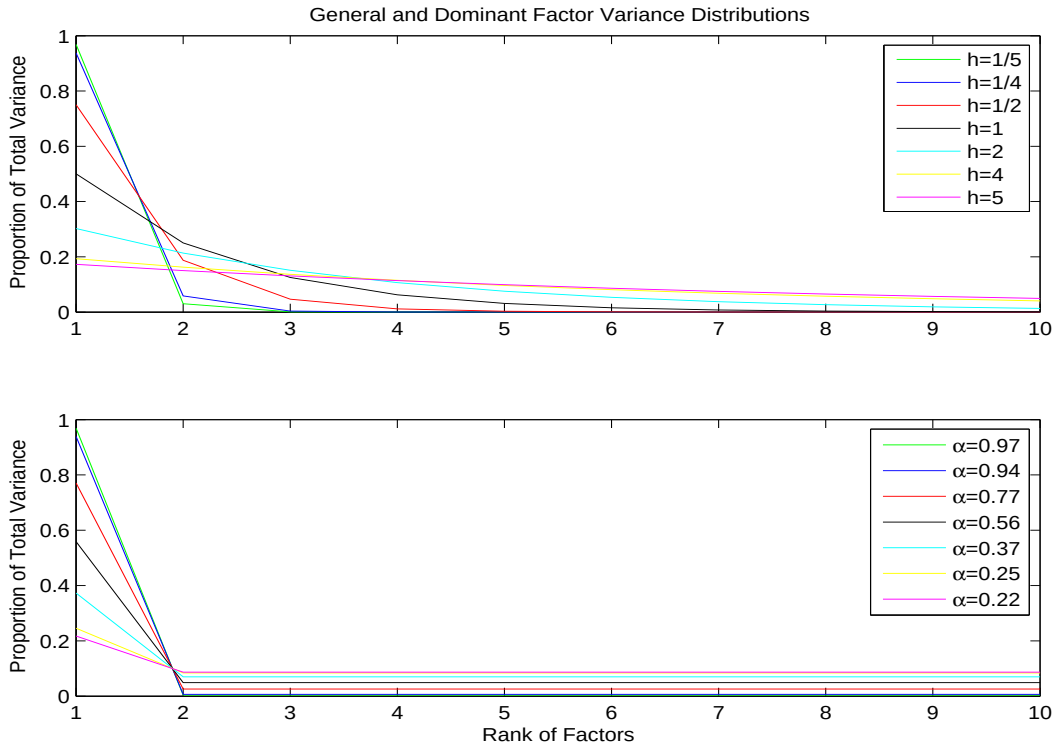

Figure 3.2: Distribution of Factor Variances under General and Dominant Factor Variance Structures (r=10)

Let $p_q(h,r)$ denote the percentage share of the total factor variance explained by the q^{th} factor generated with half-life h when there are r factors

$$\begin{aligned} p_q(h,r) &= \frac{\sigma_q^2(h)}{\sum_{q=1}^{r}\sigma_q^2(h)} \\ &= \frac{\sigma_b^2(1/2)^{(q-1)/h}}{\sum_{q=1}^{r}\sigma_b^2(1/2)^{(q-1)/h}} \\ &= \frac{2^{(1-q)/h}}{\sum_{q=1}^{r}2^{(1-q)/h}} \\ &= \frac{2^{(r-q)/h}(2^{1/h}-1)}{2^{r/h}-1} \end{aligned} \tag{3.15}$$

By definition, we have $\sum_{q=1}^{r} p_q(h,r) = 1$. Under the general variance setting for a given half-life h, the Herfindahl index[10] of $p = (p_1, ..., p_r)$ is given by

$$\begin{aligned} H_g(h,r) &= \sum_{q=1}^{r} p_q^2(h,r) \\ &= \sum_{q=1}^{r} \frac{2^{2(r-q)/h}(2^{1/h}-1)^2}{(2^{r/h}-1)^2} \\ &= \frac{(2^{r/h}+1)(2^{1/h}-1)}{(2^{r/h}-1)(2^{1/h}+1)} \end{aligned} \tag{3.16}$$

By (3.16) above, we obtain the following results:

$$\frac{\partial H_g}{\partial h} < 0 \;\; and \;\; \lim_{h\to\infty} H_g(h,r) = \frac{1}{r} \tag{3.17a}$$

$$\frac{\partial H_g}{\partial r} < 0 \;\; and \;\; \lim_{r\to\infty} H_g(h,r) = \frac{2^{1/h}-1}{2^{1/h}+1} \tag{3.17b}$$

$$\frac{\partial H_g}{\partial h \partial r} < 0 \;\; and \;\; \lim_{h,r\to\infty,\, h=cr} H_g(h,r) = 0 \tag{3.17c}$$

(3.17a) above shows that when r is kept constant, the higher is the value of h the closer are the values of factor variances to each other. Accordingly, for higher h values, we obtain lower Herfindahl index levels indicating a smaller degree of concentration. Similarly, (3.17b) indicates that for a given half-life level increasing the number of factors results in a lower degree of concentration. Finally, (3.17c) shows that letting $h = cr$, $c > 0$ and increasing both r and h decreases the Herfindahl index

[10]Herfindahl index is a concentration measure ranging from 0 to 1. 0 corresponds to virtually no concentration while 1 indicates highest possible degree of concentration. In economics, it is mainly used to measure the amount of competition among firms in a given industry where 0 corresponds to perfect competition while 1 is obtained under monopoly. For a sequence of r percentage shares $(x_1, ..., x_r)$ the Herfindahl index is calculated as $H = \sum_{q=1}^{r} x_q^2$.

indicating a lower degree of factor variance concentration.

In the factor variance structure of a factor model, higher Herfindahl values imply a greater concentration of sources of variation around a small group of factors. Therefore as Herfindahl index of factor variances increase, one would expect to observe higher δ_{abs} values on average. Table 3.1 illustrates the relation between half-life h and $\bar{\delta}_{abs}$ for $r = 10$. Column 1 considers various h values, column 3 shows the associated Herfinfdahl index value and column 4 shows the $\bar{\delta}_{abs}$ value corresponding to the choice of h.

h	$\alpha(h)$	H	$\bar{\delta}_{abs}(h)$	$\bar{\delta}_{abs}(\alpha)$
1/5	0.97	0.94	0.82	0.76
1/4	0.94	0.88	0.76	0.69
1/2	0.77	0.60	0.57	0.49
1	0.56	0.33	0.43	0.37
2	0.37	0.18	0.33	0.30
4	0.25	0.12	0.28	0.27
5	0.22	0.11	0.27	0.26

Table 3.1: Unequal Factor Variances and $\bar{\delta}_{abs}$

Dominant Factor

In factor models built to explain different sources of security returns, it is often the case that the first factor (which is generally interpreted as the "Market" factor) explains a sizable proportion of the overall variance and the remaining variance is more or less democratically distributed across the remaining factors. In order to study the distribution of delta and its relation with r in this setting, we consider a factor distribution where first (dominant) factor accounts for a significant proportion (α) of the overall variance, and the residual variance share $(1 - \alpha)$ is distributed equally among the remaining factors (Therefore each "small" factor receives a share of $(1 - \alpha)/(r - 1)$). Proportion of dominant factor in total variance is given by[11]

$$\alpha = \frac{\sigma_1^2}{\sum_{q=1}^{r} \sigma_q^2}$$

[11]Since factors are orthogonal, all cross-covariances vanish.

Rearranging we obtain

$$\begin{aligned}\sigma_1^2 &= \alpha \sum_{q=1}^{r} \sigma_q^2 \\ &= \alpha\sigma_1^2 + \alpha \sum_{q=2}^{r} \sigma_q^2 \\ &= \frac{\alpha}{1-\alpha} \sum_{q=2}^{r} \sigma_q^2\end{aligned}$$

Assuming that $\sigma_q^2 = \sigma^2$ for $q > 1$, the variance of the dominant factor is given by

$$\sigma_q^2 = \frac{\alpha(r-1)\sigma^2}{1-\alpha}$$

The bottom panel in Figure 3.2 shows the distribution of variances under the dominant factor case for different α values that correspond to h values in the top panel[12].

As in the general variance distribution case, for the dominant factor case we measure the degree of concentration of variation sources through Herfindahl index. The index is calculated as follows

$$\begin{aligned}H_d(\alpha, r) &= \alpha^2 + (r-1)\left(\frac{1-\alpha}{r-1}\right)^2 \\ &= \frac{(r-1)\alpha^2 + (1-\alpha^2)}{r-1} \\ &= \left(\frac{r}{r-1}\right)\alpha^2 - \left(\frac{2}{r-1}\right)\alpha + \left(\frac{1}{r-1}\right)\end{aligned} \tag{3.18}$$

where subscript d signifies the dominant factor case. Note that bringing (3.16) and (3.18) together we can generate a unique α level for each half-life value h

$$H_d(\alpha, r) = H_g(h, r)$$

$$\Rightarrow \left(\frac{r}{r-1}\right)\alpha^2 - \left(\frac{2}{r-1}\right)\alpha + \left(\frac{1}{r-1} - H_g(h, r)\right) = 0$$

Solving the quadratic equation above for α we obtain

$$\alpha(h, r) = \frac{1 + \sqrt{r(r-1)H_g(h, r) + r - 1}}{r} \tag{3.19}$$

[12]See equation (3.19) for the functional relation between α and h for given r value.

Second column of Table 1 shows the dominant factor variances dictated by (3.19) and different values of h. As can be seen, higher values of α correspond to lower values of h. The last column of Table 3.1 reports $\bar{\delta}_{abs}$ values corresponding to various α levels. As expected, at the same Herfindahl levels, $\bar{\delta}_{abs}$ is smaller under the dominant factor case compared to the general case. This implies that for a given $\bar{\delta}_{abs}$ value, $H_g < H_d$ since the degree of concentration is greater under the dominant factor case. For example, for $\bar{\delta}_{abs} = 0.76$ in Table 3.1, we have $H_g = 0.88 < 0.94 = H_d$.

To sum up, the central tenet for both general and dominant factor cases is that as the degree of concentration of factor variances measured by Herfindahl index increases, average delta and correlation levels in a given data set increase. The amount of change in $\bar{\delta}_{abs}$ given a unit change in H differs between general and dominant cases since the structure of underlying factor variance distributions are different.

3.3 Determining the Number of Factors with $f_{|\delta|}$

In this section we propose a new method to estimate the number of factors in a given factor structure utilizing the distribution of δ_{ij} established in Theorem 1. Given a $T \times N$ data set X, we can obtain $\hat{\delta}_{abs}$ which is a $n \times 1$ vector of unique $|\hat{\delta}|$. Let x_i be an element of $\hat{\delta}_{abs}$ with probability density function $f_{|\delta|}(x_i; r)$, $i = 1, .., n$. In Appendix B we show that under this setting the maximum likelihood estimate (MLE) of r, $\hat{r}_{mle}$, is given by the solution of the following implicit form:

$$\psi\left(\frac{\hat{r}_{mle}}{2}\right) - \psi\left(\frac{\hat{r}_{mle}-1}{2}\right) + \frac{1}{n}\sum_{i=1}^{n} ln(1-x_i^2) = 0 \tag{3.20}$$

where $\psi(.)$ is the digamma function[13]. In addition, the asymptotic variance of $\hat{r}_{mle}$ is found as

$$avar(\hat{r}_{mle}) = \frac{4}{\left[\psi^{(1)}\left(\frac{r-1}{2}\right) - \psi^{(1)}\left(\frac{r}{2}\right)\right]} \tag{3.21}$$

where $\psi^{(1)}(.)$ is the trigamma function[14]. By the asymptotic normality property of maximum likelihood estimators, the asymptotic distribution of $\hat{r}_{mle}$ is given by

$$\sqrt{n}\left(\hat{r}_{mle} - r\right) \xrightarrow{d} N(0, avar(\hat{r}_{mle})) \tag{3.22}$$

One can obtain an estimate of $avar(\hat{r}_{mle})$ by replacing the unknown parameter r with its ML estimate

$$\widehat{avar(\hat{r}_{mle})} = \frac{4}{\left[\psi^{(1)}\left(\frac{\hat{r}_{mle}-1}{2}\right) - \psi^{(1)}\left(\frac{\hat{r}_{mle}}{2}\right)\right]}$$

[13]Digamma function is defined as $\psi(a) = \frac{\Gamma'(a)}{\Gamma(a)}$ where $\Gamma(.)$ denotes the Gamma function.

[14]Trigamma function is defined as $\psi^{(1)}(a) = \frac{d}{da}\psi(a) = \frac{d^2}{da^2} ln(\Gamma(a))$

Using (3.22), the approximate small sample distribution and the 95% confidence interval of $\hat{r}_{mle}$ are found as

$$\hat{r}_{mle} \overset{a}{\sim} N(r, \frac{\widehat{avar(\hat{r}_{mle})}}{n})$$

$$CI_{95\%}(r) = \left(\hat{r}_{mle} - 1.96\sqrt{\widehat{avar(\hat{r}_{mle})}/n} \ , \ \hat{r}_{mle} + 1.96\sqrt{\widehat{avar(\hat{r}_{mle})}/n} \right)$$

Note that there is an important caveat with the estimation procedure described above: Given a data set X, one can estimate the unique absolute correlations $|\hat{\rho}_{abs}|$ directly. However, in order to obtain the sample estimate $\hat{\delta}_{abs}$ one needs the estimate of R_j^2 for each series j. R_j^2 can be obtained by regressing each series X_i on common factors F. However, at this point one needs to know "how many" common factors the factor structure has; that is, one requires the knowledge of r. This creates a vicious circle between inputs and outputs of the estimation procedure.

An intuitive solution for this problem can be as follows: Once the ρ_{abs} vector is estimated, we can start with the assumption that $r = 2$, carry out PCA to obtain the factors, apply the above MLE procedure and obtain $\widehat{avar(\hat{r}_{mle})}$. We can then repeat the same procedure for $r > 2$ up to some reasonably high r value, r^{max}, and record the asymptotic variances. At the end, we can choose the $\hat{r}_{mle}$ value with the lowest asymptotic variance.

As can be seen, the estimation procedure described above is (somehow) arbitrary and in practice computationally expensive. An alternative to this procedure could be to utilize $f_{|\rho|}(x; \theta, r)$ with ρ_{abs} that we can directly estimate from the data set X. However, this approach also suffers from the same vicious circle above: The conditioning variable $\theta = \theta_{ij} = \sqrt{R_i^2 R_j^2}$ is not known, and in order to estimate it one needs to know (or estimate) the number of factors[15].

3.4 Performance Comparison of EM and FBI

This section carries out a simulation experiment to compare the small sample performance of EM algorithm and factor-based algorithms (FBI and RFBI) under various r levels. We start with a brief introduction of the theory of EM algorithm, and its application to multivariate normal data. Then, we shortly discuss the Recursive Factor-Based Imputation (RFBI) algorithm, which is a simple extension of FBI. Next, we develop imputation performance measures that gauge the outcome of the competing algorithms from different perspectives. The final part presents and elaborate on the simulation results.

[15]Note that the support set of the distribution function for $|\rho_{ij}|$ contains the parameter of the function, θ. Therefore, one cannot even implement MLE with ρ_{abs} and $\mathrm{f}_{|\rho|}(x; \theta, r)$.

3.4.1 Expectation-Maximization Algorithm[16]

Expectation-Maximization (EM) algorithm (Dempster et al. 1977) is a very general iterative method for finding maximum likelihood estimates of parameters of an underlying distribution in incomplete data problems. The EM algorithm is used with models that involve latent or missing variables an in general unknown parameters of the model are the functions of these variables among others. With the EM algorithm the maximum likelihood estimates of the parameters of any probability distribution can be computed from incomplete data.

The EM algorithm formalizes an old idea for handling missing data problems: (1) Start with an initial estimate of missing values, (2) Replace missing values with estimated values, (3) Estimate parameters, (4) Re-estimate missing values assuming that the estimated parameters are correct. Steps (1) through (4) continue until the algorithm converges.

The steps above can be put in a more technical framework by noting that EM iteration alternates between performing expectation (E) and maximization (M) steps. In the E-step a function is created for the expectation of the log-likelihood that is evaluated using the current estimate for the parameters, while M-step computes parameters maximizing the expected log-likelihood found on the E-step. These parameter estimates are then used to determine the distribution of the latent variables in the next E-step. This recursion between E and M steps continue until the likelihood function stop changing appreciably.

There are mainly two main application areas of the EM algorithm: (i) Maximization of the likelihood is analytically intractable or numerically costly, and the existing problems can be simplified by assuming "hidden" variables. In this interpretation, it is the existence of these latent variables makes the problem "incomplete". (ii) Data has truly missing values. In this study, we consider the latter application.

Below, we first introduce the theory of EM algorithm briefly. Note that this section applies to data sets coming from any distribution function, hence the conclusions obtained would be very general. In the next section we restrict the data set to be distributed as multivariate normal, and present the workings of the algorithm in detail.

---

[16]The exposition in this part is mainly from Little and Rubin (2002).

Theory of EM Algorithm

Let $X = (Y, Z)$ be the complete data and η be the unknown parameter of interest. Assume that Y and Z represent missing and observed portions (series) of X. Then, the distribution of X, $f(X|\eta)$, can be factored as follows:

$$\begin{aligned} f(X|\eta) &= f(Y, Z|\eta) \\ &= f(Z|\eta) f(Y|Z, \eta) \end{aligned} \tag{3.23}$$

where $f(Z|\eta)$ is the density of the observed data and $f(Y|Z, \eta)$ is the density of missing data given the observed data. The likelihood function of (3.23) is given by

$$L(\eta|Y, Z) = L(\eta|Z) f(Y|Z, \eta)$$

from which we obtain the log-likelihood as follows

$$\ell(\eta|Y, Z) = \ell(\eta|Z) + lnf(Y|Z, \eta) \tag{3.24}$$

Given Z and a current estimate of η (denoted by $\eta^{(t)}$), the expectation of both sides of (3.24) over the distribution of missing data Y is found as

$$Q(\eta|\eta^{(t)}) = \ell(\eta|Z) + \int [lnf(Y|Z, \eta)] f(Y|Z, \eta = \eta^{(t)}) dY \tag{3.25}$$

where $Q(\eta|\eta^{(t)}) = \int \ell(\eta|Y, Z) f(Y|Z, \eta = \eta^{(t)}) dY$

EM iteration alternates between performing E and M steps. Basically, E-step finds the conditional expectation of the missing data given the observed data and the current (updated) estimate of the estimated parameters, $\eta^{(t)}$, and M-step performs maximum likelihood estimation as if there were no missing data and the current parameter estimate $\eta^{(t)}$ were equal to the true η.

Under the above setting, (3.25) above represents the expected complete-data likelihood found at the E-step. Given the result of E-step, M-step of EM algorithm determines the parameter estimate of the next iteration, $\eta^{(t+1)}$, by maximizing the expected complete data-likelihood (3.25)

$$Q(\eta^{(t+1)}|\eta^{(t)}) \geq Q(\eta|\eta^{(t)}) \quad , \quad \text{for all } \eta$$

It can be shown that each iteration of EM increases the log-likelihood $\ell(\eta|Z)$, and if $\ell(\eta|Z)$ is bounded, the sequence $\ell(\eta^{(t)}|Z)$ converges monotonically to the stationary value $\ell(\eta|Z)$.

EM Algorithm with Multivariate Normal Data[17]

Let $X \in \mathbb{R}^{T\times N}$ be an incomplete data matrix with T records (samples) consisting of N variables. Assume X comes from a multivariate normal distribution with mean $\mu \in \mathbb{R}^{1\times N}$ and covariance $\Sigma \in \mathbb{R}^{N\times N}$. Below we describe how EM algorithm can be applied to impute the missing values of X.

We can express each record of X $(t = 1, ..., T)$ as $x_t = (y_t, z_t)$ where $y_t \in \mathbb{R}^{1\times N_y}$ and $z_t \in \mathbb{R}^{1\times N_z}$ represent missing and available values[18][19]. Using this decomposition, we can write

$$\mu = (\mu_y, \mu_z)$$

$$\Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{bmatrix}$$

where $\mu_y \in \mathbb{R}^{1\times N_y}$ and $\mu_z \in \mathbb{R}^{1\times N_z}$ represent the mean of y and z, Σ_{yy} and Σ_{zz} are the $N_y \times N_y$ and $N_z \times N_z$ covariance matrices of y and z, and $\Sigma_{yz}(= \Sigma'_{zy})$ is the $N_y \times N_z$ cross covariance matrix between y and z, respectively.

In EM algorithm, the relation between missing and available values for each record x is modeled with the following linear regression

$$y = \mu_y - (z - \mu_z)B + \epsilon \tag{3.26}$$

where $B \in \mathbb{R}^{N_z\times N_y}$ is a matrix of regression coefficients, and $\epsilon \in \mathbb{R}^{1\times N_y}$ is the error vector (assumed to have mean zero) with unknown covariance matrix $C \in \mathbb{R}^{N_z\times N_y}$. Under the normality assumption, we obtain

$$B = \Sigma_{zz}^{-1}\Sigma_{zy} \tag{3.27a}$$

$$C = \Sigma_{yy} - \Sigma_{yz}\Sigma_{zz}^{-1}\Sigma_{zy} \tag{3.27b}$$

Note that B and C cannot be directly estimated because in equation (3.26), μ and Σ are not known. On the other hand, an estimate of μ and Σ cannot be calculated without estimating (imputing) missing values y, which require estimates of B and C. EM algorithm provides a solution by exploiting this vicious circle: It starts with (some) initial estimates of μ and Σ, and from these derives the conditional maximum likelihood estimates of B and C. Given the estimates of the regression coefficients and the

[17]The following two sections on EM and REGEM follow section 2 and 3 of Schneider (2001).

[18]Obviously, if a record does not have any missing values, we have $x_t = z_t$

[19]For the sake of notational simplicity, for the rest of this chapter we drop the subscript t when the meaning is clear.

disturbance covariance, it estimates and fills in the missing values y. Finally, using the imputed data matrix, it re-estimates μ and Σ. Dempster et al. (1977) shows that the EM algorithm process monotonically converges to a fixed point. Below we describe the working of the EM algorithm in detail.

We start the algorithm with some initial mean and variance estimates, $\hat{\mu}^{(0)}$ and $\hat{\Sigma}^{(0)}$. Let $\hat{\mu}^{(j)}$ and $\hat{\Sigma}^{(j)}$ denote the estimates of the mean vector and covariance matrix in the j^{th} iteration of the algorithm. By (3.27a) and (3.27b), we obtain

$$\hat{B}^{(j)} = \left(\hat{\Sigma}_{zz}^{(j)}\right)^{-1} \hat{\Sigma}_{zy}^{(j)} \tag{3.28a}$$

$$\hat{C}^{(j)} = \hat{\Sigma}_{yy}^{(j)} - \hat{\Sigma}_{yz}^{(j)} \left(\hat{\Sigma}_{zz}^{(j)}\right)^{-1} \hat{\Sigma}_{zy}^{(j)} \tag{3.28b}$$

where partitions of $\hat{\Sigma}^{(j)}$ are the estimates of the partitions of Σ. By (3.26), the conditional expectation of the missing values in a given record, $\hat{y}^{(j)} = E(y|z; \hat{\mu}^{(j)}, \hat{\Sigma}^{(j)})$, is given by[20]

$$\hat{y}^{(j)} = \hat{\mu}_y^{(j)} + (z - \hat{\mu}_z)\hat{B}^{(j)} \tag{3.29}$$

Therefore, at a given record the missing values y can be decomposed as

$$\begin{aligned} y &= \hat{y}^{(j)} + \hat{\epsilon}^{(j)} \\ &= \hat{\mu}_y^{(j)} + (z - \hat{\mu}_z)\hat{B}^{(j)} + \hat{\epsilon}^{(j)} \end{aligned}$$

where $\hat{\epsilon}^{(j)}$ is the residual at iteration j. After the missing values in all records x_t are imputed, the updated sample mean of the completed data is given by

$$\hat{\mu}^{(j+1)} = \frac{1}{T} \sum_{t=1}^{T} x_t$$

Let $\hat{S}_t^{(j)} = E(x_t' x_t | z; \hat{\mu}^{(j)}, \hat{\Sigma}^{(j)})$ denote the conditional expectation of the cross product $x_t' x_t$ at time t and iteration j. Then, the updated estimate of the covariance matrix is given by

$$\hat{\Sigma}^{(j+1)} = \frac{1}{T-1} \sum_{t=1}^{T} \{S_t^{(j)} - \hat{\mu}^{(j+1)\prime} \hat{\mu}^{(j+1)}\}$$

Note that the updated covariance matrix estimate has three parts: First two parts involve the

[20]Note that in the equation below (unlike y) z is known (i.e. it does not change across iterations); for this reason, z and $\hat{\mu}_z$ do not have a j superscript.

available data,

$$E(z'z|z;\hat{\mu}^{(j)},\hat{\Sigma}^{(j)}) = z'z$$

$$E(z'y^{(j)}|z;\hat{\mu}^{(j)},\hat{\Sigma}^{(j)}) = z'\hat{y}^{(j)}$$

while the final part is exclusively includes the imputed values

$$\begin{aligned} E(y^{(j)\prime}y^{(j)}|z;\hat{\mu}^{(j)},\hat{\Sigma}^{(j)}) &= E(\hat{y}^{(j)}+\epsilon^{(j)}|z;\hat{\mu}^{(j)},\hat{\Sigma}^{(j)})'E(\hat{y}^{(j)}+\epsilon^{(j)}|z;\hat{\mu}^{(j)},\hat{\Sigma}^{(j)}) \\ &= \hat{y}^{(j)\prime}\hat{y}^{(j)} + E(\epsilon_t^{(j)\prime}\epsilon^{(j)}|z;\hat{\mu}^{(j)},\hat{\Sigma}^{(j)}) \\ &= \hat{y}^{(j)\prime}\hat{y}^{(j)} + \hat{C}^{(j)} \end{aligned} \tag{3.30}$$

which is the sum of the cross-product of the imputed values and the estimated residual covariance matrix, $\hat{C}^{(j)} = Cov(y, y|z;\hat{\mu}^{(j)},\hat{\Sigma}^{(j)})$ at[21] iteration j. Therefore, updated covariance matrix is computed in the same way as the sample covariance matrix of the completed dataset, except that for each record of missing values, the estimated residual covariance matrix $\hat{C}^{(j)}$ is added to the cross-products $\hat{y}_t'\hat{y}_t$.

The next iteration of the EM algorithm is carried out with the updated mean and covariance estimates $\hat{\mu}^{(j+1)}$ and $\hat{\Sigma}^{(j+1)}$. The algorithm stops when the moment estimates $\hat{\mu}$ and $\hat{\Sigma}$ and the imputed values $\hat{y}$ stop changing appreciably.

Regularized Expectation-Maximization Algorithm

When $N > T$, $\hat{\Sigma}_{zz}$ becomes singular (rank-deficient). As a result it cannot be inverted and the estimates of B and C cannot be obtained. Therefore, when the number of variables are larger than the number of records the EM algorithm stops working. The problems involving data sets with $N > T$ are called *ill-posed* problems. Such ill-posed problems can be solved with various regularization methods, which impose additional constraints on the solution.

In order to deal with missing data problems with rank-deficient data sets, Schneider (2001) develops the Regularized Expectation-Maximization (REGEM) algorithm. REGEM is composed of the same steps as the EM algorithm with one exception: in each iteration and for each record with missing values, the inverse matrix $\hat{\Sigma}_{zz}^{-1}$ (superscript is suppressed for simplicity) in the estimate (3.28a) of the regression coefficients is replaced with a regularized inverse as follows

$$\hat{\Sigma}_{zz}^{-1} \longleftarrow \left(\hat{\Sigma}_{zz}^{-1} + s^2\hat{D}\right)^{-1} \tag{3.31}$$

[21]Note that that $\hat{C}^{(j)}$ is equal to the conditional covariance matrix of the imputation error.

where $\hat{D} = diag(\hat{\Sigma}_{zz})$ is the diagonal matrix consisting of the diagonal elements of the covariance matrix $\hat{\Sigma}_{zz}$ and scalar s is a regularization parameter. As can be seen in (3.31), regularization simply inflates the diagonal elements of $\hat{\Sigma}_{zz}$ by a factor of $(1 + s^2)$ before inverting it. The inflation of the diagonal of $\hat{\Sigma}_{zz}$ makes it nonsingular and hence invertible. This method of regularizing the inverse of a matrix, in which a regularized inverse is formed as the inverse of the sum of the matrix and a multiple of a positive definite matrix, is called *ridge regression* in the statistics literature and *Tikhonov regularization* in the literature on numerical linear algebra.

3.4.2 Recursive Factor Based Imputation Algorithm

Recursive Factor Based Imputation (RFBI) algorithm is an extension of standard FBI along the lines of the recursion principle in the EM algorithm. Note that as it is discussed in Chapter I and II, the FBI algorithm starts with a factor estimate ($\tilde{F}_z$) obtained from the complete part of the data set X, (i.e Z). Then the available portion of the missing series are regressed on the corresponding part of $\tilde{F}_z$, and the estimated values are obtained by matching the loadings obtained in this interim regression with the rows of $\tilde{F}_z$ that correspond to the missing rows of the missing variables. Finally, the ID factor estimate ($\tilde{\tilde{F}}_{ID}$) is obtained from the completed data set $\hat{X}$ via regular PCA. Therefore, under FBI one starts with $\tilde{F}_z$ and ends up with $\tilde{\tilde{F}}_{ID}$.

RFBI generalizes this procedure by allowing recursion of the ID factor estimate. Basically, RFBI calls the completed data set and the ID factor estimate as the initial completed data and the initial ID estimate; and denotes them $\tilde{\tilde{F}}_{ID}^{(1)}$ and $\hat{X}^{(1)}$, respectively. Then treating $\tilde{\tilde{F}}_{ID}^{(1)}$ in the same way we treat $\tilde{F}_z$, we obtain a new completed data set $\hat{X}^{(2)}$ and ID factor estimator $\tilde{\tilde{F}}_{ID}^{(2)}$. We repeat this process until the adjacent ID factor estimates are sufficiently close to each other. Following the EM algorithm, we let the RFBI algorithm stop once the Euclidean distance between the adjacent ID factor estimates are smaller than the Euclidean length of the last factor estimate. In the simulation experiment below, we employ RFBI as an alternative method for FBI and EM.

3.4.3 Monte Carlo Experiment

This section analyzes and compares the small sample performance of FBI, RFBI and EM imputation algorithms from different perspectives. Below, we first describe the simulation design we employ and elaborate on its details. Next, we introduce various performance measures to compare the small sample performance of the imputation methods. Finally, we carry out the simulation experiment, and present the results.

Note that in the simulation experiment we apply EM for $T > N$ and REGEM for $N > T$ cases, but report all the results in tables below under the title "EM". For EM and REGEM calculations we utilize the the MATLAB code developed by Schneider (2001)[22].

Simulation Design

We consider the following simple factor structure with equal factor (and loading) variances for generating the incomplete data set X

$$
\begin{aligned}
& X_{it} = \lambda_i' F_t + e_{it} \quad , \quad i = 1, ..., N \quad , \quad t = 1, ..., T \\
& F_t \sim iid\ N(0, \sigma_F^2 \mathbb{I}_r) \quad , \quad \lambda_i \sim iid\ N(0, \sigma_\lambda^2 \mathbb{I}_r) \\
& e_{it} \sim iid\ N\left(0\ ,\ \frac{1 - R_i^2}{R_i^2} \sum_{q=1}^{r} \lambda_{iq}^2 \sigma_F^2\right) \\
& cov(F_t, e_{it}) = 0 \quad \text{for all } i, t
\end{aligned}
\tag{3.32}
$$

where $\mathbb{I}_r$ is the $r \times r$ identity matrix, σ_F^2 and σ_λ^2 are the variances of each loading and factor, respectively (we arbitrarily set them 2 and 3, respectively), and R_i^2 is the percentage contribution of the systematic component to total variance (i.e. coefficient of determination) for series i. Note that setting the variance of e_{it} in the way above guarantees that the coefficient of determination of each series is exactly equal to R_i^2.

Since both factors and idiosyncratic errors are distributed iid normal, each series in X is normally distributed (given the loadings). Therefore, we can directly apply the setting in section 3.4.1 for the EM algorithm. Note that by creating series that are normally distributed with an underlying factor structure, we level the playing field between EM and FBI algorithms. In this sense, the simulation experiment in this section evaluates both methods with the data structures that they are most applicable to.

For all the simulation experiments we consider in this section, we set $N = 500$. We want to compare FBI against both EM and REGEM algorithms; therefore we set T to be 400 (so that $T < N$ and we use REGEM) and 600 (so that $T > N$ and we use EM). For the sake of simplicity, we let all R_i^2 be equal to the same value, R^2, and we consider two cases: (a) $R^2 = 0.1$, which leads to very noisy series due to very limited systematic component contribution to the overall variance, (b) $R^2 = 0.6$, which implies moderate systematic component variation and much less noise.

[22]Please note that the EM-REGEM imputation MATLAB code can be accessed at http://www.clidyn.ethz.ch/imputation/

In generating the data set X using the above setting, we employ a special sampling method that makes sample means, variances and covariances "exactly" match their population counterparts[23]. This method makes it possible to obtain very accurate results with smaller sample sizes and iterations.

Similar to the simulation experiment in Chapter 1, we create a missing block on the northwest corner of the data panel. We let X have 50% missingness (i.e. half of the values in X are missing). Therefore, we have

$$\frac{T_m}{T} = \frac{N_m}{N} = \sqrt{0.5} = 0.7$$

where T_m and N_m are the number of missing rows (samples) and columns (series) in X. Note that with this choice of missingness ratio, the number of missing and available (N_a) columns are $500(0.7) = 350$ and $500(1 - 0.7) = 150$, respectively. Similarly the number of missing and available (T_a) rows are 280 and 120 when $T = 400$, and 420 and 180 when $T = 600$, respectively.

We consider $r = \{2, 5, 10, 25\}$. We choose the maximum r to be equal to 25 since we want to test the relative performance of FBI when $\zeta = \frac{r}{min(T_m, N_m)}$ is large[24].

As in the previous chapters, we can partition the incomplete data set X as

$$X = \begin{bmatrix} Y^m & Z^m \\ Y^a & Z^a \end{bmatrix}$$

where Y^m is the $T_m \times N_m$ missing portion of X while the rest are observed. The completed data set $\hat{X}$ is obtained by simply replacing Y^m with its estimate $\hat{Y}^m$ that is imputed with FBI, RFBI or EM algorithms. Let y_i^m, $\hat{y}_i^m$ $(i = 1, .., N_m)$ and z_k^m $(k = 1, ..., N_a)$ be some $T_m \times 1$ series in Y^m, $\hat{Y}^m$ and Z^m, respectively. In the Monte Carlo experiment, we generate 100 distinct copies of X, create missingness and impute it to obtain $\hat{X}$ using competing imputation methods. We then calculate the performance measures described below section for each copies, and report the median of 100 copies for each method.

---

[23]Note that the orthogonality conditions dictated by the factor structure above (i.e. factors are orthogonal to each other, loadings are orthogonal to each other, and factors and idiosyncratic errors are orthogonal to each other) can be achieved only when $T > N$. When $T < N$, numerically it is impossible to generate N series that are not correlated to each other

[24]FBI and RFBI algorithms' performances are directly determined by the accuracy of the factor and interim loading estimates.

Consistency and asymptotic normality of the factor and interim loading estimates ($\hat{F}_{ID}$ and $\hat{\lambda}_{ID}$), and the FBI estimate of missing values ($\hat{y}$) implicitly require ζ to approach to 0. Note that the maximum value ζ can take is 1 (since estimation of r factors require at least r available rows or columns). For $T = 400$ and $T = 600$, we have $\zeta = \frac{25}{min(120,150)} = 0.21$ and $\zeta = \frac{25}{min(180,150)} = 0.17$, respectively. Both of these values are significantly different from zero and provide ideal test cases for small sample performance.

Imputation Performance Measures

In order to capture the performance of FBI, RFBI and EM in different dimensions, we compare their accuracy of imputation from three different perspectives:

(i) Imputed Values: We measure the accuracy of the missing value estimates with the average Frobenius norm of the difference between missing values and their estimates as follows:

$$m_1 = med\left\{\frac{||\hat{Y}^m - Y^m||}{T_m N_m}\right\} \tag{3.33}$$

Note that the numbers obtained m1 are not cardinally important[25], but they are ordinally comparable. In this regard, the lower is the value of m_1, the better is the performance of a given imputation method.

(ii) Second Moment of the Imputed Values: Each imputation algorithm we employ is a form of conditional mean imputation; that is, they express missing value as $y_{it} = \hat{y}_{it} + \hat{e}_{it}$ and impute it with its conditional mean $\hat{y}_{it} = E(y_{it}|I_t)$ where I_t is the information set available as of time t. In other words, missing values are replaced with estimates coming from the *center* of y_{it}'s distribution, and the imputation algorithm does not account for the variation of y_{it} that comes from $\hat{e}_{it}$. Since $E(y_{it}|I_t)$ explains roughly R^2 of the variation[26], in an ideal setting we expect the variance of imputed values to be around $100R^2\%$ of the true latent variance. That is, we have $\sigma^2_{\hat{y}} \approx R^2\sigma^2_y$ where $\sigma^2_{\hat{y}}$ and σ^2_y are the variances of imputed and missing values, respectively. Since R^2 is in general less than 1, this leads to downward bias in variance estimation. Accordingly, our first performance check in the second moment category considers the ratio of the variance of the imputed values to the true variance to uncover the magnitude of downward bias.

$$m_2 = med\left\{\frac{1}{N_m}\sum_{i=1}^{N_m}\frac{\sigma^2_{\hat{y}_i}}{\sigma^2_{y_i}}\right\} \tag{3.34}$$

In a given copy, we calculate the variance ratio for each missing series and average them out. Then, m_2 is calculated as the median over 100 copies. Other things remaining the same, the closer is the value of m_2 to R^2, the better the imputation algorithm.

---

[25]Since the difference would be affected by the mean and variance of factors and loadings.

[26]We say "roughly" because $E(y_{it}|I_t)$ is not obtained from a proper regression of y_i on a set of explanatory variables (such a regression is obviously infeasible since y_i is not fully observed). Therefore, $E(y_{it}|I_t)$ and $\hat{e}_{it}$ are not necessarily orthogonal, and $E(y_{it}|I_t)$ does not necessarily explain $100 * R^2\%$ of the total variation.

A good imputation algorithm should deliver imputed values that are also highly correlated to the missing values they estimate. To this end, we propose the following performance measure:

$$m_3 = med\left\{\frac{1}{N_m}\sum_{i=1}^{N_m} cor(\hat{y}_i^m, y_i^m)\right\} \tag{3.35}$$

Similar to m_2, in a given copy we calculate the mean of correlations over missing series, and report the median of the averages over the copies. In the perfect case, the correlation between missing an imputed values would be 1 . Therefore, the closer the value of m_3 to 1, the better is the imputation algorithm.

Note that m_3 is closely related to m_2

$$\begin{aligned} cor(\hat{y}_{it}, y_{it}) &= \frac{cov(\hat{y}_{it}, y_{it})}{\sqrt{\sigma_{\hat{y}}^2}\sqrt{\sigma_y^2}} \\ &= \frac{cov(\hat{y}_{it}, \hat{y}_{it} + \hat{e}_{it})}{\sigma_{\hat{y}}\sigma_y} \\ &= \frac{\sigma_{\hat{y}}^2 + cov(\hat{y}_{it}, \hat{e}_{it})}{\sigma_{\hat{y}}\sigma_y} \\ &= \frac{\sigma_{\hat{y}}}{\sigma_y} + \frac{cov(\hat{y}_{it}, \hat{e}_{it})}{\sigma_{\hat{y}}\sigma_y} \end{aligned} \tag{3.36}$$

As can be seen in (3.36), unlike m_2, m_3 accounts for the covariance between imputed values and the imputation error.

For each imputation algorithm, imputed values $\hat{y}_i$ are obtained conditional to the available ones. That is, imputed values are "functions" of available values. As a result, one would expect an inflated correlation between imputed and available values compared to the the correlation between missing and available values. In other words, imputation of missing values tend to create an upward bias for the cross correlation missing and available values. A good imputation method should be able to preserve the correlation structure between missing an available values by minimizing this upward bias. We consider the following measure to test this feature:

$$m_4 = med\left\{\frac{1}{N_m N_a}\sum_{i=1}^{N_m}\sum_{k=1}^{N_a} \frac{cor(\hat{y}_i^m, z_k^m)}{cor(y_i^m, z_k^m)}\right\} \tag{3.37}$$

As can be seen in (3.37), for each copy we calculate $NmNa$ cross-correlations (among each missing and available values) and take their average. The value for m_4 is given by the median of average across copies. As in the case of m_3, the closer the value of m_4 to 1, the better the performance of the imputation algorithm.

(iii) Accuracy of the Factor Estimate obtained from Completed Data: Let $\tilde{\tilde{F}}$ denote the factor estimate obtained from $\hat{X}$ that is generated with one of the imputation methods we consider. As in Chapter 1, we compare the accuracy of the factor estimators by comparing their trace R^2 of multivariate regression of $\tilde{\tilde{F}}$ on F given by

$$m_5 = med\left\{R^2_{\tilde{\tilde{F}},F}\right\} \tag{3.38}$$

As we discuss in Chapter 1, the higher the value of m_5, the more accurate the factor estimator.

Results

We present the results for each performance measure in a separate table. Each table reports the calculated measures for FBI, RFBI and EM for different r, T, N, R^2 and MR combinations. Throughout the simulation experiment, we let $N = 500$ and $MR = 50\%$. Each table is composed of four parts determined by R^2 and T (whose values are given immediately below each part), and results in each part is sorted with respect to r.

Before starting to discuss the performance measures, we elaborate on the sample estimate of mean absolute correlation. Table 3.2 reports $\mu_{|\delta|}$, $\mu_{|\rho|}$, $\bar{\rho}_{abs}$ and[27] the percentage estimation error between $\mu_{|\rho|}$ and $\bar{\rho}_{abs}$. For $R^2 = 0.1$ and $T = 400$, the estimation error is very notable. On the other hand, when either R^2 or T (or both) is increased, the error decreases significantly. Additionally, absolute error generally increases with r for all R^2 and T levels. We can explain these observations as follows: r is the number of sources of variation for a given series. As r is increased, the estimation problem becomes more complicated because the structure that is to be estimated gets more complex. Additionally, raising r while keeping T and N constant, the effective degrees of freedom in estimating r decreases. When these are coupled with low informativeness caused by $R^2 = 0.1$, estimate of absolute correlation become very noisy[28]. A rise in T increases the effective degrees of freedom, and an increase in R^2 decreases the contribution of the noise component in each series; both of these modifications lead to more accurate sample estimate for absolute correlation.

In Table 3.2, last two rows of the first part require particular attention since they have particularly large errors (31.7% and 31.6%). They come from very noisy data sets; accordingly the performance measures obtained for these cases contain large estimation error. Therefore, while comparing the

[27]Note that we calculate $\mu_{|\rho|}$ simply as $\mu_{|\delta|}R^2$.

[28]Additionally, when $T < N$ it is impossible to generate $N = 500$ series whose systematic and idiosyncratic parts are uncorrelated. As a results, R^2 of $N - T$ series are not equal to the intended R^2 value. In order to minimize the effect of this problem, we refine these $N - T$ series by keeping only the ones which are $\pm 1.5R^2$ band.

performance measures of competing imputation algorithms below, we will keep in mind the noise embedded in these two cases and tend to downweight the results corresponding to them.

r	$\mu_{\|\delta\|}$	$\mu_{\|\rho\|}$	$\bar{\rho}_{abs}$	$error$
2	0.6366	0.0637	0.0663	8.6%
5	0.3750	0.0375	0.0426	15.1%
10	0.2587	0.0259	0.0371	31.7%
25	0.1612	0.0161	0.0618	31.6%
$R^2 = 0.1$, $T = 400$				
2	0.6366	0.0637	0.0635	-0.2%
5	0.3750	0.0375	0.0373	-0.5%
10	0.2587	0.0259	0.0256	-1.6%
25	0.1612	0.0161	0.0157	-3.8%
$R^2 = 0.1$, $T = 600$				
2	0.6366	0.3820	0.3836	0.5%
5	0.3750	0.2250	0.2262	1.2%
10	0.2587	0.1552	0.1565	2.1%
25	0.1612	0.0967	0.0999	3.8%
$R^2 = 0.6$, $T = 400$				
2	0.6366	0.3820	0.3814	1.1%
5	0.3750	0.2250	0.2240	-0.5%
10	0.2587	0.1552	0.1538	-1.4%
25	0.1612	0.0967	0.0943	-3.9%
$R^2 = 0.6$, $T = 600$				

Table 3.2: Population and sample estimates of $|\rho|$ and the estimation error

Well-informed about the properties of the imputed data sets from which we calculate the performance measures, now we are ready to compare the competing imputation algorithms below.

(i) Imputed Values Table 3.3 reports the performance results for the estimated missing values as measured by m_1. For all R^2 and T combinations, the average error in imputation monotonically increases with r for all methods. Additionally, increasing R^2 and T greatly improves the performance of all imputation methods.

When $R^2 = 0.1$, FBI and EM have very comparable imputation errors. On the other hand, when R^2 is increased to 0.6 FBI starts to outperform EM, and the relative strength of FBI tends to increase with r.

RFBI has very similar performance with FBI for $R^2 = 0.6$, but underperforms it for $R^2 = 0.1$. The reason for this can be the propagation and amplification of noise in the RFBI iterations in the low R^2

case.

r	FBI	$RFBI$	EM
2	0.0375	0.0377	0.0387
5	0.0603	0.0614	0.0611
10	0.0874	0.0916	0.0857
25	0.1386	0.1534	0.1241
$R^2 = 0.1$, $T = 400$			
2	0.0273	0.0274	0.0283
5	0.0440	0.0448	0.0448
10	0.0639	0.0678	0.0632
25	0.1085	0.1385	0.1001
$R^2 = 0.1$, $T = 600$			
2	0.0102	0.0102	0.0113
5	0.0166	0.0167	0.0194
10	0.0247	0.0248	0.0300
25	0.0444	0.0456	0.0559
$R^2 = 0.6$, $T = 400$			
2	0.0074	0.0074	0.0081
5	0.0120	0.0120	0.0137
10	0.0176	0.0176	0.0208
25	0.0310	0.0317	0.0408
$R^2 = 0.6$, $T = 600$			

Table 3.3: Results for the performance measure m_1

(ii) Variances and Correlations of the Imputed Values Table 3.4 presents the results for measure m_2. For all the cases, as r increases m_2 decreases indicating that variance underestimation gets worse with r. For EM, the main reason behind this observation is the decreasing relations (measured by absolute correlation) between missing and available variables. On the other hand, the main reason for FBI (and RFBI) is the increase in number of factors (both in an absolute sense and relative to N and T) makes their estimation less accurate. As a result, imputation accuracy decreases and variance underestimation worsens. In addition, relative values T and N has almost no effect on the results for all the imputation methods.

Note that for $R^2 = 0.1$ m_2 of EM vanishes for all r level; that is, the relative variance of imputed values becomes zero. Similarly, when R^2 is increased to 0.6, it becomes for high *values*. Therefore EM imputation is prone to completely changing the second moment structure of noisy incomplete data sets. On the other hand, FBI and RFBI perform much better than EM for all r,R^2, T and N levels.

As we mention above, the maximum achievable m_2 value is R^2, and both FBI and RFBI get very close to this upper limit for majority of the cases we consider.

r	FBI	$RFBI$	EM
2	0.07	0.09	0.00
5	0.06	0.09	0.00
10	0.04	0.08	0.00
25	0.02	0.06	0.00
$R^2 = 0.1$, $T = 400$			
2	0.07	0.09	0.00
5	0.05	0.09	0.00
10	0.03	0.08	0.00
25	0.02	0.03	0.00
$R^2 = 0.1$, $T = 600$			
2	0.59	0.59	0.35
5	0.57	0.59	0.27
10	0.55	0.59	0.17
25	0.51	0.58	0.00
$R^2 = 0.6$, $T = 400$			
2	0.59	0.59	0.38
5	0.58	0.59	0.30
10	0.56	0.59	0.23
25	0.50	0.58	0.00
$R^2 = 0.6$, $T = 600$			

Table 3.4: Results for the performance measure m_2

Table 3.5 shows the results for measure m_3. As can be seen, R^2 is the major determinant of the correlation between missing and available values. As in the case of m_2 above, in general imputation becomes less accurate as r increases for all the algorithms considered. Similar to m_2, relative values of T and N do not play a major role.

When the R^2 is high, all three algorithms have similar performances for low r values. However, for $r = 25$ factor-based methods notably outperforms the EM algorithm. For $R^2 = 0.1$, FBI and RFBI have better performance than EM for all r and T values. Once again this shows that EM performs poorly under high r and high noise.

Table 3.6 presents the results for measure m_4. Similar to m_3 results above, the most important determinant for m_4 is R^2; for each method under consideration m_4 is much closer to 1 for $R^2 = 0.6$ than that for $R^2 = 0.1$. In general, cross-correlation overestimate becomes more pronounced as r

r	FBI	$RFBI$	EM
2	0.25	0.26	0.17
5	0.20	0.21	0.04
10	0.12	0.15	-0.03
25	0.02	0.08	0.01
$R^2 = 0.1$, $T = 400$			
2	0.26	0.27	0.20
5	0.20	0.22	0.04
10	0.11	0.15	-0.11
25	-0.04	0.03	-0.24
$R^2 = 0.1$, $T = 600$			
2	0.76	0.76	0.76
5	0.75	0.75	0.74
10	0.72	0.72	0.70
25	0.63	0.64	0.55
$R^2 = 0.6$, $T = 400$			
2	0.77	0.77	0.77
5	0.75	0.75	0.75
10	0.73	0.73	0.72
25	0.66	0.67	0.59
$R^2 = 0.6$, $T = 600$			

Table 3.5: Results for the performance measure m_3

increases. As in the case of m_2 and m_3, relative values of T and N do not affect the measure.

For most of the cases considered in Table 3.6, factor-based methods outperform EM algorithm; the outperformance is particularly recognizable when R^2 is 0.6.

In summary, comparison results for m_2, m_3 and m_4 show that factor based approach preserves the second moment structure of the data set much better compared to the EM algorithm. The relative performance of FBI and RFBI is particularly marked for high r and low R^2.

(iii) Accuracy of the Factor Estimate obtained from Completed Data Table 3.7 shows the comparison of factor estimation accuracy measured by m_5 across the imputation algorithms we consider. As expected, increasing R^2 improves the factor estimation (i.e. m_5 gets closer to 1) while factor estimation worsens as r increases. Relative values of T and N is important for $R^2 = 0.1$, but does not affect performance when R^2 is 0.6 for any of the imputation algorithms.

For most of the cases in Table 3.7, FBI and RFBI performs better than EM, and the outperformance is more pronounced for higher R^2 and r values. For $R^2 = 0.6$FBI and RFBI values arevery close to the

r	FBI	$RFBI$	EM
2	2.94	2.94	3.14
5	2.90	2.88	2.86
10	2.89	2.75	2.68
25	2.99	2.45	17.20
$R^2 = 0.1$, $T = 400$			
2	2.93	2.95	3.18
5	2.89	2.90	2.90
10	2.78	2.77	2.66
25	2.68	2.72	2.67
$R^2 = 0.1$, $T = 600$			
2	1.29	1.29	1.30
5	1.30	1.30	1.32
10	1.32	1.31	1.35
25	1.36	1.35	1.45
$R^2 = 0.6$, $T = 400$			
2	1.29	1.29	1.30
5	1.30	1.30	1.31
10	1.31	1.31	1.34
25	1.36	1.35	1.44
$R^2 = 0.6$, $T = 600$			

Table 3.6: Results for the performance measure m_4

maximum possible m_5 value of 1 and are identical. This points out that the only iteration employed by the FBI algorithm is able to extract most of the relevant information and there is no efficiency resulted from extra recursion utilized in RFBI.

The results for m_1 through m_5 show that other things remaining the same, performance of FBI and EM algorithm are in general adversely affected by the increase in the number of factors. This finding is expected since both imputation algorithms exploit the information embedded in the correlation structure, and increases in r lead to a deterioration in the informativeness of correlations. On the other hand, above results indicate that in small sample FBI algorithm proves to be more robust to increases in r (or decreases in informativeness of correlations) than the EM algorithm, and it outperforms the EM algorithm on significant majority of the cases considered under different performance measures.

These findings verify our initial conjecture that ... (from introduction)...

r	FBI	$RFBI$	EM
2	0.80	0.79	0.74
5	0.76	0.75	0.72
10	0.70	0.70	0.68
25	0.73	0.74	0.69
$R^2 = 0.1$, $T = 400$			
2	0.81	0.81	0.74
5	0.70	0.69	0.65
10	0.56	0.53	0.54
25	0.33	0.28	0.34
$R^2 = 0.1$, $T = 600$			
2	0.98	0.98	0.97
5	0.97	0.97	0.94
10	0.95	0.95	0.90
25	0.91	0.91	0.81
$R^2 = 0.6$, $T = 400$			
2	0.99	0.99	0.98
5	0.97	0.97	0.95
10	0.96	0.96	0.92
25	0.90	0.90	0.79
$R^2 = 0.6$, $T = 600$			

Table 3.7: Results for the performance measure m_5

3.5 Limitations and Further Study

The current study can be extended in many directions. Below we present a brief list of possible extensions:

i. Large sample behavior of the EM algorithm is already known. Once the asymptotic distribution of FBI is derived, EM and FBI algorithms can be compared in terms of their large sample properties under different r, R^2 and MR levels.

ii. We derived the distribution function $f_{|\delta|}(x; r)$ in section **??** under certain simplifying assumptions (equal factor variances, iid normality of loadings, orthogonality of factors and idiosyncratic errors etc...). Among these assumptions, the most restricting one is the equal-variance assumption which is highly unlikely to be satisfied in practice. Future study can extend the distribution function $f_{|\delta|}(x; r)$ with a more flexible and realistic factor variance structure.

iii. Estimating r via using $f_{|\delta|}(x;r)$ or $f_{|\rho|}(x;\theta,r)$ require the knowledge of r, and this creates a vicious circle. Although there are ways to estimate r under these setting, they prove to be hard to implement and costly. Future study can focus on alternative formulations that use easily obtained sample observations and can break the vicious circle mentioned above.

3.6 Conclusion

In this chapter we discuss the relation between factor structure and imputation performance through the medium of the correlation structure of the data set. To this end we first analyze the relation between the factor and correlation structures, and show that increasing the number of factors in the system leads to a decrease in the average absolute correlation ($\mu_{|\rho|}$) under both equal and unequal factor variance cases. $\mu_{|\rho|}$ is an simple way to summarize the degree of comovement among series in a data set. Since the imputation algorithms exploit the relation between missing and available series, a decrease in $\mu_{|\rho|}$ due to increasing r adversely affects their performance.

In simulation experiments, we verify the above reasoning and quantify the effect of the changes in r on $\mu_{|\rho|}$, and the effect of $\mu_{|\rho|}$ on imputation performance under many different scenarios. We show that FBI is on par with EM for high $\mu_{|\rho|}$ values, and significantly outperforms it for low $\mu_{|\rho|}$ values measured in a variety of dimensions (even under the unrealistic assumption that the series follow a multivariate normal distribution). This indicates that FBI is more robust to higher r (lower $\mu_{|\rho|}$) values than EM. In summary, FBI should be preferred to EM for monotone missing data if it is established that the data set under consideration admits a factor structure.

3.7 Appendix 3.1: Probability Density Derivations for Equal Variance Case

3.7.1 Proof of Theorem 1: Probability Density of $|\delta_{ij}|$

Let x represent some $|\delta_{ij}|$ obtained from the data set X. By (3.5), it is easy to see that x is a random draw from $|z_k|/\sqrt{\sum_{q=1}^{r} z_q^2}$ where $z_q \sim iid\ N(0, \sigma_z^2)$, $q = 1, ..., r$, and k is en element in $\{1, ..., r\}$.

Using the distribution of z_q we can write

$$\left(\frac{z_q}{\sigma_z}\right)^2 = c_q$$

where c_q is a chi-square random variable with 1 degrees of freedom. Rearranging the above equation we obtain

$$z_q^2 = \sigma_z^2 c_q \tag{3.39}$$

Now consider

$$x = \frac{|z_k|}{\sqrt{\sum_{q=1}^{r} z_q^2}}$$

Rearranging the expression above, we obtain

$$\begin{aligned} x &= \frac{1}{\sqrt{\frac{1}{z_k^2}\left(z_k^2 + \sum_{q\neq k}^{r} z_q^2\right)}} \\ &= \frac{1}{\sqrt{1 + \sigma_z^2 \sum_{q\neq k}^{r} c_q/\sigma_z^2 c_1}} \\ &= \frac{1}{\sqrt{1 + c_{r-1}/c_1}} \\ &= \frac{1}{\sqrt{1+y}} \end{aligned} \tag{3.40}$$

where c_{r-1} is a chi-squared variable with $(r-1)$ degrees of freedom and y is a random variable with F-distribution with parameters $(r-1)$ and 1. The density of y is given by

$$f_y(y; r-1, 1) = \begin{cases} \frac{(r-1)^{(r-1)} y^{(r-1)/2-1}}{B((r-1)/2,1/2)[1+(r-1)y]^{r/2}} & \text{for } y \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.41}$$

where $B((r-1)/2, 1/2)$ is the beta function with parameters $(r-1)/2$ and $1/2$. We can express y in terms of x as follows:

$$y(x) = \frac{1-x^2}{(r-1)x^2}$$

Given the above expression, we obtain the Jacobian of the transformation as

$$J_{yx} = \frac{dy(x)}{dx} = \frac{-2}{(r-1)x^3}$$

Then, the density of y is calculated as follows

$$\begin{aligned}
f_x(x;r) &= f_y(y(x); r-1, 1) J_{yx} \\
&= \frac{(r-1)^{(r-1)} \left(\frac{1-x^2}{(r-1)x^2}\right)^{(r-1)/2-1} \frac{(-2)}{(r-1)x^3}}{B((r-1)/2, 1/2) \left[1 + (r-1)\left(\frac{1-x^2}{(r-1)x^2}\right)\right]^{r/2}} \\
&= \frac{-2(r-1)^{(r-1)/2)+(3-r)/2-1}(1-x^2)^{(r-3)/2} x^{3-r} x^{-3}}{B((r-1)/2, 1/2) \left[1 + \frac{1-x^2}{x^2}\right]^{r/2}} \\
&= \frac{-2(1-x^2)^{(r-3)/2} x^{-r}}{B((r-1)/2, 1/2) \left[\frac{1}{x^2}\right]^{r/2}} \\
&= \frac{-2(1-x^2)^{(r-3)/2}}{B((r-1)/2, 1/2)}
\end{aligned} \tag{3.42}$$

For $y = 0$ and $y = \infty$ we obtain $x = 1$ and $x = 0$, respectively. Rearranging (3.42) and renaming the distribution function of $|\delta|$ as $f_{|\delta|}(.)$, for $r > 1$ we obtain

$$f_{|\delta|}(x;r) = \begin{cases} \frac{2(1-x^2)^{(r-3)/2}}{B((r-1)/2, 1/2)} & \text{for } x \in [0,1] \\ 0 & \text{otherwise} \end{cases} \tag{3.43}$$

Note that variances of z_q do not enter the density function of x. This shows that scaling the loading variances by the same constant does not affect the distribution of $|\delta|$.

Moment generating function of $|\delta|$ is found as follows

$$\begin{aligned}
M_{|\delta|}(t) = E(e^{tx}) &= \int_0^1 \frac{2e^{tx}(1-x^2)^{(r-3)/2}}{B((r-1)/2, 1/2)} dx \\
&= \frac{\sqrt{\Pi} 2^{(r/2-1)} t^{(1-r/2)} \Gamma((r-1)/2) J((r-2)/2, t) + L(r/2-1, t)}{B((r-1)/2, 1/2)}
\end{aligned} \tag{3.44}$$

where J and L are Bessel function of first kind and modified Struve function, respectively.

Mean and variance of $|\delta|$ are obtained as follows

$$\begin{aligned}\mu_{|\delta|} = E(|\delta|) &= \int_0^1 \frac{2x(1-x^2)^{(r-3)/2}}{B((r-1)/2,1/2)}dx \\ &= \frac{2}{(r-1)B((r-1)/2,1/2)}\end{aligned} \tag{3.45}$$

$$\begin{aligned}\sigma^2_{|\delta|} = E((|\delta|-\mu_{|\delta|})^2) &= \int_0^1 \frac{(x-\mu_x)^2 2(1-x^2)^{(r-3)/2}}{B((r-1)/2,1/2)}dx \\ &= \frac{2^r\left[\Gamma((r+1)/2)^2\Pi - 2\Gamma(r/2+1)\Gamma(r/2)\right]}{(r-1)B((r-1)/2,1/2)\Gamma(r+1)\Pi}\end{aligned} \tag{3.46}$$

3.7.2 Proof of Corollary 1: Probability Density of $|\rho_{ij}|$

For notational simplicity let $x = |\rho|$ and $w = |\delta|$. Then, the relation between $|\rho|$ and $|\delta|$ in (3.4) can be rewritten as

$$w(x) = \frac{x}{\theta} \tag{3.47}$$

where $0 \leq \theta \leq 1$. Therefore the Jacobian of the transformation is given by

$$J_{wx} = \frac{dw(x)}{dx} = \frac{1}{\theta}$$

Using the J_{wx} and the distribution of $|\delta|$, we can obtain the distribution of $|\rho|$ for a given θ as follows

$$\begin{aligned}f_x(x) &= f_w(x)J_{wx} \\ &= \frac{2(1-\left(\frac{x}{\theta}\right)^2)^{(r-3)/2}}{B((r-1)/2,1/2)}\frac{1}{\theta} \\ &= \frac{2(\theta^2-x^2)^{(r-3)/2}\theta^{3-r}}{B((r-1)/2,1/2)\theta} \\ &= \frac{2(\theta^2-x^2)^{(r-3)/2}}{\theta^{r-2}B((r-1)/2,1/2)}\end{aligned} \tag{3.48}$$

For $w = 0$ and $w = 1$ we have $x = 0$ and $x = \theta$, respectively. Rearranging and renaming (3.48), we obtain the distribution of $|\rho|$ for $r > 1$ as follows

$$f_{|\rho|}(x;\theta,r) = \begin{cases} \frac{2(\theta^2-x^2)^{(r-3)/2}}{\theta^{r-2}B((r-1)/2,1/2)} & \text{for } x \in [0,\theta] \\ 0 & \text{otherwise} \end{cases} \tag{3.49}$$

3.8 Appendix 3.2: Maximum Likelihood Estimation of r

Let $\hat{\delta}_{abs} = \{x_1, x_2, , x_n\}$ be a $n \times 1$ random sample vector of unique absolute deltas obtained from the data set X. We assume that X has an equal-variance factor structure; therefore, x_i has the probability density function $f_{|\delta|}(x_i; r)$, $i = 1, 2, ..., n$ described in Theorem 1. Then, the likelihood function of x is given by

$$\begin{aligned} L(r|x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} f_{|\delta|}(x_i; r) \\ &= \prod_{i=1}^{n} \frac{2(1-x_i^2)^{(r-3)/2}}{B((r-1)/2, 1/2)} \end{aligned} \tag{3.50}$$

Using (3.50), the log-likelihood function of x can be written as

$$\begin{aligned} \ell(r|\boldsymbol{x}) &= \sum_{i=1}^{n} ln\left(\frac{2(1-x_i^2)^{(r-3)/2}}{B((r-1)/2, 1/2)}\right) \\ &= \sum_{i=1}^{n} \left(ln(2) - ln(B((r-1)/2, 1/2)) + \frac{(r-3)}{2} ln(1-x_i^2)\right) \\ &= nln(2) - nln(B((r-1)/2, 1/2)) + \frac{(r-3)}{2} \sum_{i=1}^{n} ln(1-x_i^2) \end{aligned} \tag{3.51}$$

Differentiating $\ell(r|\boldsymbol{x})$ with respect to r, we obtain

$$\begin{aligned} \frac{\partial \ell(r|\boldsymbol{x})}{\partial r} &= -\frac{n}{2(B((r-1)/2, 1/2)} \left(\frac{\Gamma'((r-1)/2)}{\Gamma((r-1)/2)} - \frac{\Gamma'(r/2)}{\Gamma(r/2)}\right) B((r-1)/2, 1/2) + \frac{1}{2} \sum_{i=1}^{n} ln(1-x_i^2) \\ &= -\frac{n}{2}\left[\psi\left(\frac{r-1}{2}\right) - \psi\left(\frac{r}{2}\right)\right] + \frac{1}{2} \sum_{i=1}^{n} ln(1-x_i^2) \end{aligned} \tag{3.52}$$

where $\Gamma(.)$ and $\psi(.)$ are the gamma and digamma functions, respectively[29]. Equating (3.52) to 0, we obtain the following implicit form for $\hat{r}_{mle}$, the maximum likelihood estimator of r:

$$\psi\left(\frac{\hat{r}_{mle}}{2}\right) - \psi\left(\frac{\hat{r}_{mle} - 1}{2}\right) + \frac{1}{n} \sum_{i=1}^{n} ln(1-x_i^2) = 0 \tag{3.53}$$

Using (3.53) one cannot obtain an explicit functional form $\hat{r}_{mle}(\delta_a bs)$, but the estimate $\hat{r}$ can still be calculated numerically.

---

[29] In differentiating the beta function we applied the chain rule $\frac{\partial B((r-1)/2,1/2)}{\partial r} = \frac{\partial B((r-1)/2,1/2)}{\partial [(r-1)/2]} \frac{\partial [(r-1)/2]}{\partial r} = \frac{1}{2} \frac{\partial B((r-1)/2,1/2)}{\partial [(r-1)/2]}$, and used the following property of the beta function: $\frac{\partial B(a,b)}{\partial a} = \left[\frac{\Gamma'(a)}{\Gamma(a)} - \frac{\Gamma'(a+b)}{\Gamma(a+b)}\right] B(a,b)$

Note that the score function associated with the observation x_i is given by $S(r|x_i) = \frac{\partial \ell(r|x_i)}{\partial r}$. In our case, we can obtain the score function for x_i from (3.52) as

$$S(r|x_i) = \frac{1}{2}\left[\psi\left(\frac{r-1}{2}\right) - \psi\left(\frac{r}{2}\right)\right] + \frac{1}{2}ln(1-x_i^2)$$

The information matrix of the sample is given by $I(r|\delta_{abs}) = nvar(S(r|x_i)$. In order to calculate information matrix for the sample δ_{abs} we need the following lemma.

Lemma 1. *Let x_i have the density function $f_{|\delta|}(x_i, r)$. Then, the density function of $y_i = ln(1-x_i^2)$ is given by:*

$$f_{|y|}(y_i; r) = \begin{cases} \frac{e^{y(r-1)/2}}{B((r-1)/2,1/2)(1-e^y)^(1/2)} & \text{for } y \in (-\infty, 0] \\ 0 & \text{otherwise} \end{cases}$$

Additionally, the variance of y_i is given as follows

$$var(y_i) = \left[\psi^{(1)}\left(\frac{\hat{r}_{mle}-1}{2}\right) - \psi^{(1)}\left(\frac{\hat{r}_{mle}}{2}\right)\right]$$

Using transformation $y_i = ln(1-x_i^2)$ and Lemma 1, the information matrix of δ_{abs} is calculated as

$$\begin{aligned} I(r|\delta_{abs}) &= n\ var\left(S(r|x_i)\right) \\ &= n\ var\left(\frac{1}{2}\left[\psi\left(\frac{r-1}{2}\right) - \psi\left(\frac{r}{2}\right)\right] + \frac{1}{2}ln(1-x_i^2)\right) \\ &= \frac{n}{4}var\left(ln(1-x_i^2)\right) \\ &= \frac{n}{4}var\left(y_i\right) \\ &= \frac{n}{4}\left[\psi^{(1)}\left(\frac{r-1}{2}\right) - \psi^{(1)}\left(\frac{r}{2}\right)\right] \end{aligned} \tag{3.54}$$

Using the above information the asymptotic variance of the maximum likelihood estimator of r is given by

$$\begin{aligned} avar(\hat{r}_{mle}) &= \frac{1}{n}I(r|\delta_{abs})^{-1} \\ &= \frac{4}{\left[\psi^{(1)}\left(\frac{r-1}{2}\right) - \psi^{(1)}\left(\frac{r}{2}\right)\right]} \end{aligned} \tag{3.55}$$

BIBLIOGRAPHY

Afifi, A. A., and Elashoff, R. M. (1966): "Missing Observations in Multivariate Statistics: I. Review of the Literature," *Journal of American Statistical Association*, 61, 595-604.

—— (1967): "Missing Observations in Multivariate Statistics: II. Point Estimation in Simple Linear Regression," *Journal of American Statistical Association*, 62, 10-29.

—— (1969): "Missing Observations in Multivariate Statistics: III. Large Sample Analysis of Simple Linear Regression," *Journal of American Statistical Association*, 64, 337-358.

Anderson, T.W. (2002): *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135-172.

Bai, J., and Ng, S. (2006): "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74, 1133-1150.

—— (2004): "Confidence Intervals for Diffusion Index Forecasts with a Large Number of Predictors," Working Paper, retrieved from http://ideas.repec.org/p/wpa/wuwpem/0408006.html

—— (2008): "Large Dimensional Factor Analysis," *Foundations and Trends in Econometrics*, 3, 89-163.

Banbura, M., and Modugno, M. (2010): "Maximum Likelihood Estimation of Factor Models on Data Sets with Arbitrary Pattern of Missing Data," European Central Bank Working Paper Series No: 1189.

Bernanke, B., and Boivin, J. (2003): "Monetary Policy in a Data Rich Environment," *Journal of Monetary Economics*, 50, 525-546.

—— (2005): "Factor Augmented Vector Autoregressions (FVARs) and the Analysis of Monetary Policy," *Journal of Econometrics*, 132, 169-194.

Boivin, J., and Ng, S. (2006): "Are more data always better for factor analysis?," *Quarterly Journal of Economics*, 120, 387-422.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997): *The Econometrics of Financial Markets*. New Jersey: Princeton University Press.

Chamberlain, G., and Rothschild, M. (1983): "Arbitrage, Factor Structure and Mean- Variance Analysis in Large Asset Markets," *Econometrica*, 51, 1281-1304.

Connor, G., and Korajczyk, R. (1986): "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis," *Journal of Financial Economics*, 15, 373-394.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977): "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, 39, 1-38.

Christadoro, R., Forni, M, Reichlin, L., and Giovanni, V. (2002): "A Core Inflation Index for the Euro Area", manuscript, www.dynfactor.org

Giannone, D., Reichlin, L., and Sala, L. (2002): "Tracking Greenspan: Systemic and Unsystemic Monetary Policy Revisited", manuscript, www.dynfactor.org

Hamilton, J. D. (1994): *Time Series Analysis*. New Jersey: Princeton University Press.

Hayashi, F. (2000): *Econometrics*. New Jersey: Princeton University Press.

Jolliffe, I.T. (2002): *Principal Component Analysis*. New York: Springer.

Kelejian, H. H. 1969): "Missing Observations in Multivariate Regression: Efficiency of a First-Order Method," *Journal of American Statistical Association*, 64, 1609-1616.

Lehman, E. L. (1999): *Elements of Large Sample Theory*. New York: Springer.

Little, R.J.A. 1992): "Regression with Missing X's: A Review", *Journal of American Statistical Association*, 87, 1227-1237.

Little, R.J.A., and Rubin, D.B. (2002): *Statistical Analysis with Missing Data*. New York: Wiley.

Schneeweiss, H., and Mathes, H. (1995): "Factor Analysis and Principal Components," *Journal of Multivariate Analysis*, 55, 105-124.

Schneider, T. (2001): "Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values," *Journal of Climate*, 14, 853-871.

Schumacher, C. (2005): "Forecasting German GDP Using Alternative Factor Models Based on Large Data Sets," Deutsche Bundesbank Discussion Paper 24/2005

Stock, J., and Watson, M.D. (2002a): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167-1179.

—— (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20, 147-162.

Srivastava, V. K., and Toutenburg, H. (2005): "On the First Order Regression Procedure of Estimation for Incomplete Regression Models," *Statistical Papers*, 46, 303-307.

Toutenburg, H., Srivastava, V. K., and Shalab, C.H (2005): "Estimation of Parameters in Multiple Regression with Missing Covariates Using a Modified First Order Regression Procedure," *Annals of Economics and Finance*, 6, 289-301.

Tsay, Ruey, S. (2005): *Analysis of Financial Time Series*. New York: Wiley.

Walczak, B., and Massart, D. L. (2001): "Dealing with Missing Data: Part I," *Chemometrics and Intelligent Laboratory Systems*, 58, 15-27.

—— (2001): "Dealing with Missing Data: Part II," *Chemometrics and Intelligent Laboratory Systems*, 58, 29-42.

Zivot, E., and Wang, J. (2006): *Modeling Financial Time Series with S-PLUS*. New York: Springer.