5-14-2017

ESSAYS ON MISSPECIFICATION IN HIGH DIMENSIONAL ECONOMETRICS AND ASSET PRICING

JAEWOO OH
Syracuse University

Abstract

This dissertation comprises three essays that examine misspecification issues in high dimensional econometrics and asset pricing. The first two essays theoretically diagnose the misdetection risk of the number of factors in high dimensional factor models and propose procedures for correcting such misspecification. In particular, the second essay extends the first one, which focuses on over-detection, to under-detection so that it formulates a non-asymptotic bound on the overall misdetection probability of the number of factors and decides the optimal penalization to minimize its upper bound. The third essay revisits the Recovery theorem of Ross (2015) on the identification of the physical probability distribution of stock returns. It suggests a novel procedure for applying the theorem to the Gaussian affine term structure but empirically verifies that the physical probability is falsely identified by the Recovery theorem. From such misspecification, however, we learn that term premia can be decomposed into nearly constant short-term premia regarding transitory shocks and highly volatile long-term premia regarding martingale shocks. This result finally demonstrates that long-term risk matters for asset pricing.

ESSAYS ON MISSPECIFICATION IN HIGH DIMENSIONAL ECONOMETRICS AND ASSET PRICING

by

Jaewoo Oh

B.A., Yonsei University, 2000
M.P.A., Syracuse University, 2013

Dissertation
Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics

Syracuse University
June 2017

Acknowledgements

I would like to express my sincerest gratitude to my primary advisor Chihwa Kao for his invaluable assistance and guidance without which this dissertation would not have been possible. Special thanks go to my co-advisor Yoonseok Lee. He is not only an academic advisor but also an excellent mentor. I am also thankful to committee members Badi Baltagi, Bill Horrace, Jan Ondrich and Peter Wilcoxen for their sincere interest in my work and many insightful suggestions. I will never forget my friends Sanggon Na, Jaeyoon Lee and Seokchae Hwang, the Economics Department staff, and my many colleagues whose support helped me overcome all the difficulties during my Ph.D. study. I would like to thank my parents Jinhaeng Kim and Hyunsub Oh, and my parents-in-law Jinsook Ryu and Hwuinam Moon who have always loved me unconditionally.

Finally, this dissertation is lovingly dedicated to my brilliant and outrageously loving and supportive wife Soohyun. You have been patient with me when I'm frustrated, you celebrate with me when I make even tiny things, and you are there whenever I need you. You have been my best cheerleader.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation examines misspecification issues in two contexts: (i) signal (or equivalently factor) detection in high dimensional factor models and (ii) the identification of the physical probability distribution of stock returns in the asset pricing literature.

The first essay revisits the panel information criteria (IC) proposed by Bai and Ng (2002), which is a popular estimator for the number of factors in high dimensional factor models, and studies its over-detection risk in finite samples. First, we analyze the finite sample performance of IC by computing the over-detection probability bound. In particular, we specify the asymptotic over-detection condition of IC in terms of eigenvalues coming from pure noise and then derive the computable formula for a non-asymptotic upper bound on the overestimation probability by adopting random matrix theory. We show that unless the sample size is sufficiently large, the overestimation probability is not negligible even for the case in which factors have strong explanatory power. Second, we show that for small sample sizes the over-detection risk of IC is significantly reduced by the degrees of freedom adjustment in the penalty of the original criteria. Finally, we propose modified information criteria (MIC) as a practical guide to improving the finite sample performance of IC. Simulations show that our MIC outperforms IC for the case with weakly serially or cross-sectionally correlated errors as well as i.i.d. errors.

The second essay examines the misdetection risk of the panel information criteria (IC) proposed by Bai and Ng (2002) for detecting the number of factors in high dimensional factor models and examines the optimal penalty to minimize an upper bound on the misdetection

probability of the IC estimator in finite samples. This study extends the first chapter, which analyzed the finite sample performance of the IC estimator regarding its over-detection risk, to the comprehensive misdetection risk considering under-detection risk as well. We derive the computable formula for a non-asymptotic upper bound on the misdetection probability by employing recent results from random matrix theory. Using the formula, we analyze the misdetection risk of the IC estimator and achieve the minimum upper bound of the misdetection probability by finding the optimal weight for the penalty function. Our numerical examples suggest that modified criteria with the optimized penalization improve the finite sample performance of the original IC estimator.

In my third essay, we revisit the Recovery theorem on the identification of the physical probability distribution of stock returns, proposed by Ross (2015). First, its applicability in fixed-income markets is considered. We suggest a new procedure for applying the Recovery theorem to the Gaussian affine term structure. As a result, we can recover a particular probability distribution and decompose forward rates into the investors' short-rate expectations and term premia under this recovered probability measure. Next, the reliability of the Recovery theorem is examined. In particular, we study its misspecification issue in line with the claim of misspecified recovery by Borovička, Hansen, and Scheinkman (2015). Our empirical result verifies that what Ross really recovers is not the physical probability but the long-term risk-neutral probability which absorbs compensation for exposure to permanent shocks. In consequence, we can decompose forward term premia into nearly constant short-term risk premia associated with transitory shocks and highly volatile long-term risk premia corresponding to permanent shocks. Finally, we find that a secular decline in forward rates is mostly attributed to investors' short-rate expectations under the long-term risk-neutral probability measure, and all important variations in term premia can be captured by long-term risk premia. Concisely, long-term risk matters for asset pricing.

Chapter 2

On the Over-detection Probability of the Number of Factors

2.1 Introduction

This chapter examines the issue of the misdetection of the number of factors in large dimensional panels. Our analysis focuses on a popular estimator for the number of factors based on a model selection problem, the panel information criteria (IC) proposed by Bai and Ng (2002). In particular, we address the following questions: (i) how to diagnose the over-detection risk of the IC estimator theoretically, and (ii) how to improve the finite sample performance of IC when its misdetection risk is not negligible.

To diagnose the over-detection risk of the IC estimator, we formulate and compute the upper bound on the probability of overdetecting the number of factors by adopting theoretical results from random matrix theory (e.g., Geman, 1980; Tracy and Widom, 1996; Johnstone, 2001; Baik, Arous, and Péché, 2005; Baik and Silverstein, 2006; Ledoux, 2007; Paul, 2007; Karoui, 2008; Ma, 2012). Our analysis is inspired by the digital signal processing literature regarding signal detection analysis (e.g., Kritchman and Nadler, 2009; Nadler, 2008, 2010). To increase the precision of the estimate in finite samples, we improve the penalty for overfitting of the original criteria by adjusting degrees of freedom for the number of factors. This approach is motivated by previous studies on model selection criteria (e.g., Ng and Perron, 2005; Nadler, 2010).

Large dimensional datasets contain not only important signals but also irrelevant disturbances, namely noise. The beauty of factor analysis such as principal components analysis (PCA) is to provide an efficient data reduction device for big data analysts. That is, when the true number of factors is given, PCA reduces a large number of variables to a small number of factors while preserving most of the information in the original data; however, the true number of factors is unknown in large factor models and consequently should be estimated. Thus, if the estimate of factor numbers is misspecified, the benefits of data reduction can be undermined. Specifically, when the number of factors is overestimated, users suffer from the loss of degrees of freedom. In this regard, Onatski (2015) examined the consequences of the misspecified number of factors for the loss of asymptotic efficiency in the principal

components estimator.

Such misspecification is particularly an issue in small samples. Several researchers have already proposed asymptotically consistent estimators for the number of factors (e.g., Bai and Ng, 2002; Kritchman and Nadler, 2009; Onatski, 2010, 2012; Ahn and Horenstein, 2013; Choi and Jeong, 2013; Harding, 2013); however, their estimators tend to over or under detect the number of factors to some extent in finite samples. Bai and Ng (2002) provided simulation evidence for the misdetection of their IC estimator. Besides, a few simulation studies show that misspecification is likely to get worse if errors are serially or/and cross-sectionally correlated, or if the explanatory power of the factors does not strongly dominate the explanatory power of the idiosyncratic components (e.g., Onatski, 2010; Greenaway-McGrevy, Han and Sul, 2012; Ahn and Horenstein, 2013). On the other hand, there is no computable guidance on how frequently misspecification occurs subject to different sample sizes. As a consequence, it is theoretically unknown how to improve the finite sample performance of existing estimators.

In this chapter, we derive the computable formula for an upper bound on the over-detection probability of the IC estimator by employing some results from random matrix theory. By using this formula, we can analyze the detection performance of the IC estimator in finite samples. This chapter provides practical users with the numerical examples of over-detection probability bounds subject to various sample sizes and numbers of factors. These examples show that when the sample size is not sufficiently large, there exists a non-negligible overestimation risk even for the case in which each factor has a nontrivial contribution to variation in the data. Moreover, this chapter provides practitioners with a practical guide to correcting such misspecification. We show that the degrees of freedom adjustment in the penalty term of the original IC criteria leads to improved penalization for overfitting and consequently decreases the overestimation probability substantially. The over-detection probability bounds of such modified criteria are also measured by our formula. The results indicate that for the case with i.i.d. errors, our modified estimator performs better than the

original IC estimator when the sample size is small. Moreover, Monte Carlo simulations show that it also outperforms the IC estimator in the presence of weak serial or cross-sectional correlation, or both in the error components.

The rest of the chapter is organized as follows. In Section 2.2, we describe our factor model and assumptions. Section 2.3 introduces the panel information criteria (IC) for the number of factors of Bai and Ng (2002) and proposes its eigenvalue representation. Section 2.4 presents an asymptotic expression for the overestimation probability of the IC estimator. Section 2.5 reviews recent results from random matrix theory as mathematical preliminaries. We derive the computable formula for an upper bound on the overestimation probability and analyze the detection performance of IC for finite values of both dimensions in Section 2.6. Section 2.7 proposes a modified estimator and shows its better performance in small samples via Monte Carlo simulations as well as theoretical computations. Section 2.8 provides a summary and discussion. All the proofs are given in the Appendix.

A word on notation. Ordinary limits are denoted by $\rightarrow$ while almost sure convergence, also known as convergence with probability one (w.p.1), is denoted by $\xrightarrow{a.s}$. Convergence in distribution is denoted by $\xrightarrow{d}$. Orders of magnitude for a sequence converging in probability are denoted by O_p and o_p. $tr(A)$ is the trace of a matrix A. The transpose operator is denoted by a prime symbol as in $A^{'}$. I_p denotes the identity matrix of order p. An estimate of a parameter ϑ is denoted by $\hat{\vartheta}$. $x \sim D$ means that a random variable x has the probability distribution D. The Gaussian distribution with mean μ and covariance Σ is denoted by $\mathcal{N}(\mu, \Sigma)$ while the Chi-squared distribution with n degrees of freedom is denoted by $\chi^2(n)$. *i.i.d.* means that a random variable is independent and identically distributed. *ln* denotes a natural logarithm. $Pr(X)$ is the probability of an event X.

2.2 Model

2.2.1 Large Dimensional Factor Model

In this chapter, we study the following standard factor model as described in Bai and Ng (2002). Let x_{it} be the real-valued observed data for the i-th cross-section unit at time t, for $i = 1, \ldots, p$, and $t = 1, \ldots, n$. Note that we denote the cross-sectional and temporal dimensions of the data by p and n, respectively, instead of N and T, to be consistent with the literature on random matrix theory. Consider the factor representation of the data of the form

$$x_{it} = \lambda_i' f_t + e_{it}, \tag{2.2.1}$$

where f_t is an $r \times 1$ vector of the factors, λ_i is an $r \times 1$ vector of factor loadings, and r is the *true* number of factors. $\lambda_i' f_t$ is the common component and e_{it} is the idiosyncratic error. Factors, factor loadings and the idiosyncratic components are not observable. Moreover, the true number of factors is unknown beforehand.

In vector notation, (2.2.1) can be written as a p-dimension time series with n observations:

$$\underset{(p\times 1)}{x_t} = \underset{(p\times r)}{\Lambda}\ \underset{(r\times 1)}{f_t} + \underset{(p\times 1)}{e_t}, \tag{2.2.2}$$

where $x_t = (x_{it}, \ldots, x_{pt})'$ is a $p \times 1$ vector of real-valued cross-section observations at time t, $\Lambda = (\lambda_1, \ldots, \lambda_p)'$ is a $p \times r$ factor loading matrix composed of r linearly independent vectors, and $e_t = (e_{it}, \ldots, e_{pt})'$ is a p-dimensional real-valued vector.

In matrix notation, the model is given by

$$\underset{(p\times n)}{X} = \underset{(p\times r)}{\Lambda}\ \underset{(r\times n)}{F'} + \underset{(p\times n)}{e}, \tag{2.2.3}$$

where $X = (x_1, \ldots, x_n)$, $F = (f_1, \ldots, f_n)'$, and $e = (e_1, \ldots, e_n)$.

Assumptions First, suppose that f_t is the zero mean random vector and independent of e_t. Both f_t and λ_i have positive definite covariance matrices Σ_F and Σ_Λ, respectively, so that each is of full rank, r. These assumptions imply that each factor has a nontrivial contribution to variance of x_t as in Bai and Ng (2002).

Next, for technical reasons, we assume that the errors e_{it} are independently and identically normally distributed, where σ is the unknown noise variance. Throughout this chapter, we assume $\sigma = 1$ without loss of generality since the overestimation probability bound is eventually given by the ratio of eigenvalues and consequently σ terms are cancelled out in this ratio.

The assumption of the i.i.d. errors enables us to employ some results from random matrix theory in order to derive the overestimation probability bound of the IC estimator. Random matrix theory studies the limiting behaviors of the eigenvalues and eigenvectors of the sample covariance matrix in a large dimensional framework. Especially, of all theoretical results from random matrix theory, a result on the non-asymptotic exponential bound of the largest eigenvalue is necessary for our study; however, it has been established only for Gaussian i.i.d. errors (see Section 2.5). To the best of our knowledge, such a result is not currently available for the more general covariance structure of the idiosyncratic terms. This chapter is not the first to assume i.i.d. errors in the literature on large dimensional factor models. For example, by using random matrix theory under the assumption of Gaussian i.i.d. errors, Onatski (2007) studied on the estimation of large factor models with weak factors, and Moon and Weidner (2015) analyzed large dimensional panels with unknown number of factors as interactive fixed effects. Moreover, this assumption is not too restrictive since it is sufficient enough to capture the main idea of large factor models. In the presence of strong factors, all important variations in the data should be captured by factors; hence, empirical studies on large factor models with strong factors do not typically specify the complicated correlation structure of the idiosyncratic terms (Harding, 2013). As a consequence, i.i.d. errors, along with strong factors, enable us to focus on the over-detection risk rather than the under-

detection risk of the IC estimator. While some of the techniques employed in our over-detection analysis are likely to be used to analyze the underestimation probability of IC as well, the under-detection risk is beyond the scope of this chapter.

In contrast, Bai and Ng (2002) allow for weak serial and cross-sectional dependence in the idiosyncratic components. In this regard, we examine the possibility that our theoretical result based on random matrix theory is extended to the case with the more general covariance structure of the error terms. First, we sketch the idea of how to formulate the overestimation probability bound for the case with non-i.i.d. errors; however, we leave a rigorous solution for future research while describing nontrivial difficulties (see Section 2.6). Next, we explore the finite sample performance of our modified criteria in the presence of weak correlation in the error terms through a Monte Carlo simulation study. The results show that the modified criteria lead to better performance even for the case with weakly serially or/and cross-sectionally correlated errors (see Section 2.7).

Third, for discussions related to random matrix theory, we consider the joint limit asymptotics where both n and p approach infinity simultaneously subject to $\frac{p}{n} \to c$, for $c \in [0, \infty)$. It is standard in the literature on large dimensional random matrices. By this assumption, sample eigenvalues corresponding to the error components remain bounded. Even though we assume the population eigenvalues of the error components to be bounded, their sample eigenvalues will diverge to infinity when p increases faster than n (Onatski, 2005).

Lastly, the true number of factors r is fixed regardless of n and p. The fixed r is generally assumed in the literature on the detection of the number of factors (e.g., Bai and Ng, 2002; Onatski, 2010, 2012; Ahn and Horenstein, 2013; Choi and Jeong, 2013; Harding, 2013).

2.2.2 Spiked Population Covariance Model

This subsection delineates the model structure by using the eigenvalue decomposition.

Let us decompose p eigenvalues of the population covariance matrix of x_t into two parts: (i) one coming from the systemic component and (ii) the other coming from the error terms.

Under the assumptions mentioned above, the population covariance matrix can be written as $\Sigma = \Psi + \Omega$, where Ψ is the covariance matrix of the common component and Ω is the error covariance matrix. Let $\{\psi_j\}_{j=1}^{r}$ denote r eigenvalues of Ψ which have non-zero finite values for all j with a decreasing order, that is, $\psi_1 \geq \psi_2 \geq \ldots \geq \psi_r > 0$. Besides, p eigenvalues of Ω are each equal to one since $\sigma = 1$. Then, p population eigenvalues of Σ are

$$(\psi_1 + 1,\ \psi_2 + 1,\ \ldots,\ \psi_r + 1,\ 1,\ 1, \ldots,\ 1). \tag{2.2.4}$$

Similarly, in the unknown basis B of $\mathbb{R}^p$, the population covariance matrix Σ takes a diagonal form

$$B'\Sigma B = diag\ (\psi_1, \ldots, \psi_r, 0, \ldots, 0) + I_p, \tag{2.2.5}$$

where B is a p-dimensional orthogonal matrix, that is, a $p \times p$ matrix composed of p eigenvectors corresponding to the eigenvalues of the population covariance matrix, Σ. The literature on random matrix theory refers to a covariance structure like (2.2.5) as a spiked population covariance model (Johnstone, 2001; Baik and Silverstein, 2006).

Note that while each factor has a nontrivial contribution to the data, the idiosyncratic term is an irrelevant disturbance so that it does not affect the data systematically. In this sense, f_t and e_t can be referred to as signals and noise, respectively, as in the literature on signal processing. Throughout this chapter, these insightful terms – signals and noise – are more often used than factors and errors. Thus, the eigenvalues of Ψ can be called *noise-free* population signal eigenvalues because Ψ is of rank r, while the eigenvalues of Ω are considered as *pure noise* eigenvalues.

Now, let S_n denote the sample covariance matrix of the n observations x_t from the model (2.2.2),

$$S_n = \frac{1}{n} \sum_{t=1}^{n} x_t x_t', \tag{2.2.6}$$

which is a $p \times p$ matrix with n samples of p-dimensional mean zero vectors, and let $\{\ell_j\}_{j=1}^{p}$ denote its eigenvalues, which are decreasingly ordered, $\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p$. For later use, we

also define a tail statistic by the ratio of the $(r+1)$th largest eigenvalue of S_n to the average of its last $p-r$ eigenvalues:

$$U_{p-r} = \frac{\ell_{r+1}}{\frac{T_{p-r}}{p-r}}, \tag{2.2.7}$$

where T_{p-r} is the sum of the last $p-r$ eigenvalues of S_n (i.e., $T_{p-r} = \sum_{j=r+1}^{p} \ell_j$). Especially when $r = 0$, the denominator equals the average trace of S_n (i.e., $\frac{1}{p}T_p = \frac{1}{p}tr(S_n)$). Note that U_{p-r} does not depend on the unknown noise variance, σ. Hence, as aforementioned, we assume $\sigma = 1$ without loss of generality.

2.3 Detection of the Number of Factors

2.3.1 *IC* estimator

Bai and Ng (2002) set up the detection of the number of factors as a model selection problem. They proposed the panel information criteria (IC) as follows:

$$IC(k) = \ln S(k) + k \cdot G(p,n), \tag{2.3.1}$$

where k is an arbitrary number such that $k < \min\{p,n\}$, $G(p,n)$ denotes the penalty function for overfitting, and $S(k)$ is the sum of squared residuals divided by pn such that

$$S\left(k\right) = \frac{1}{pn}\sum_{i=1}^{p}\sum_{t=1}^{n}(x_{it} - \widetilde{\lambda}_i^{'k}\widetilde{f}_t^{k})^2. \tag{2.3.2}$$

$\widetilde{f}_t^{k}$ and $\widetilde{\lambda}_i^{'k}$ denote estimated factors and loadings by the principal components method given the number of factors k, respectively. Then, the estimator for the true number of factors (IC estimator) is obtained by minimizing (2.3.1), namely that

$$\hat{k}_{IC} = \underset{0\leq k\leq kmax}{\arg\min} IC(k),$$

where $kmax$ is a bounded integer which is a maximum possible number of factors prespecified by users such that $r \leq kmax$. The IC estimator was proven to be consistent, namely that

$$\lim_{n,p\to\infty} \Pr(\hat{k}_{IC} = r) = 1,$$

if (1) $G(p,n) \to 0$ and (2) $C_{pn}^2 G(p,n) \to \infty$ as $n, p \to \infty$, where $C_{pn} = \min\{\sqrt{p}, \sqrt{n}\}$. That is, in the joint limit $n, p \to \infty$, the probability limit with which this model selection criterion selects the true number of factors converges to one if the penalty factor asymptotically converges to zero at an appropriate rate. Also, Bai and Ng propose specific formulations of the penalty factor to be used in practice: $G_1(p,n) = \left(\frac{p+n}{pn}\right) \ln\left(\frac{pn}{p+n}\right)$, $G_2(p,n) = \left(\frac{p+n}{pn}\right) \ln C_{pn}^2$, and $G_3(p,n) = \frac{\ln C_{pn}^2}{C_{pn}^2}$. Finally, they consider the following three criteria associated with three penalty terms:

$$IC_1(k) = \ln S(k) + k \cdot G_1(p,n) = \ln S(k) + k \cdot \left(\frac{p+n}{pn}\right) \ln\left(\frac{pn}{p+n}\right); \tag{2.3.3}$$

$$IC_2(k) = \ln S(k) + k \cdot G_2(p,n) = \ln S(k) + k \cdot \left(\frac{p+n}{pn}\right) \ln C_{pn}^2; \tag{2.3.4}$$

$$IC_3(k) = \ln S(k) + k \cdot G_3(p,n) = \ln S(k) + k \cdot \frac{\ln C_{pn}^2}{C_{pn}^2}. \tag{2.3.5}$$

Eigenvalue representation In this chapter, we work with random matrix theory to derive the upper bound on the overestimation probability of IC. To do so, the first step is to represent IC in terms of eigenvalues. If A is a square $p \times p$ matrix, then the trace of A is the same as the sum of the eigenvalues of A. Using this fact, IC (2.3.1) can be rewritten as follows:

Lemma 2.1. *Let $\{\ell_j\}_{j=1}^p$ denote p eigenvalues of a sample covariance matrix of the n observations x_t defined in (2.2.6), which are decreasingly ordered, $\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p$. Then, the panel information criteria (2.3.1) as proposed in Bai and Ng (2002) can be written as*

$$IC(k) = \ln\left(\frac{1}{p}\sum_{j=k+1}^{p} \ell_j\right) + k \cdot G(p,n), \tag{2.3.6}$$

where k is an arbitrary number such that $k < \min\{p, n\}$, and $G(p, n)$ is the penalty function for overfitting.

As a result, IC is written in terms of only the last $(p - k)$ sample eigenvalues without the first k sample eigenvalues.

2.3.2 Overestimation of the IC estimator

In what follows, we specify a mathematical condition for the overestimation of IC and its over-detection probability in terms of only the last $(p - r)$ sample eigenvalues based on Lemma 2.1. This chapter particularly focuses on the situation when IC overestimates the true number of factors by *exactly* one factor rather than multiple factors. Here we give a brief discussion on this approach. First, the theoretical part of this chapter assumes that the explanatory power of signals is strong and errors are i.i.d; therefore, we focus on the analysis of over-detection performance rather than under-detection performance. Next, the population eigenvalues are assumed to be decreasingly ordered. Under the same assumption, various studies based on random matrix theory investigate the ratio or difference of two *adjacent* sample eigenvalues to propose a consistent estimator for the true number of factors, see Onatski (2010), and Ahn and Horenstein (2013). It implies that a difference in the explanatory power of two adjoining factors governs the detection performance of the estimator. We also consider various works which studied the signal detection performance of the classical information criteria such as the Akaike information criterion (e.g., Zhang, Wong, and Reilly, 1989; Nadler, 2010). It was shown that overestimation by exactly one signal dominates the misdetection risk of the information criteria.

For conceptual simplicity, suppose that the criterion (2.3.6) is minimized at $r+1$, where r is the true number of factors. Then, since the IC estimator, $\hat{k}_{IC}$, is defined as the minimizer of $IC(k)$ over a range of values for k, the IC estimator overdetects the true number of factors by *exactly* one factor, namely that $\hat{k}_{IC} = r + 1$. We hence specify a condition for

overestimation by one factor:

$$\triangle IC(1) = IC(r) - IC(r+1) > 0, \tag{2.3.7}$$

where $IC(r) = \ln\left(\frac{1}{p}\sum_{j=r+1}^{p}\ell_j\right) + r \cdot G(p,n)$ and $IC(r+1) = \ln\left(\frac{1}{p}\sum_{j=r+2}^{p}\ell_j\right) + (r+1) \cdot G(p,n)$. Correspondingly, the overestimation probability of IC is specified as follows:

Lemma 2.2. *Suppose that IC (2.3.6) is minimized at $r+1$, where r is the true number of factors. Let $\{\ell_j\}_{j=1}^{p}$ denote the eigenvalues of a sample covariance matrix, S_n, of the n observations x_t defined in (2.2.6), which are decreasingly ordered, $\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p$. Also, we denote by T_{p-r} the sum of the last $p-r$ eigenvalues of S_n. Then, the IC estimator overestimates the true number of factors by exactly one factor if $\triangle IC(1) > 0$ with $\triangle IC(1)$ given by (2.3.7). Thus, the probability with which the number of factors would be overestimated by exactly one factor takes the form*

$$\Pr\left(\triangle IC(1) > 0\right) = \Pr\left(\ln\frac{T_{p-r}}{T_{p-r-1}} - G(p,n) > 0\right), \tag{2.3.8}$$

where $T_{p-r} = \sum_{j=r+1}^{p}\ell_j$, $T_{p-r-1} = \sum_{j=r+2}^{p}\ell_j$, and $G(p,n)$ is the penalty function of IC.

To apply random matrix theory to our analysis, the next step is to express a condition (2.3.7) for the overestimation of IC and its overestimation probability (2.3.8) in terms of *pure noise* sample eigenvalues. Eventually, they will be represented by a tail statistic (2.2.7) which is a function of pure noise eigenvalues. Before moving on, we can show that (2.3.8) is easily approximated by a tail statistic using the log inequality, $\log(1-x) \leq -x$ for $x \in [0,1)$. That is,

$$\Pr\left(\frac{\ell_{r+1}}{T_{p-r}} > G(p,n)\right) \tag{2.3.9}$$

since $\ln\left(\frac{T_{p-r}}{T_{p-r-1}}\right) = -\ln\left(\frac{T_{p-r-1}}{T_{p-r}}\right) = -\ln\left(1 - \frac{\ell_{r+1}}{T_{p-r}}\right) \geq \frac{\ell_{r+1}}{T_{p-r}}$. Both (2.3.8) and (2.3.9) imply that the overestimation probability is defined in terms of only the last $p-r$ eigenvalues of the sample covariance matrix; that is, it is not a function of the first r eigenvalues of S_n.

This implication is essential for this chapter because the probability limit of (2.3.9) can be analyzed by using results from random matrix theory regarding the limiting behaviors of eigenvalues coming from pure noise components. It should be noted, however, that ℓ_{r+1} and T_{p-r} are not truly coming from pure noise. Since the space spanned by the signal–plus–noise subspace eigenvectors contains both signals and noise, ℓ_{r+1} contains not only contributions of noise but also those of signals and the interactions between signals and noise (for details, see Nadler, 2008, Theorem 2.1, p. 2802). Thus, the above argument (2.3.9) is given only for illustrative purposes, but it is not good enough for our analysis based on random matrix theory, regardless of how good the approximation is.

In the next section, we derive a more suitable expression for the overestimation probability to employ random matrix theory. It can be written in terms of the pure noise eigenvalues by constructing a Wishart matrix whose entries are Gaussian i.i.d. noise.

2.4 Overestimation Probability

Following Nadler (2008, 2010), this section shows that the overestimation probability (2.3.8) can be asymptotically specified by $p-r$ pure noise eigenvalues which are independent of r signal eigenvalues. Theoretically, $p-r$ pure noise eigenvalues can be identified as the eigenvalues of a $p-r$ dimensional Wishart matrix with identity covariance matrix. Here we first define related terms and introduce preliminary results.

Definition 2.1. ***Wishart matrix*** *(Silverstein, 1985; Johnstone, 2001): Let A denote a $p \times n$ matrix whose A_t are i.i.d. $\mathcal{N}(0, \Sigma_A)$ random vectors, and let $H = \frac{1}{n}AA'$. Then, the random matrix H is commonly referred to as a Wishart matrix, and $nH = AA'$ is said to have the Wishart distribution, $W_p(n, \Sigma_A)$. For the null case in which $\Sigma_A = I_p$, H is especially referred to as a Wishart matrix with identity covariance matrix.*

Furthermore, one can obtain the following result based on the standard distribution theory, which states that the squared norm of n standard normally distributed variables has

the Chi-squared distribution with n degrees of freedom.

Remark 2.1. *(Rao, 1973, p. 534) Under Definition 2.1, let* $nH \sim W_p(n, \Sigma_A)$. *Let* Y *be any* $p \times 1$ *fixed vector such that* $Y'A_t \sim \mathcal{N}(0, \sigma^2)$, *where* $\sigma^2 = Y'\Sigma_A Y$. *Then,* $nY'HY \sim \sigma^2 \cdot \chi^2(n)$.

Remark 2.1 can be extended to the following result:

Remark 2.2. *Suppose* $nH \sim W_p(n, \Sigma_A)$. *Let* a_j *denote the* j*-th eigenvalue of* H, *and let* Y *denote a* $p \times 1$ *eigenvector corresponding to* a_j *such that* $Y'A_t \sim \mathcal{N}(0, 1)$. *Then, by Remark 2.1,* $a_j \sim \chi^2(n)/n$ *and* $\sum_{j=1}^{p} a_j \sim \chi^2(np)/n$. *Also,* $E(a_j) = 1$, $Var(a_j) = 2/n$, $E(\sum_{j=1}^{p} a_j) = p$, *and* $Var(\sum_{j=1}^{p} a_j) = 2p/n$ *so that* $a_j = 1 + O_p(\sqrt{1/n})$ *and* $\sum_{j=1}^{p} a_j = p + O_p(\sqrt{p/n})$.

As seen before, $B'\Sigma B = \text{diag}(\psi_1 + 1, \ldots, \psi_r + 1, 1, \ldots, 1)$, where $B = (b_1, \ldots, b_p)$ is an orthogonal matrix which diagonalizes the population covariance matrix, Σ. For $j = 1, \ldots, p$, each column b_j is the eigenvector corresponding to the j-th population eigenvalue of Σ. Now, let us consider a new p-dimensional matrix $\widetilde{B} = (b_1, \ldots, b_r, \tilde{d}_{r+1}, \ldots, \tilde{d}_p)$ whose vectors are linearly independent. As before, the first r column vectors, $\{b_i\}_{i=1}^{r}$, are the r eigenvectors corresponding to the first r population eigenvalues, $\{\psi_i + 1\}_{i=1}^{r}$. On the other hand, the last $p - r$ column vectors, $\{\tilde{d}_j\}_{j=r+1}^{p}$, diagonalize the lower right sub-matrix of $\widetilde{B}' S_n \widetilde{B}$. Then, in the basis $\widetilde{B}$, S_n has the following form:

$$\widetilde{B}' S_n \widetilde{B} = \left[\begin{array}{ccc|ccc} \rho_{11} & \cdots & \rho_{1r} & & & \\ \vdots & \ddots & \vdots & & L' & \\ \rho_{r1} & \cdots & \rho_{rr} & & & \\ \hline & & & \tilde{\ell}_{r+1} & & \emptyset \\ & L & & & \ddots & \\ & & & \emptyset & & \tilde{\ell}_p \end{array}\right]. \tag{2.4.1}$$

In matrix (2.4.1), $\{\rho_{ii}\}_{i=1}^{r}$ are sample variances in the directions b_i corresponding to the first r population eigenvalues, that is, $\rho_{ii} = b_i' \left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right) b_i$ such that $\rho_{ii} \sim \left(\frac{\psi_i + 1}{n}\right) \chi^2(n)$. Next, $\{\widetilde{\ell}_j\}_{j=r+1}^{p}$ are the $p - r$ diagonal elements of a lower right sub-matrix in (2.4.1), that is, $\widetilde{\ell}_j = \tilde{d}_j' \left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right) \tilde{d}_j)$. In the basis $\widetilde{B}$, this lower right sub-matrix is given by the

projection of S_n onto the only noise subspace, which is independent of the projection of S_n onto the signal subspace; therefore, it does not contain any signal contributions. Accordingly, this $p-r$ dimensional sub-matrix is considered as the random realization of a Wishart matrix with identity covariance matrix, and its diagonal elements are considered as the sample eigenvalues of this Wishart matrix; that is, pure noise eigenvalues. Thus, $\widetilde{\ell}_j \sim \chi^2(n)/n$ by Remark 2.2. Meanwhile, another sub-matrix L contains the interaction terms between signals and noise. If we denote by η_{ij} each element of L, then $\eta_{ij} = \tilde{d}_j' \left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right) b_i$ for $i = 1, \ldots, r$ and $j = r+1, \ldots, p$.

So far, we have identified pure noise eigenvalues, $\{\widetilde{\ell}_j\}_{j=r+1}^{p}$. Now, we rewrite (2.3.8) in terms of $\widetilde{\ell}_j$. O'leary and Stewart (1990) refer to matrices such as (2.4.1) as arrow-head matrices; especially, they consider such matrices with one element of ρ in the upper left sub-matrix, that is, the case with $r = 1$. They derived the explicit formula for computing the eigenvalues and eigenvectors of symmetric arrow-head matrices, which is a function of ρ, η and $\widetilde{\ell}$ (O'leary and Stewart, 1990, Theorem 2.1; Nadler, 2008, p.2807). Also, Nadler (2010) extended their results to the case with $r > 1$. We obtain an approximate expansion for ℓ_j by employing results from the literature mentioned above.

Lemma 2.3. *Consider the model (2.2.2). Let $\{\psi_i\}_{i=1}^{r}$ denote the first r eigenvalues of the p-by-p population covariance matrix such that $\psi_1 \geq \psi_2 \geq \ldots \geq \psi_r > 0$, and $\psi_i = O(1)$. Let $\{\ell_j\}_{j=r+1}^{p}$ denote the last $p-r$ eigenvalues of a sample covariance matrix, S_n, of the n observations x_t defined in (2.2.6), which are decreasingly ordered, $\ell_{r+1} \geq \ell_{r+2} \geq \ldots \geq \ell_p$. Also, as described in matrix (2.4.1), ρ_{ii}, $\widetilde{\ell}_j$ and η_{ij} denote the i-th sample variance, the j-th sample eigenvalue of a Wishart matrix with identity covariance matrix, and an interaction term between signals and noise, respectively. Then, as $n \to \infty$, ℓ_j is represented in terms of*

ρ_{ii}, $\widetilde{\ell}_j$ *and* η_{ij} *as follows:*

$$\ell_j = \widetilde{\ell}_j - \frac{1}{n}\sum_{i=1}^{r} \frac{(\sqrt{n}\,\eta_{ij})^2}{\rho_{ii} - \widetilde{\ell}_j} + o_p\left(\frac{1}{n}\right) \tag{2.4.2}$$

$$= \widetilde{\ell}_j \left(1 - \frac{M_r}{n} - \frac{\sqrt{r}}{n} Z_j\right) + o_p\left(\frac{1}{n}\right), \tag{2.4.3}$$

where $M_r = \sum_{i=1}^{r} \frac{\psi_i+1}{\psi_i}$, $Z_j = \frac{1}{\sqrt{r}} \sum_{i=1}^{r} \frac{\psi_i+1}{\psi_i}(\kappa_{ij}^2 - 1)$, *and* $\kappa_{ij} = \frac{\sqrt{n}\,\eta_{ij}}{(\rho_{ii}\widetilde{\ell}_j)^{1/2}}$.

Indeed, the sum of the last $p - r$ *sample eigenvalues,* T_{p-r}, *is represented by*

$$T_{p-r} = \widetilde{T}_{p-r} \left(1 - \frac{M_r}{n} - \frac{\sqrt{r}}{n} \frac{1}{\widetilde{T}_{p-r}} \sum_{j=r+1}^{p} \widetilde{\ell}_j Z_j\right) + o_p\left(\frac{1}{n}\right), \tag{2.4.4}$$

where $\widetilde{T}_{p-r} = \sum_{j=r+1}^{p} \widetilde{\ell}_j$.

Lemma 2.3 shows that the j-th sample eigenvalue, ℓ_j, is approximately the same as the product of the j-th pure noise eigenvalue, $\widetilde{\ell}_j$, and additional terms which contain signal eigenvalues and the interaction terms. Now we obtain the first contribution of this chapter based on this result. Our asymptotic expression for the overestimation probability of IC is explicitly identified by only pure noise eigenvalues so that it is asymptotically independent of signal eigenvalues.

Theorem 2.1. *Let* W *be a* $(p-r)\times(p-r)$ *Wishart matrix with identity covariance matrix. The largest eigenvalue of* W *is denoted by* $\ell_1(W)$, *and the sum of* $p - r$ *eigenvalues of* W *is denoted by* $Tr(W)$. *Assuming that IC (2.3.6) is minimized at* $r + 1$, *where* r *is the true number of factors, the IC estimator overestimates the true number of factors by exactly one factor. Then, under the conditions of the Lemma 2.3, asymptotically as* $n \to \infty$, *the overestimation probability of IC in the presence of* r *factors is given by*

$$\Pr(\triangle IC(1) > 0) = \Pr\left(\frac{\ell_1(W)}{Tr(W)} - \xi_{n,p} > 0\right) + O_p\left(\frac{1}{n}\right), \tag{2.4.5}$$

where $\xi_{n,p} = -1 + \sqrt{1 + 2G(p,n)}$, *and* $G(p,n)$ *is the penalty function of IC.*

Note that since a $p-r$ dimensional lower right sub-matrix of (2.4.1) is considered as the random realization of W, the largest eigenvalue of W, $\ell_1(W)$, is equivalent to the first pure noise eigenvalue, $\widetilde{\ell}_{r+1}$. Also, $Tr(W)$ is equivalent to the sum of pure noise eigenvalues, $\widetilde{T}_{p-r}$.

Hitherto, we derived the asymptotic expression for the overestimation probability of IC in terms of a tail statistic with only pure noise eigenvalues independent of the signal eigenvalues. The following sections explore the second contribution of this chapter – namely, determining a non-asymptotic upper bound on the over-detection probability in finite samples. This analysis is highly related to random matrix theory since the overestimation probability (2.4.5) can be pinned down by using the limiting distribution of the largest eigenvalue of a Wishart matrix with identity covariance matrix.

2.5 Mathematical Preliminaries

The main tools used in our analysis are recent results from random matrix theory regarding the largest eigenvalue of a pure noise matrix. In this section, we review the idea and relevant results of random matrix theory. In a concise manner, random matrix theory is sort of special limiting laws to deal with high dimensional statistics. It is well known that classical limit theorems for a fixed dimension (large n with fixed p) are not sufficient enough to analyze large dimensional panels (large n and large p); specifically, the sample covariance matrix is no longer a good approximation to the population covariance matrix when the population size is large and comparable with the sample size (for details, see Baik and Silverstein, 2006; Bai and Silverstein, 2010). In addition, as Anderson (2003) showed, as $n \to \infty$ with fixed p, the largest eigenvalue of the sample covariance matrix is consistent for the largest eigenvalue of the population covariance matrix; however, it is no longer true in large dimensions (Geman, 1980; Johnstone, 2001). For this reason, new theorems are required to study a random covariance matrix and corresponding eigenvalues in a large dimensional framework; as a response, random matrix theory provides such new limiting laws.

Random matrix theory typically digs into the following topics: (i) the joint distribution of all eigenvalues of a Wishart matrix; (ii) the distribution of its extreme eigenvalues, especially the largest one and the smallest one; and more recently, (iii) a non-asymptotic bound on the largest eigenvalue of a Wishart matrix for finite values of p and n. Now, we summarize the main results of random matrix theory. By definition 2.1 and Remark 2.2, let $H = AA'/n$ denote a $p \times p$ Wishart matrix with identity covariance matrix, where A is a $p \times n$ matrix with real valued Gaussian i.i.d. entries, and let a_j denote the j-th sample eigenvalue with a decreasing order, for $j = 1, \ldots, p$.

First, Geman (1980) showed that in the joint limit $n, p \to \infty$, with $\frac{p}{n} \to c \leqslant 1$, the empirical distribution of eigenvalues given by $F_p(h)$ converges to a non-random distribution function $F(h)$, which has the support of $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$ with a probability one. Then, the largest eigenvalue of H converges to the upper bound on the support of the limiting distribution with a probability one. That is, for any real h,

$$F_p(h) = \frac{1}{p}\{\text{number of } \ a_j \leq h\} \xrightarrow{a.s} F(h),$$

and a density is given by $f(h) = \frac{1}{2\pi hc}\sqrt{(\beta - h)(h - \alpha)}$ for $\alpha \leq h \leq \beta$, where $\alpha = (1-\sqrt{c})^2$ and $\beta = (1+\sqrt{c})^2$. Then,

$$a_1 \xrightarrow{a.s} (1+\sqrt{c})^2. \tag{2.5.1}$$

Johnstone (2001) derived the limiting distribution of the largest eigenvalue of a real-valued Wishart matrix with identity covariance matrix. Specifically, call

$$\begin{aligned} n_1 &= \max\{n, p\} - 1, \quad p_1 = \min\{n, p\}, \\ \mu^o_{n,p} &= \frac{1}{n}(\sqrt{n_1} + \sqrt{p_1})^2, \\ \sigma^o_{n,p} &= \frac{1}{n}(\sqrt{n_1} + \sqrt{p_1})\left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{p_1}}\right)^{1/3}, \end{aligned}$$

and TW_β is the Tracy-Widom distribution of order β, it was shown that in the joint limit

$n, p \to \infty$ with $\frac{p}{n} \to c \in (0, \infty)$, the distribution of the largest eigenvalue of H converges to a Tracy-Widom distribution

$$\frac{a_1 - \mu^o_{n,p}}{\sigma^o_{n,p}} \xrightarrow{d} TW_1, \tag{2.5.2}$$

where TW_1 is the Tracy-Widom distribution of order 1 corresponding to real-valued observations. Also, for any real h, it can be written as

$$\Pr\left(\frac{a_1 - \mu_{n,p}}{\sigma_{n,p}} \leq h\right) \to TW_1(h), \tag{2.5.3}$$

where $TW_1(h)$ is the Tracy-Widom CDF which is defined in terms of the Airy function (for details, see Tracy and Widom, 1996; Johnstone, 2001). The above result is applied for both situations in which $n \geq p$ as well as $n < p$.

Karoui (2008) generalized results in Johnstone (2001) to the following: (i) with the same centering and scaling, (2.5.2) still holds when $\frac{p}{n}$ or $\frac{n}{p} \to 0$; (ii) further, (2.5.2) holds for the τ largest eigenvalues, where τ is a fixed integer such that $\tau > 1$; and (iii) the Tracy-Widom approximation is reasonable even when one of the dimensions is small. Although the generic rate of convergence of the left side of (2.5.3) to $TW_1(h)$ is $O(\min\{n, p\}^{-1/3})$, small modifications in a centering parameter $\mu^o_{n,p}$ and a scaling parameter $\sigma^o_{n,p}$ lead to $O(\min\{n, p\}^{-2/3})$ errors. Along the line of Karoui (2008), Ma (2012) particularly suggested that in the joint limit $n, p \to \infty$ with $\frac{p}{n} \to c \in [0, \infty]$,

$$\left|\Pr\left(\frac{a_1 - \mu_{n,p}}{\sigma_{n,p}} \leq h\right) - TW_1(h)\right| = O(\min\{n, p\}^{-2/3}), \tag{2.5.4}$$

with modified centering and scaling parameters:

$$\mu_{n,p} = \frac{1}{n}\left(\sqrt{n-\frac{1}{2}} + \sqrt{p-\frac{1}{2}}\right)^2;$$
$$\sigma_{n,p} = \frac{1}{n}\left(\sqrt{n-\frac{1}{2}} + \sqrt{p-\frac{1}{2}}\right)\left(\frac{1}{\sqrt{n-\frac{1}{2}}} + \frac{1}{\sqrt{p-\frac{1}{2}}}\right)^{1/3}.$$

Recently, Nadler (2011) applied the above results to a tail statistic. Let U denote the ratio of the largest sample eigenvalue of H to the average of its p eigenvalues (i.e., $U = p \cdot a_1/T_p$). Then, in the joint limit $n, p \to \infty$ with $\frac{p}{n} \geq 0$, the distribution of U also converges to the TW distribution:

$$\frac{U - \mu_{n,p}}{\sigma_{n,p}} \xrightarrow{d} TW_1. \tag{2.5.5}$$

The convergence rate to the TW distribution is also known as $O(\min\{n,p\}^{-2/3})$. Intuitively, the asymptotic property of U is equivalent to a_1 in the sense that the denominator of U has a negligible remainder with respect to that of a_1 because $a_1 = 1 + O(1/\sqrt{n})$ and $T_p/p = 1 + O(1/\sqrt{np})$. Building on this result, we can show that in the joint limit $n, p \to \infty$ with $\frac{p}{n} \geq 0$, the overestimation probability of IC given by Theorem 2.1 is also approximated by the TW distribution. Especially, for the case with no signal,

$$\Pr(\triangle IC(1) > 0) = \Pr\left(\frac{\ell_1(W)}{Tr(W)} > \xi_{n,p}\right) \longrightarrow 1 - TW_1(h), \tag{2.5.6}$$

where $\ell_1(W)/Tr(W) = U_p/p$, and $h = (p \cdot \xi_{n,p} - \mu_{n,p})/\sigma_{n,p}$.

In this chapter, however, we analyze the detection performance of IC by providing an explicit non-asymptotic bound on the overestimation probability rather than the above approximate analysis. Our analysis relies strongly on the results in Ledoux (2007). Ledoux provided the following non-asymptotic bound on the largest eigenvalue of a Wishart matrix with identity covariance matrix. For some constant $M > 0$, $\varepsilon > 0$, and $n \geq 1$,

$$\Pr\left(a_1 \geq (1+\sqrt{\bar{c}})^2 + \varepsilon\right) \leq M \exp\left(-n \min\{\varepsilon, \varepsilon^{3/2}\}/M\right), \tag{2.5.7}$$

where $\bar{c} = p/n$ for finite values n and p (Ledoux, 2007, Proposition 2.2). As an extension of (2.5.7), Kritchman and Nadler (2009) and Nadler (2010) showed that for all values of n and p,

$$\Pr\left(a_1 \geq (1+\sqrt{\bar{c}})^2 + \varepsilon\right) \leq \exp\left(-n J_{LAG}(\varepsilon)\right), \tag{2.5.8}$$

where

$$J_{LAG}(\varepsilon) = \int_1^x (x-y) \frac{(1+\bar{c})y + 2\sqrt{\bar{c}}}{(y+B)^2} \frac{dy}{\sqrt{y^2-1}}$$

with $\bar{c} = p/n$, $x = 1 + (\varepsilon/2\sqrt{\bar{c}})$, and $B = (1+\bar{c})/2\sqrt{\bar{c}}$.

Note that all the above results are stated for the case with no signal. Nonetheless, these results can be generalized to the case where r signals exist. In particular, the largest $(r+1)$th eigenvalue in our spiked covariance model defined in (2.2.5) asymptotically follows the TW distribution with parameters: n and $p-r$ (Baik and Silverstein, 2006; Paul, 2007; Karoui, 2008). (2.5.8) can be also applied to a spiked covariance model with r signals (Kritchman and Nadler, 2009); in this case, $\bar{c}$ is adjusted to $(p-r)/n$.

2.6 Non-asymptotic Bound on Overestimation Probability

2.6.1 Main Result

In this section, we finally derive a non-asymptotic bound on the overestimation probability of *IC* based on previous discussions. Specifically, by applying a result from random matrix theory (2.5.8) to our expression of the overestimation probability of *IC* (2.4.5), we provide the following theorem:

Theorem 2.2. *Consider the model (2.2.2) and the panel information criteria (IC) defined in (2.3.1). Suppose that the IC estimator overestimates the true number of factors by exactly one factor, namely that IC is minimized at $r+1$. Then, for finite values of n and p, a*

non-asymptotic upper bound on the overestimation probability of IC by exactly one factor is given by

$$\Pr(\triangle IC(1) > 0) \leq \exp\left(\frac{-(p-r)s^2}{4}\right) + \exp\left(-\frac{4n}{3}(\bar{c})^{1/4}\left((p-r)\left(1-\frac{s}{\sqrt{n}}\right)\xi_{n,p} - (1+\sqrt{\bar{c}})^2\right)^{3/2}\right). \tag{2.6.1}$$

This non-asymptotic bound is appropriate for any positive value of s chosen by a user such that

$$\sqrt{n} - \frac{1}{\xi_{n,p}\sqrt{p-r}}\left(3+\sqrt{\bar{c}}+\frac{1}{\sqrt{\bar{c}}}\right) < s < \sqrt{n} - \frac{1}{\xi_{n,p}\sqrt{p-r}}\left(2+\sqrt{\bar{c}}+\frac{1}{\sqrt{\bar{c}}}\right), \tag{2.6.2}$$

where $\bar{c} = \frac{p-r}{n}$ and $\xi_{n,p} = -1 + \sqrt{1+2G(p,n)}$. Also, (2.6.1) holds for all the formulations of the penalty function $G(p,n)$ which are specified in (2.3.3), (2.3.4), and (2.3.5).

Theorem 2.2 provides users with a simple diagnostic tool for the misspecification of the number of factors. It discloses numerically how maximally overestimation occurs so long as users know the temporal and cross-sectional size of the data. Recall that $\bar{c}$ and $\xi_{n,p}$ are functions of n and p. Also, the appropriate value of s depends on n and p. In practice, the user can choose the value of s such that it can minimize the upper bound defined in (2.6.1) as long as it satisfies (2.6.2).

Remarks on the case with non-i.i.d. errors As aforementioned, results on the deviation inequalities of the largest eigenvalue from random matrix theory, which are shown in (2.5.7) and (2.5.8), are currently only available for the case with Gaussian i.i.d. errors. Moreover, results from Nadler (2010), which are used to obtain Lemma 2.3 and Theorem 2.1, are also only feasible under the assumption of the Gaussian i.i.d. error components. In the presence of weak serial and cross-sectional dependence, a different approach is hence needed to analyze a bound on the over-detection probability; however, there are nontrivial hurdles. Although a rigorous solution is beyond the scope of this chapter, we instead sketch some

ideas and difficulties for future research.

1. *Consider a specific covariance structure as proposed in Ma (2003) and Stein (2005), a spatio-temporal covariance model:* $e = R_p^{1/2} U Q_n^{1/2}$. *$U$ is a $p \times n$ matrix whose entries are Gaussian i.i.d. Also, a $p \times p$ matrix R_p and an $n \times n$ matrix Q_n are positive definite matrices capturing cross-sectional and serial correlation in e, respectively. This model has been used in previous studies on signal detection (e.g., Onatski, 2010; Ahn and Horenstein, 2013; Harding, 2013)*

2. *Let $\psi_\tau(A)$ denote the τ-th eigenvalue of a matrix A with a decreasing order. For $i, j = 1, \ldots, p$ and $t, s = 1, \ldots, n$, if R_p and Q_n are symmetric toeplitz matrices with entries of $\rho_R^{|i-j|}$ and $\rho_Q^{|t-s|}$, respectively, then asymptotic bounds on their extreme eigenvalues are known in the literature (Grenander and Szegö, 1958, p.147–154; Gray, 2006, Lemma 4.1): as $n, p \to \infty$, $\psi_1(R_p) \to \frac{1+\rho_R}{1-\rho_R}$, $\psi_p(R_p) \to \frac{1-\rho_R}{1+\rho_R}$, $\psi_1(Q_n) \to \frac{1+\rho_Q}{1-\rho_Q}$, and $\psi_n(Q_n) \to \frac{1-\rho_Q}{1+\rho_Q}$.*

3. *Since Theorem 2.1 is no longer applicable, we instead consider (2.3.9),* $\Pr\left(\frac{\ell_{r+1}}{T_{p-r}} > G_{p,n}\right)$, *where the inequality is only a necessary condition for overestimation by exactly one factor. Some known results on eigenvalue inequalities may be used to derive a bound on the probability that this necessary condition holds (e.g., Anderson and Gupta, 1963, Corollary 2.2.1; Rao, 1963, p.64; Horn and Johnson, 1991, Theorem 3.3.16).*

By following the above steps, we could formulate an expression for the overestimation probability bound in terms of asymptotic bounds on the extreme eigenvalues of R_p, Q_n and U, when there is no signal. Note that, however, this bound may not be fine enough since the approximation error seems to be quite large. It could be attributed to (i) quite loose eigenvalue-inequalities used in our analysis or/and (ii) the fact that we derived a probability bound associated with only a necessary condition. Besides, this bound was only available for the case with no signal. Finally, more acceptable solutions are left for future work. Potential

improvements might be attained by using tighter eigenvalue-inequalities or by analyzing a more acceptable expression for the overestimation condition.

2.6.2 Detection Performance of the *IC* estimator

The finite sample performance of the IC estimator has been studied by Monte Carlo simulations in the literature. It was shown that IC tends to overdetect the true number of factors for the case with relatively small sample sizes. For example, the experiments of Bai and Ng (2002) showed that the over-detection risk is non negligible for the case with small sample sizes $(n, p) \in \{(10, 50), (10, 100), (20, 100), (100, 10), (100, 20)\}$ when factors are not sufficiently strong, and such overdetection occurs for both cases with weakly correlated errors and Gaussian i.i.d. errors. There are additional simulation studies, which obtained the same results, allowing the presence of weak factors and weak correlation in the error components (e.g., Ahn and Horenstein, 2013; Onatski, 2010). In the above simulation studies, however, the results for the case with strong factors and i.i.d. Gaussian errors were not reported.

Accordingly, in this subsection, we theoretically analyze the finite sample performance of the IC estimator for the case with strong factors and i.i.d. Gaussian errors. Using Theorem 2.2, we compute non-asymptotic upper bounds on the overestimation probability of the IC estimator corresponding to various sample sizes and estimated numbers of factors. In each case, an appropriate positive number s was chosen by minimizing an upper probability bound on the interval (2.6.2). Main results are presented in Table 2.1. Each cell displays an upper bound on the overestimation probability of the IC estimator corresponding to each value of n, p and $\hat{k}_{IC}$, and the choice of a penalty function. Following the experiments of Bai and Ng (2002), we consider small sample sizes such that $\max\{n, p\} \in \{50, 60, 75, 100, 200\}$ and $\min\{n, p\} \in \{10, 15, 20, 25, 50\}$. In a few cases, an upper bound was not available since there was no positive value of s which satisfied (2.6.2).

Table 2.1 shows that for quite a few cases with small sample sizes, the computed bounds on the overestimation probability of the IC estimator are not negligible, say over 50%. This

Table 2.1: Detection Performance of the IC estimator (I.I.D. Errors)

	$r=0$			$r=1$			$r=2$		
(n,p)	IC_1	IC_2	IC_3	IC_1	IC_2	IC_3	IC_1	IC_2	IC_3
(50,10)	1.0256	0.4403	n.a	1.8276	1.0309	n.a	n.a	1.9685	n.a
(50,15)	0.1139	0.0068	1.1898	0.3620	0.0376	n.a	0.8541	0.1615	n.a
(50,20)	0.0049	0.0000	1.0724	0.0210	0.0002	1.5879	0.0768	0.0014	n.a
(60,10)	0.6247	0.2070	1.1891	1.1934	0.8002	n.a	n.a	1.5334	n.a
(60,15)	0.0211	0.0011	0.6127	0.1082	0.0093	1.0661	0.4056	0.0605	1.8685
(60,20)	0.0003	0.0000	0.3195	0.0021	0.0000	0.7165	0.0124	0.0002	1.0746
(75,10)	0.1975	0.0555	0.6697	0.8808	0.4054	1.3870	1.8811	1.1264	n.a
(75,15)	0.0012	0.0000	0.0720	0.0128	0.0009	0.3259	0.0939	0.0114	0.9460
(75,20)	0.0000	0.0000	0.0115	0.0000	0.0000	0.0577	0.0006	0.0000	0.2243
(100,10)	0.0185	0.0045	0.0870	0.2529	0.0953	0.6738	1.0892	0.7881	1.7878
(100,15)	0.0000	0.0000	0.0007	0.0002	0.0000	0.0107	0.0049	0.0005	0.1061
(200,10)	0.0000	0.0000	0.0000	0.0002	0.0000	0.0008	0.0410	0.0185	0.1039
(10,50)	0.7059	0.1788	n.a	0.7705	0.2153	n.a	0.8330	0.2573	n.a
(15,50)	0.0318	0.0006	1.0039	0.0445	0.0010	1.0170	0.0616	0.0016	1.0616
(20,50)	0.0006	0.0000	0.9576	0.0011	0.0000	1.0074	0.0020	0.0000	1.0318
(10,60)	0.2728	0.0468	0.9433	0.3137	0.0581	0.9753	0.3588	0.0717	0.9978
(15,60)	0.0019	0.0000	0.3084	0.0029	0.0000	0.3675	0.0042	0.0000	0.4335
(20,60)	0.0000	0.0000	0.1403	0.0000	0.0000	0.1830	0.0000	0.0000	0.2355
(10,75)	0.0335	0.0038	0.2680	0.0405	0.0049	0.3013	0.0487	0.0063	0.3374
(15,75)	0.0000	0.0000	0.0083	0.0000	0.0000	0.0111	0.0000	0.0000	0.0147
(20,75)	0.0000	0.0000	0.0007	0.0000	0.0000	0.0010	0.0000	0.0000	0.0015
(10,100)	0.0003	0.0000	0.0051	0.0004	0.0000	0.0062	0.0005	0.0000	0.0074
(15,100)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
(10,200)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Note: This table reports an upper bound on the overestimation probability of the IC estimator, $\Pr(\triangle IC(1) > 0)$ defined in Theorem 2.1, subject to the true number of factors $r \in \{0, 1, 2\}$ and the choice of panel information criteria. Upper bounds are computed by the formula (2.6.1) depending on various sample sizes (n, p). We consider sample sizes (n, p) such that $\max\{n, p\} \in \{50, 60, 75, 100, 200\}$ and $\min\{n, p\} \in \{10, 15, 20, 25, 50\}$. Three different panel information criteria, $IC_1(k)$, $IC_2(k)$ and $IC_3(k)$, are defined in (2.3.3), (2.3.4) and (2.3.5), respectively. If a probability bound is less than 1.0×10^{-4}, we simply put a zero. In some cases, we report an upper bound which is larger than one because it helps compare the magnitude of over-detection risks. *"n.a"* ("Not Applicable") indicates that an appropriate positive number of s which satisfies (2.6.2) is not available in this case.

result says that even when the explanatory power of factors are strong and the error components are i.i.d, the over-detection risk is not negligible for the case with small samples. Hence, it provides additional evidence of the overdetection of IC for finite samples. In addi-

Table 2.1 (Continued)

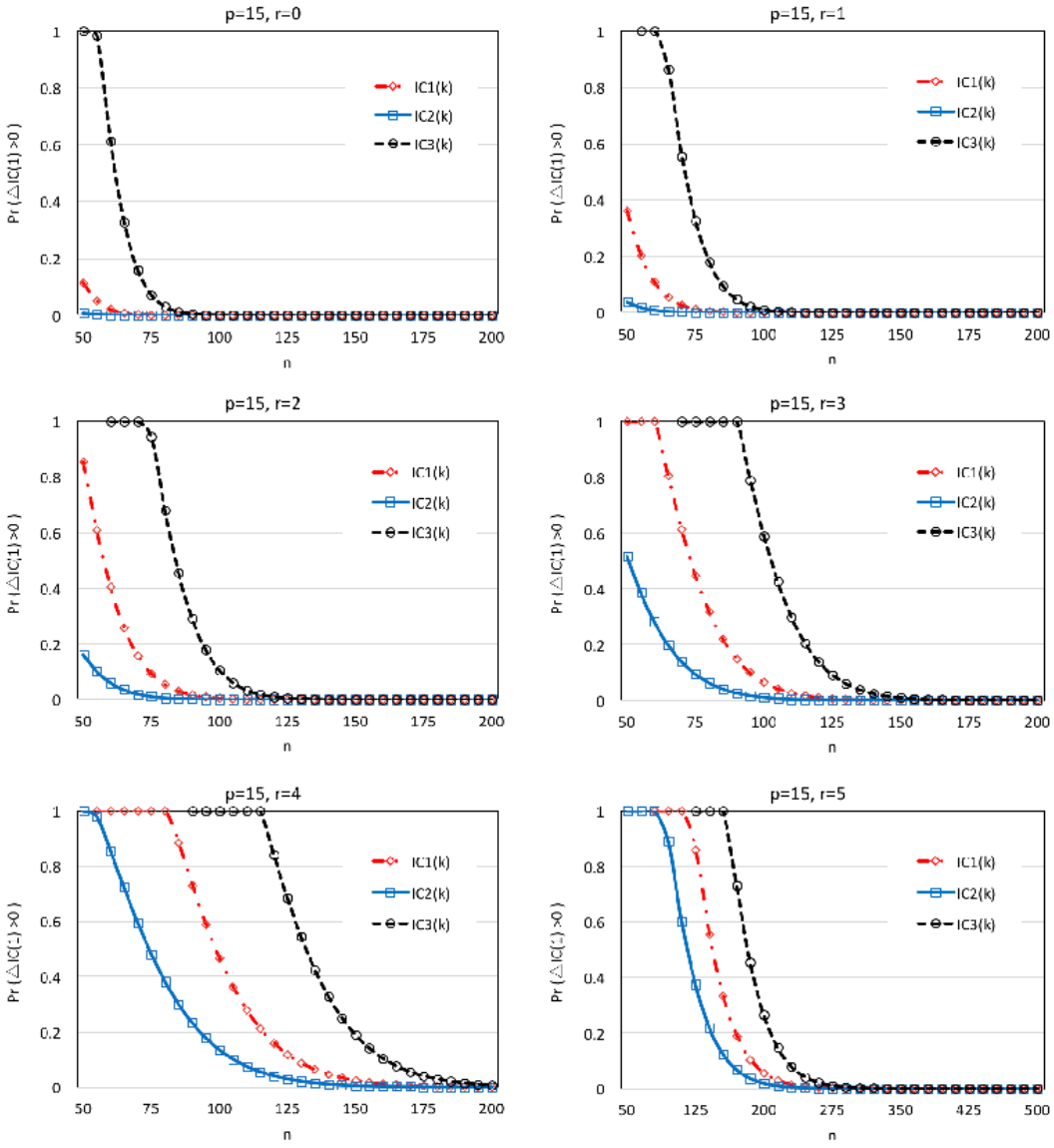
	$r=3$			$r=4$			$r=5$		
(n,p)	IC_1	IC_2	IC_3	IC_1	IC_2	IC_3	IC_1	IC_2	IC_3
(50,15)	1.2309	0.5168	n.a	n.a	1.0344	n.a	n.a	1.8167	n.a
(50,20)	0.2339	0.0082	n.a	0.5758	0.0384	n.a	1.0259	0.1461	n.a
(50,25)	0.0115	0.0000	n.a	0.0406	0.0002	n.a	0.1240	0.0011	n.a
(60,15)	1.0016	0.2808	n.a	1.6445	0.8549	n.a	n.a	1.4372	n.a
(60,20)	0.0585	0.0017	1.7174	0.2173	0.0114	n.a	0.6127	0.0615	n.a
(60,25)	0.0010	0.0000	1.1721	0.0055	0.0000	1.8839	0.0251	0.0002	n.a
(75,15)	0.4475	0.0941	1.6053	1.0719	0.4816	n.a	n.a	1.1159	n.a
(75,20)	0.0052	0.0001	0.6488	0.0359	0.0015	1.1068	0.1810	0.0140	1.9739
(75,25)	0.0000	0.0000	0.2084	0.0002	0.0000	0.5469	0.0016	0.0000	1.0368
(100,15)	0.0654	0.0110	0.5898	0.4660	0.1346	1.3252	1.2246	0.7908	n.a
(100,20)	0.0000	0.0000	0.0220	0.0010	0.0000	0.1442	0.0137	0.0008	0.6050
(200,15)	0.0000	0.0000	0.0000	0.0007	0.0001	0.0072	0.0547	0.0179	0.2643
(10,50)	0.8911	0.3052	n.a	0.9420	0.3592	n.a	0.9820	0.4192	n.a
(15,50)	0.0841	0.0026	1.1854	0.1134	0.0041	1.4509	0.1509	0.0065	1.8413
(20,50)	0.0034	0.0000	1.1129	0.0058	0.0000	1.3208	0.0097	0.0000	1.6985
(10,60)	0.4078	0.0879	1.0044	0.4608	0.1073	1.0159	0.5173	0.1301	1.0502
(15,60)	0.0062	0.0001	0.5059	0.0090	0.0002	0.5837	0.0129	0.0003	0.6653
(20,60)	0.0000	0.0000	0.2987	0.0000	0.0000	0.3733	0.0002	0.0000	0.4590
(10,75)	0.0585	0.0080	0.3764	0.0699	0.0101	0.4180	0.0832	0.0128	0.4623
(15,75)	0.0000	0.0000	0.0195	0.0000	0.0000	0.0255	0.0001	0.0000	0.0332
(20,75)	0.0000	0.0000	0.0023	0.0000	0.0000	0.0035	0.0000	0.0000	0.0051
(10,100)	0.0006	0.0000	0.0089	0.0007	0.0000	0.0108	0.0009	0.0000	0.0129
(15,100)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
(10,200)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Note: This table reports an upper bound on the overestimation probability of the IC estimator, $\Pr(\triangle IC(1) > 0)$ defined in Theorem 2.1, subject to the true number of factors $r \in \{3, 4, 5\}$ and the choice of panel information criteria. If a probability bound is less than 1.0×10^{-4}, we simply put a zero. In some cases, we report an upper bound which is larger than one because it helps compare the magnitude of over-detection risks. *"n.a"* ("Not Applicable") indicates that an appropriate positive number of s which satisfies (2.6.2) is not available in this case.

tion, Figure 2.1 plots an upper bound on the overestimation probability of the IC estimator for the cases with $p = 15$ and increasing n from 50 to 200, while Figure 2.2 depicts the cases with $n = 10$ and increasing p from 50 to 200. For each value of $r \in \{0, 1, 2, 3, 4, 5\}$, each panel compares the performances of three different panel information criteria: $IC_1(k)$, $IC_2(k)$, and $IC_3(k)$.

In these Figures, we can see that the findings from Table 2.1 are true for all the formula-

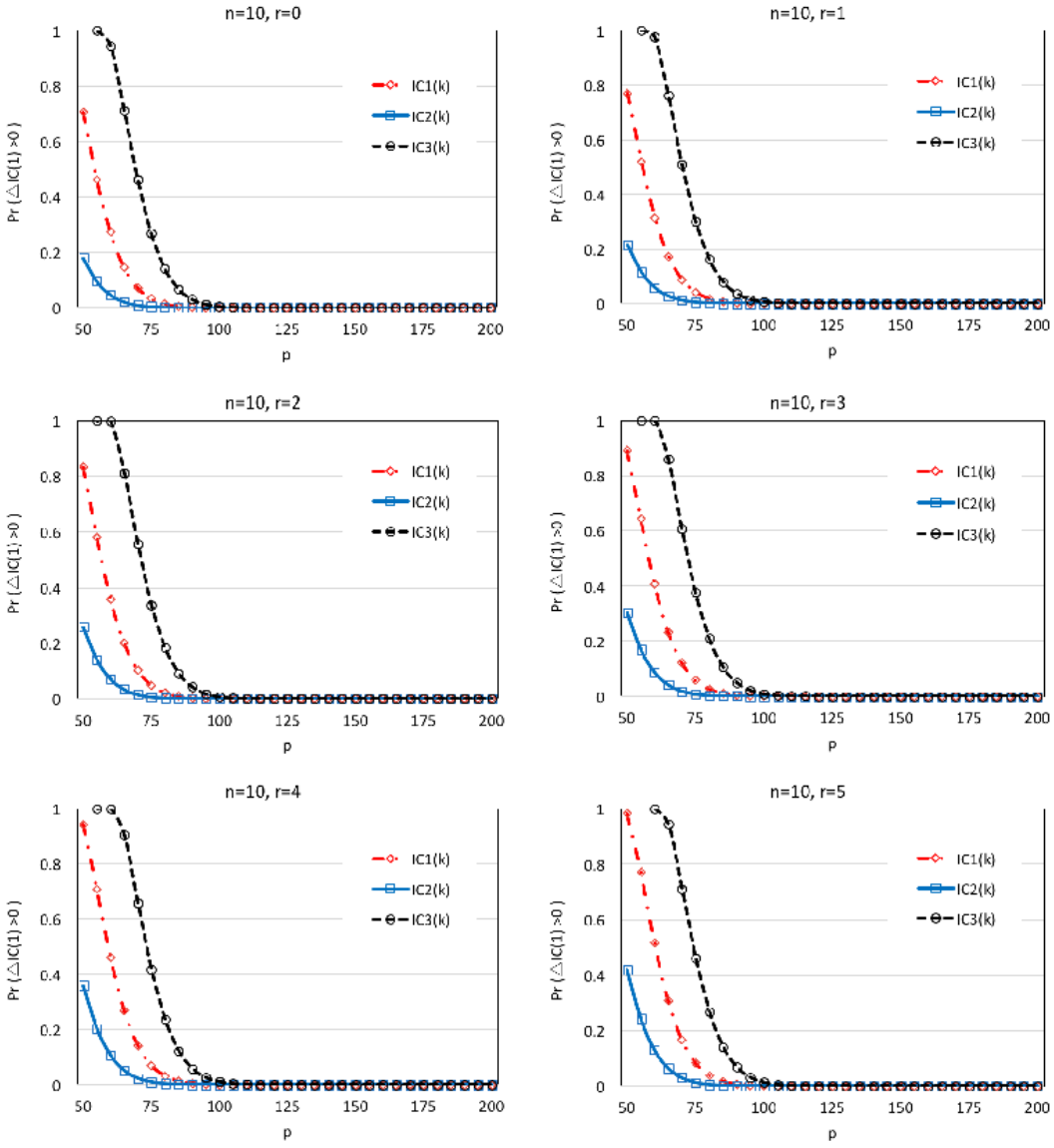
Figure 2.1: Detection Performance of the IC estimator (I.I.D. Errors, $n > p$)



Note: This plots an upper bound on the overestimation probability of the IC estimator, $\Pr(\triangle IC(1) > 0)$ defined in Theorem 2.1. A bound is computed by the formula (2.6.1). We consider the true number of factors $r \in \{0, 1, 2, 3, 4, 5\}$ such that $r = \hat{k}_{IC} - 1$. We only present the case with $p = 15$ and increasing sample sizes from 50 to 200 (Note, when $r = 5$, the maximum number of n is set to 500). Each panel compares the detection performances of three different panel information criteria, $IC_1(k)$, $IC_2(k)$ and $IC_3(k)$ which are defined in (2.3.3), (2.3.4) and (2.3.5), respectively.

tions of the penalty function. When we choose $G_3(p, n)$ as a penalty function, or equivalently when we use $IC_3(k)$, however, upper bounds on the overestimation probability are particularly high. On the other hand, we obtain much lower bounds for the case with $G_2(p, n)$ than

Figure 2.2: Detection Performance of the IC estimator (I.I.D. Errors, $p > n$)



Note: This plots an upper bound on the overestimation probability of the IC estimator, $\Pr(\triangle IC(1) > 0)$ defined in Theorem 2.1. A bound is computed by the formula (2.6.1). We consider the true number of factors $r \in \{0, 1, 2, 3, 4, 5\}$ such that $r = \hat{k}_{IC} - 1$. We only present the case with $n = 10$ and increasing p from 50 to 200. Each panel compares the detection performances of three different panel information criteria, $IC_1(k)$, $IC_2(k)$ and $IC_3(k)$ which are defined in (2.3.3), (2.3.4) and (2.3.5), respectively.

other formulations. Such performance differences can be explained as follows. In finite samples, $\frac{p+n}{pn} > \frac{1}{p}$ and $\ln p > \ln\left(\frac{pn}{p+n}\right)$; therefore, $G_3(p,n) < G_2(p,n)$, and $G_1(p,n) < G_2(p,n)$. It implies that $\left(\frac{pn}{p+n}\right)$ provides a small-sample correction to the asymptotic convergence rate

of p so that $G_2(p, n)$ is a higher penalty for overfitting (Bai and Ng, 2002). Consequently, $IC_2(k)$ yields the lowest overestimation probability among three panel information criteria; however, such differences become negligible as the sample size grows.

We can also see that the overestimation probability given the sample size tends to increase as the number of factors r grows. As Nadler (2010) pointed out, the reason stems from a decrease in the effect of the error components. Recall that we assume $\hat{k}_{IC} = r+1$. As r increases so that $\hat{k}_{IC}$ increases as well, the dimension of a noise subspace $p - \hat{k}_{IC}$ shrinks; consequently, the effect of the idiosyncratic components weakens, whereas the relative explanatory power of signals is likely to be overly inflated.

Obviously, when the sample size is not sufficiently small, we obtain nearly zero upper bound (not reported here). In particular, when n is greater than 200, we obtain practically zero overestimation probability bounds, say less than 10^{-5} in most cases.

2.7 Modified Information Criteria

2.7.1 Improved Penalty for Overfitting

In this section, we provide a practical guide for users who may worry about the overdetection of IC in their empirical research. We demonstrate here that a simple modification of IC (called *modified criteria*), which gives an increase in the penalty for overfitting, leads to a negligible over-detection risk in finite samples and consequently a substantial improvement of detection performance. First, by using Theorem 2.2, we compute theoretical upper bounds on the overestimation probability of the modified criteria. As a consequence, we show the better performance of the modified criteria than IC for the case with Gaussian i.i.d. errors. Next, via Monte Carlo simulations, we also analyze the detection performance of the modified criteria for the case with weak serial or/and cross-sectional dependence of the error terms.

As seen before, the IC estimator often results in a non-negligible overestimation proba-

bility for the case with small samples. Obviously, this result raises an interesting question of how to make this over-detection risk negligible. Here is a clue to the answer. As Hallin and Liška (2007) and Ahn and Horenstein (2013) pointed out, the penalty function defined by Bai and Ng (2002) is not unique since it is only required to satisfy certain asymptotic conditions for the consistency of the IC estimator; for example, any fixed scalar multiple of $G(p, n)$ still satisfies the asymptotic conditions. Their finite sample properties are different, however, due to a scalar multiple. Such notions imply that we can improve the finite sample performance of IC by simply modifying its penalty term while preserving its asymptotic consistency. Nadler (2010) applied this idea to the Akaike information criterion (AIC) for signal detection; specifically, after the original penalty term in the criterion is multiplied by an arbitrary constant, this *modified* AIC yields better performance.

This chapter adopts a different approach to improve the penalty for overfitting. In our modified criteria, degrees of freedom in the penalty term are adjusted for the number of factors because the effective dimension is $p - k$ rather than p in the presence of a strictly positive number of factors. Our approach is in line with Ng and Perron (2005) on the sensitivity of model selection criteria to sample sizes and degrees of freedom in finite samples. They consider different penalty terms by various degrees of freedom adjustments; as a consequence, they show that the lag-length selected by the AIC or the Bayesian information criterion (BIC) is quite sensitive to degrees of freedom adjustments. Since there has been no definitive guide for such an adjustment, Ng and Perron (2005) instead provide a practical guide for practitioners through extensive experiments. In particular, they consider the following adjustments: $p - k$, $p - 2k$, and $p - kmax$. In fact, they also consider the case in which the sum of squared residuals is adjusted for degrees of freedom; that is, the sum of squared residuals is divided by $(p - k)n$, $(p - 2k)n$, or $(p - kmax)n$ rather than pn. In our study, however, the latter option is not considered since the formula for the overestimation probability bound of IC given by (2.6.1) is not affected by the denominator of $S(k)$ defined in (2.3.2).

Recall the original IC given by (2.3.1) as proposed by Bai and Ng (2002),

$$IC(k) = \ln S(k) + k \cdot G(p,n),$$

where k is an arbitrary number ($k < \min\{p,n\}$), and $S(k)$ is the sum of squared residuals is divided by pn. $G(p,n)$ is the penalty function which has three different forms: $G_1(p,n) = \left(\frac{p+n}{pn}\right)\ln\left(\frac{pn}{p+n}\right)$; $G_2(p,n) = \left(\frac{p+n}{pn}\right)\ln C_{pn}^2$; and $G_3(p,n) = \frac{\ln C_{pn}^2}{C_{pn}^2}$, where $C_{pn} = \min\{\sqrt{p},\sqrt{n}\}$. Now, we denote by MIC our modified panel information criteria. Then, MIC has the form

$$MIC(k) = \ln S(k) + k \cdot mG(p,n,k), \tag{2.7.1}$$

where $mG(p,n,k)$ is a new penalty factor which modifies $G(p,n)$ by degrees of freedom adjustment. Moreover, the *modified* estimator for the true number of factors (hereafter, MIC estimator) is defined as the minimizer of $MIC(k)$ over a range of values for k, namely that

$$\hat{k}_{_{MIC}} = \underset{0\le k\le kmax}{\arg\min} MIC(k). \tag{2.7.2}$$

To sum up, the only difference between IC and MIC is a penalty function. For this reason, a non-asymptotic bound on the overestimation probability of MIC is the same as that of IC except for a penalty function. Under the conditions in Theorem 2.2, it is given by

$$\begin{aligned}\Pr(\triangle MIC(1) > 0) \le & \exp\left(\frac{-(p-r)s^2}{4}\right) + \\ & \exp\left(-\frac{4n}{3}(\bar{c})^{1/4}\left((p-r)\left(1-\frac{s}{\sqrt{n}}\right)\widetilde{\xi}_{n,p,k} - (1+\sqrt{\bar{c}})^2\right)^{3/2}\right),\end{aligned} \tag{2.7.3}$$

where $\bar{c} = \frac{p-r}{n}$ and $\widetilde{\xi}_{n,p,k} = -1 + \sqrt{1 + 2\cdot mG(p,n,k)}$. Obviously, this bound is appropriate for any positive value of s chosen by a user such that

$$\sqrt{n} - \frac{1}{\widetilde{\xi}_{n,p,k}\sqrt{p-r}}\left(3 + \sqrt{\bar{c}} + \frac{1}{\sqrt{\bar{c}}}\right) < s < \sqrt{n} - \frac{1}{\widetilde{\xi}_{n,p,k}\sqrt{p-r}}\left(2 + \sqrt{\bar{c}} + \frac{1}{\sqrt{\bar{c}}}\right).$$

In particular, we consider the following modified penalty function which has obviously three different choices corresponding to three original penalty terms.

Definition 2.2 (***Modified penalty function***). *Let $mG(p,n,k)$ denote a modified penalty function. It has three different choices given by*

$$mG_1(p,n,k) = \left(\frac{N+n}{Nn}\right)\ln\left(\frac{pn}{p+n}\right); \tag{2.7.4}$$

$$mG_2(p,n,k) = \left(\frac{N+n}{Nn}\right)\ln C_{pn}^2; \tag{2.7.5}$$

$$mG_3(p,n,k) = \frac{\ln C_{pn}^2}{C_{Nn}^2}, \tag{2.7.6}$$

where $N = p - \alpha k > 0$ with a fixed strictly positive integer α, $C_{pn} = \min\{p,n\}$, and $C_{Nn} = \min\{N,n\}$.

Note that the above modified penalty function is designed in order to provide a small-sample correction to the original IC estimator while preserving its consistency. Our degrees of freedom adjustment leads to an increase in the penalty term of the original IC. $mG(p,n,k)$ is higher than $G(p,n)$ when $k>0$ since we have $\frac{N+n}{Nn} > \frac{p+n}{pn}$. Note that $mG_3(p,n,k)$ gives a higher penalty than $G_3(p,n)$ only when $n > N$.

Finally, we define the modified panel information criteria, MIC, in relation to the above three modified penalty terms:

$$MIC_1(k) = \ln S(k) + k \cdot mG_1(p,n,k); \tag{2.7.7}$$

$$MIC_2(k) = \ln S(k) + k \cdot mG_2(p,n,k); \tag{2.7.8}$$

$$MIC_3(k) = \ln S(k) + k \cdot mG_3(p,n,k). \tag{2.7.9}$$

Here we explore in more detail some properties of our modified penalty function. First, $mG(k)$ is strictly convex in k. For given n and p, $mG(k)$ is a twice differentiable function of k, and its second derivative is non negative on the interval $[0, kmax]$. The strict convexity

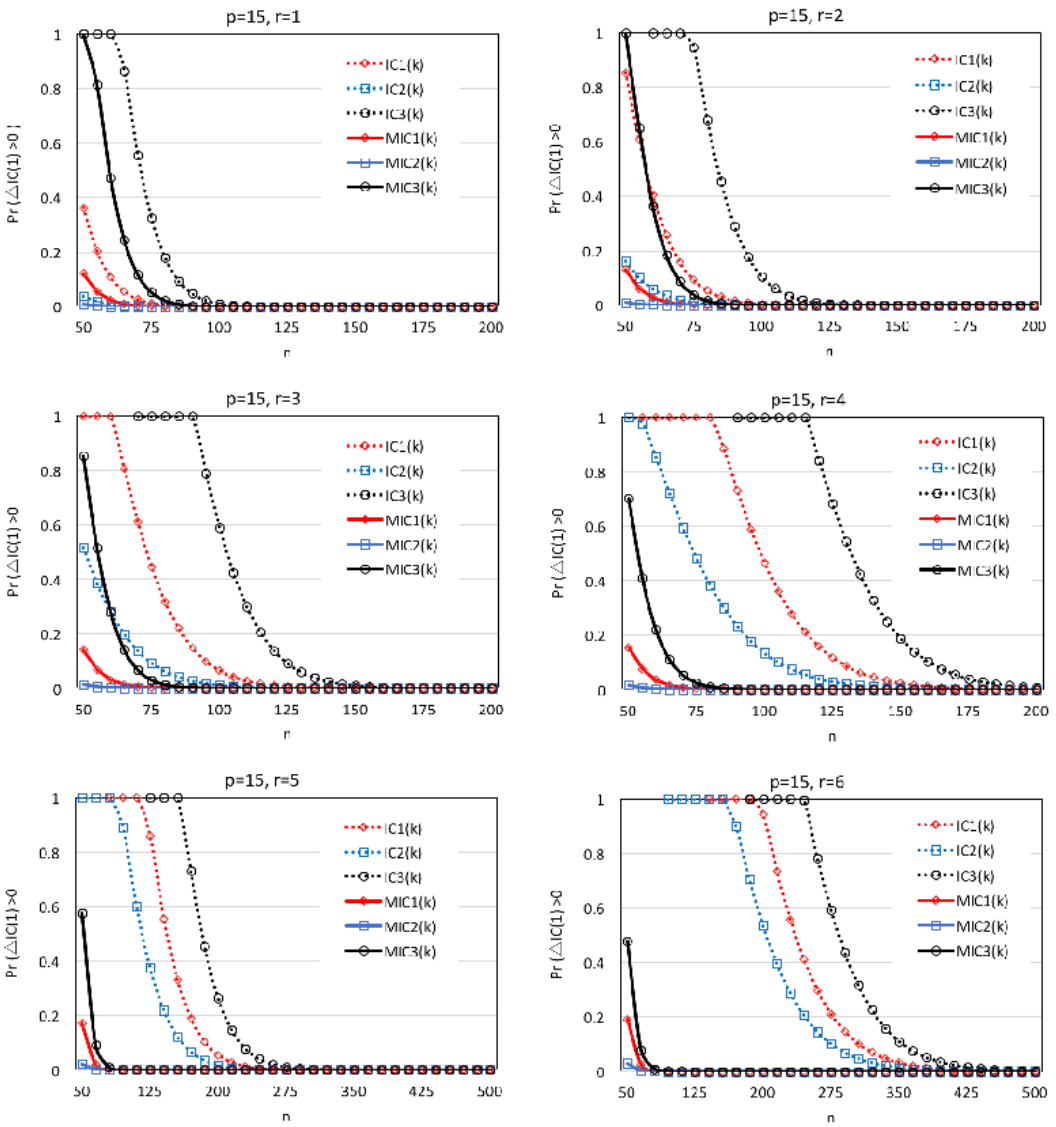
of the penalty and the squared error loss leads to the strictly convex optimization problem (2.7.2) so that a unique solution (a global minimum) exists. Second, α governs the magnitude of improved penalization. A large α leads to an increase in the penalty for overfitting, given fixed n, p and k. Lastly, our modified penalty factor also satisfies the asymptotic conditions for the consistency of the estimator: (i) $mG(p,n,k) \to 0$, and (ii) $C_{pn}^2 \cdot mG(p,n,k) \to \infty$ as $n, p \to \infty$ because k is fixed regardless of n and p. Thus, the *MIC* estimator is consistent, namely that $\lim_{n,p\to\infty} \Pr(\hat{k}_{MIC} = r) = 1$.

2.7.2 Detection Performance of the *MIC* estimator

Now, as a counterpart to the performance analysis of the *IC* estimator in Section 2.6.2, we examine the finite sample performance of the *MIC* estimator by using the formula for a non-asymptotic bound on the overestimation probability of *MIC* given by (2.7.3). Note that this theoretical analysis is only feasible for the case with Gaussian i.i.d. errors. In the next section, we perform more general analyses allowing the serially or/and cross-sectionally correlated error components through Monte Carlo experiments.

First, for the case with $n > p = 15$, Figure 2.3 and 2.4 compare the detection performances of the original *IC* and the modified criteria, *MIC*, given the true number of factors $r \in \{1,2,3,4,5,6\}$ and $kmax = 8$. Here we consider three different versions of *MIC* corresponding to the choice of a penalty function: MIC_1, MIC_2, and MIC_3. As depicted in these Figures, *MIC* yields much lower overestimation probabilities than *IC* in all cases. In particular, Figure 2.3 considers *MIC* with $\alpha = 1$ ($N = p-k$), while Figure 2.4 considers *MIC* with $\alpha = 2$ ($N = p-2k$) which leads to a higher penalty for overfitting. Consequently, the overestimation probability falls more substantially in Figure 2.4 than in Figure 2.3 across all choices of penalty terms and various numbers of factors. Moreover, as r grows (so that $\hat{k}_{IC}$ increases), the performance improvement becomes significant. Especially for the case with $r \geq 3$, it results in nearly zero probabilities. Even for the case with $r \in \{1,2\}$, upper bounds fall below 50%. As the sample size increases, however, the difference between *IC* and *MIC*

Figure 2.3: Performance Comparison between MIC ($\alpha = 1$) and IC (I.I.D. Errors, $n > p$)

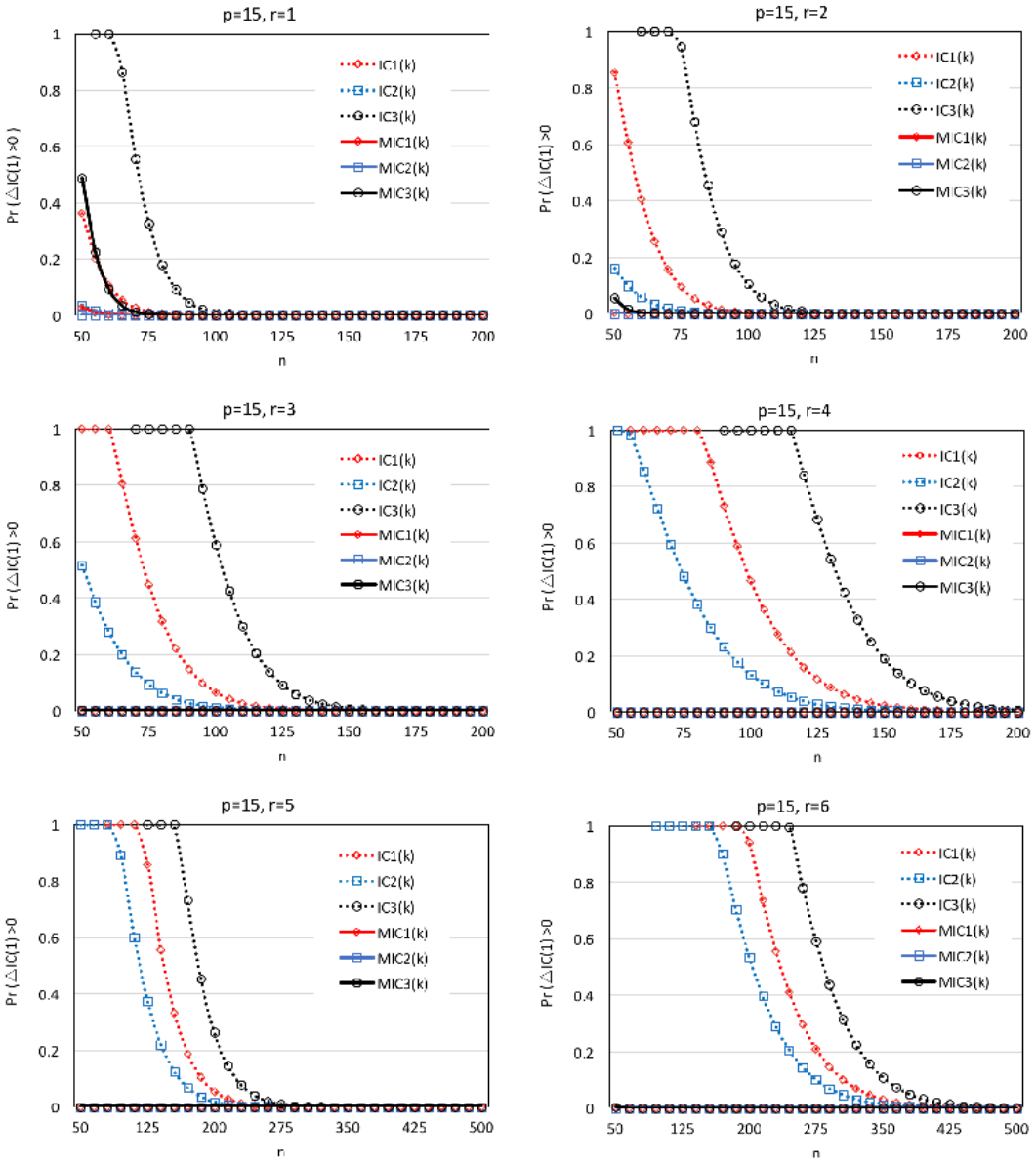


Note: This compares an upper bound on the overestimation probability of the IC estimator computed by (2.6.1) with that of the MIC estimator computed by (2.7.3). For the analysis of MIC estimator, we set $\alpha = 1$. We only present the case of $p = 15$ subject to $r \in \{1, 2, 3, 4, 5, 6\}$ which is the true number of factors and increasing sample sizes from 50 to 200. (Note, when $r \in \{5, 6\}$, the maximum number of n is 500). Each panel plots the performances of three different original criteria, $IC_1(k)$, $IC_2(k)$ and $IC_3(k)$ which are defined in (2.3.3), (2.3.4) and (2.3.5), respectively, along with the performances of three different modified criteria, $MIC_1(k)$, $MIC_2(k)$ and $MIC_3(k)$ which are defined in (2.7.7), (2.7.8) and (2.7.9), respectively.

becomes negligible since the original IC already yields sufficiently low probability bounds of overestimation for the case with large sample sizes.

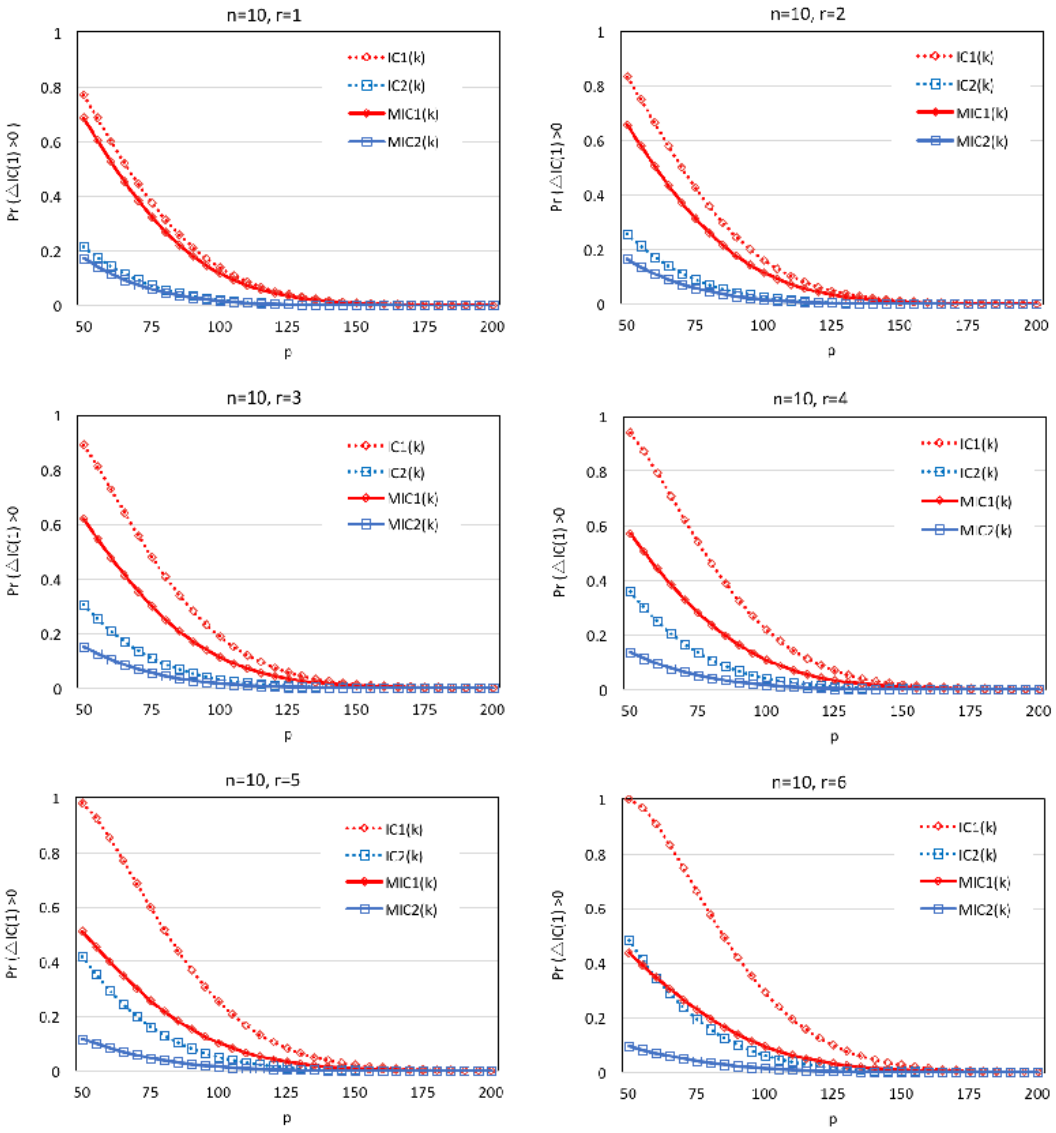
Second, for the case with $p > n = 10$, Figure 2.5 compares the detection performances of

Figure 2.4: Performance Comparison between MIC ($\alpha = 2$) and IC (I.I.D. Errors, $n > p$)



Note: This compares an upper bound on the overestimation probability of the IC estimator computed by (2.6.1) with that of the MIC estimator computed by (2.7.3). For the analysis of MIC estimator, we set $\alpha = 2$. We only present the case of $p = 15$ subject to $r \in \{1, 2, 3, 4, 5, 6\}$ which is the true number of factors and increasing sample sizes from 50 to 200. (Note, when $r \in \{5, 6\}$, the maximum number of n is 500). Each panel plots the performances of three different original criteria, $IC_1(k)$, $IC_2(k)$ and $IC_3(k)$ which are defined in (2.3.3), (2.3.4) and (2.3.5), respectively, along with the performances of three different modified criteria, $MIC_1(k)$, $MIC_2(k)$ and $MIC_3(k)$ which are defined in (2.7.7), (2.7.8) and (2.7.9), respectively.

Figure 2.5: Performance Comparison between MIC ($\alpha = 3$) and IC (I.I.D. Errors, $p > n$)



Note: This compares an upper bound on the overestimation probability of the IC estimator computed by (2.6.1) with that of the MIC estimator computed by (2.7.3). For the analysis of MIC estimator, we set $\alpha = 3$. We only present the case of $n = 10$ subject to $r \in \{1, 2, 3, 4, 5, 6\}$ which is the true number of factors and increasing p from 50 to 200. Each panel plots the performances of two different original criteria, $IC_1(k)$ and $IC_2(k)$ which are defined in (2.3.3) and (2.3.4), respectively, along with the performances of two different modified criteria, $MIC_1(k)$ and $MIC_2(k)$ which are defined in (2.7.7) and (2.7.8), respectively.

IC and MIC with $\alpha = 3$, given the true number of factors $r \in \{1, 2, 3, 4, 5, 6\}$ and $kmax = 8$. Here we consider the cases in which MIC_1 and MIC_2 are used. Obviously, MIC_3 is not considered here since $mG_3 = G_3$ when $N > n$. Figure 2.5 shows that MIC yields lower

Table 2.2: Performance Comparison between MIC ($\alpha = 2$) and IC (I.I.D. Errors, $n > p$)

	$r = 1$						$r = 2$					
(n, p)	IC_1	IC_2	IC_3	MIC_1	MIC_2	MIC_3	IC_1	IC_2	IC_3	MIC_1	MIC_2	MIC_3
(50,10)	1.8276	1.0309	n.a	0.3581	0.0962	0.7882	n.a	1.9685	n.a	0.0351	0.0069	0.0611
(50,15)	0.3620	0.0376	n.a	0.0290	0.0014	0.4882	0.8541	0.1615	n.a	0.0042	0.0002	0.0558
(50,20)	0.0210	0.0002	1.5879	0.0013	0.0000	0.4350	0.0768	0.0014	n.a	0.0002	0.0000	0.0652
(60,10)	1.1934	0.8002	n.a	0.1210	0.0291	0.3186	n.a	1.5334	n.a	0.0060	0.0011	0.0109
(60,15)	0.1082	0.0093	1.0661	0.0034	0.0001	0.0932	0.4056	0.0605	1.8685	0.0003	0.0000	0.0049
(60,20)	0.0021	0.0000	0.7165	0.0000	0.0000	0.0462	0.0124	0.0002	1.0746	0.0000	0.0000	0.0031
(75,10)	0.8808	0.4054	1.3870	0.0189	0.0041	0.0575	1.8811	1.1264	n.a	0.0004	0.0000	0.0007
(75,15)	0.0128	0.0009	0.3259	0.0001	0.0000	0.0043	0.0939	0.0114	0.9460	0.0000	0.0000	0.0000
(75,20)	0.0000	0.0000	0.0577	0.0000	0.0000	0.0006	0.0006	0.0000	0.2243	0.0000	0.0000	0.0000
(100,10)	0.2529	0.0953	0.6738	0.0006	0.0001	0.0021	1.0892	0.7881	1.7878	0.0000	0.0000	0.0000
(100,15)	0.0002	0.0000	0.0107	0.0000	0.0000	0.0000	0.0049	0.0005	0.1061	0.0000	0.0000	0.0000
(200,10)	0.0002	0.0000	0.0008	0.0000	0.0000	0.0000	0.0410	0.0185	0.1039	0.0000	0.0000	0.0000

	$r = 3$						$r = 4$					
(n, p)	IC_1	IC_2	IC_3	MIC_1	MIC_2	MIC_3	IC_1	IC_2	IC_3	MIC_1	MIC_2	MIC_3
(50,15)	1.2309	0.5168	n.a	0.0003	0.0000	0.0018	n.a	1.0344	n.a	0.0000	0.0000	0.0000
(50,20)	0.2339	0.0082	n.a	0.0000	0.0000	0.0041	0.5758	0.0384	n.a	0.0000	0.0000	0.0000
(50,25)	0.0115	0.0000	n.a	0.0000	0.0000	0.0100	0.0406	0.0002	n.a	0.0000	0.0000	0.0005
(60,15)	1.0016	0.2808	n.a	0.0000	0.0000	0.0000	1.6445	0.8549	n.a	0.0000	0.0000	0.0000
(60,20)	0.0585	0.0017	1.7174	0.0000	0.0000	0.0000	0.2173	0.0114	n.a	0.0000	0.0000	0.0000
(60,25)	0.0010	0.0000	1.1721	0.0000	0.0000	0.0001	0.0055	0.0000	1.8839	0.0000	0.0000	0.0000
(75,15)	0.4475	0.0941	1.6053	0.0000	0.0000	0.0000	1.0719	0.4816	n.a	0.0000	0.0000	0.0000
(75,20)	0.0052	0.0001	0.6488	0.0000	0.0000	0.0000	0.0359	0.0015	1.1068	0.0000	0.0000	0.0000
(75,25)	0.0000	0.0000	0.2084	0.0000	0.0000	0.0000	0.0002	0.0000	0.5469	0.0000	0.0000	0.0000
(100,15)	0.0654	0.0110	0.5898	0.0000	0.0000	0.0000	0.4660	0.1346	1.3252	0.0000	0.0000	0.0000
(100,20)	0.0000	0.0000	0.0220	0.0000	0.0000	0.0000	0.0010	0.0000	0.1442	0.0000	0.0000	0.0000
(200,10)	1.0508	0.7857	1.2705	0.0000	0.0000	0.0000	n.a	n.a	n.a	0.0003	0.0003	0.0003

Note: This table compares an upper bound on the overestimation probability of the IC estimator computed by (2.6.1) with that of the MIC estimator computed by (2.7.3) depending on various sample sizes (n, p) such that $n > p$. For the analysis of MIC estimator, we set $\alpha = 2$. We consider sample sizes (n, p) such that $n \in \{50, 60, 75, 100, 200\}$ and $p \in \{10, 15, 20, 25, 50\}$, but a few cases which show negligible probability bounds are not reported here. Three different panel information criteria of the original IC, $IC_1(k)$, $IC_2(k)$ and $IC_3(k)$, are defined in (2.3.3), (2.3.4) and (2.3.5), respectively. Three different panel information criteria of the MIC, $MIC_1(k)$, $MIC_2(k)$ and $MIC_3(k)$, are defined in (2.7.7), (2.7.8) and (2.7.9), respectively. If a probability bound is less than 1.0×10^{-4}, we simply put a zero. In some cases, we report an upper bound which is larger than one because it helps compare the magnitude of over-detection risks. *"n.a"* ("Not Applicable") indicates that an appropriate positive number of s which satisfies (2.6.2) is not available in this case.

Table 2.3: Performance Comparison between MIC ($\alpha = 3$) and IC (I.I.D. Errors, $p > n$)

	$r = 1$						$r = 2$					
(n, p)	IC_1	IC_2	IC_3	MIC_1	MIC_2	MIC_3	IC_1	IC_2	IC_3	MIC_1	MIC_2	MIC_3
(10,50)	0.7705	0.2153	n.a	0.6859	0.1734	n.a	0.8330	0.2573	n.a	0.6579	0.1646	n.a
(15,50)	0.0445	0.0010	1.0170	0.0276	0.0005	1.0170	0.0616	0.0016	1.0616	0.0228	0.0004	1.0616
(20,50)	0.0011	0.0000	1.0074	0.0005	0.0000	1.0074	0.0020	0.0000	1.0318	0.0003	0.0000	1.0318
(10,60)	0.3137	0.0581	0.9753	0.2688	0.0469	0.9753	0.3588	0.0717	0.9978	0.2616	0.0462	0.9978
(15,60)	0.0029	0.0000	0.3675	0.0018	0.0000	0.3675	0.0042	0.0000	0.4335	0.0016	0.0000	0.4335
(20,60)	0.0000	0.0000	0.1830	0.0000	0.0000	0.1830	0.0000	0.0000	0.2355	0.0000	0.0000	0.2355
(10,75)	0.0405	0.0049	0.3013	0.0342	0.0040	0.3013	0.0487	0.0063	0.3374	0.0346	0.0042	0.3374
(15,75)	0.0000	0.0000	0.0111	0.0000	0.0000	0.0111	0.0000	0.0000	0.0147	0.0000	0.0000	0.0147
(20,75)	0.0000	0.0000	0.0010	0.0000	0.0000	0.0010	0.0000	0.0000	0.0015	0.0000	0.0000	0.0015
(10,100)	0.0004	0.0000	0.0062	0.0003	0.0000	0.0062	0.0005	0.0000	0.0074	0.0003	0.0000	0.0074
(15,100)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
(10,200)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

	$r = 3$						$r = 4$					
(n, p)	IC_1	IC_2	IC_3	MIC_1	MIC_2	MIC_3	IC_1	IC_2	IC_3	MIC_1	MIC_2	MIC_3
(10,50)	0.8911	0.3052	n.a	0.6203	0.1523	n.a	0.9420	0.3592	n.a	0.5717	0.1364	n.a
(15,50)	0.0841	0.0026	1.1854	0.0178	0.0003	1.1854	0.1134	0.0041	1.4509	0.0129	0.0002	1.4509
(20,50)	0.0034	0.0000	1.1129	0.0002	0.0000	1.1129	0.0058	0.0000	1.3208	0.0000	0.0000	1.3208
(10,60)	0.4078	0.0879	1.0044	0.2509	0.0446	1.0044	0.4608	0.1073	1.0159	0.2364	0.0421	1.0159
(15,60)	0.0062	0.0001	0.5059	0.0013	0.0000	0.5059	0.0090	0.0002	0.5837	0.0011	0.0000	0.5837
(20,60)	0.0000	0.0000	0.2987	0.0000	0.0000	0.2987	0.0000	0.0000	0.3733	0.0000	0.0000	0.3733
(10,75)	0.0585	0.0080	0.3764	0.0345	0.0042	0.3764	0.0699	0.0101	0.4180	0.0341	0.0043	0.4180
(15,75)	0.0000	0.0000	0.0195	0.0000	0.0000	0.0195	0.0000	0.0000	0.0255	0.0000	0.0000	0.0255
(20,75)	0.0000	0.0000	0.0023	0.0000	0.0000	0.0023	0.0000	0.0000	0.0035	0.0000	0.0000	0.0035
(10,100)	0.0006	0.0000	0.0089	0.0004	0.0000	0.0089	0.0007	0.0000	0.0108	0.0004	0.0000	0.0108
(15,100)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
(10,200)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Note: This table compares an upper bound on the overestimation probability of the IC estimator computed by (2.6.1) with that of the MIC estimator computed by (2.7.3) depending on various sample sizes (n, p) such that $p > n$. For the analysis of MIC estimator, we set $\alpha = 3$. We consider sample sizes (n, p) such that $n \in \{10, 15, 20, 25, 50\}$ and $p \in \{50, 60, 75, 100, 200\}$, but a few cases which show negligible probability bounds are not reported here. Three different panel information criteria of the original IC, $IC_1(k)$, $IC_2(k)$ and $IC_3(k)$, are defined in (2.3.3), (2.3.4) and (2.3.5), respectively. Three different panel information criteria of the MIC, $MIC_1(k)$, $MIC_2(k)$ and $MIC_3(k)$, are defined in (2.7.7), (2.7.8) and (2.7.9), respectively. If a probability bound is less than 1.0×10^{-4}, we simply put a zero. In some cases, we report an upper bound which is larger than one because it helps compare the magnitude of over-detection risks. *"n.a"* ("Not Applicable") indicates that an appropriate positive number of s which satisfies (2.6.2) is not available in this case.

overestimation probabilities than IC; especially, upper bounds decrease more sharply as the number of factors increases. For the case with $r \geq 5$, upper bounds fall below 50%. Similarly to the case with $n > p$, the performance improvement becomes negligible as p increases.

More detailed results are reported in Table 2.2 for the case with $n > p$ and $\alpha = 2$ while Table 2.3 for the case with $p > n$ and $\alpha = 3$. By and large, our modified criteria helps users control over-detection risk when the sample size is small.

2.7.3 Simulation Study

Our performance analysis based on the computable formula for a probability bound is no longer feasible in the presence of serially or/and cross-sectionally correlated error terms. Thus, here we investigate the small sample performance of the MIC estimator for the cases with more general error covariance structures through Monte Carlo simulations.

For our simulation exercises, we generate $1,000$ replications of data produced by the following data-generating process:

$$\begin{aligned} x_{it} &= \sum_{j=1}^{r} \lambda_{ij} f_{jt} + \sqrt{\theta} e_{it}; \\ e_{it} &= \sqrt{\frac{1-\rho^2}{1+2J\beta^2}}\, \varepsilon_{it}; \quad \varepsilon_{it} = \rho \varepsilon_{i,t-1} + \upsilon_{it} + \sum_{j \neq 0,\, j=-J}^{J} \beta \upsilon_{i-j,t}, \end{aligned} \tag{2.7.10}$$

where λ_{ij} and υ_{it} are all drawn from $\mathcal{N}(0,1)$. The factors f_{jt} are drawn from normal distributions with zero means. The same data generating process has been used in Bai and Ng (2002) and Onatski (2010). The magnitude of serial correlation is governed by ρ, and the magnitude of cross-sectional correlation is specified by β. As in Onatski (2010), we set $J = 8$ so that each cross-section unit is correlated with the $16 (= 2J)$ adjacent cross-section units. Further, as in Ahn and Horenstein (2013), we normalize the idiosyncratic components e_{it} so that their variances are equal to 1. The parameter θ controls the relative strength of noise to a signal. When $\text{var}(f_{jt}) = 1$, θ is the same as the inverse of the signal to noise ratio (SNR) of *each* factor since $\theta = \text{var}(\sqrt{\theta}\, e_{it})/\text{var}(f_{jt})$. Thus, we can change SNRs of all factors

by only adjusting the value of θ while fixing variances of factors at 1. Following previous studies, we consider four different correlation structure of the idiosyncratic components: (A) i.i.d. errors ($\rho = \beta = 0$); (B) weakly serially correlated errors ($\rho = 0.5$ and $\beta = 0$); (C) weakly cross-sectionally correlated errors ($\rho = 0$ and $\beta = 0.2$); and (D) both weakly serially and cross-sectionally correlated errors ($\rho = 0.3$ and $\beta = 0.1$). Moreover, in this simulation study, we consider an n-dimension system with p cross-sectional observations as in Bai and Ng (2002).

Our simulation consists of two experiments with different levels of SNR. The first experiment is to examine the finite sample performance of the MIC estimator in the presence of sufficiently strong factors. In particular, we consider the case in which all factors have strong explanatory power by setting $\theta = 0.2$ (SNR=5). Further, we also investigate how the covariance structure of errors affects the detection performance of the MIC estimator. In the second experiment, we consider relatively weaker factors. We set $\theta = 1$ (SNR=1) which implies that the factors explain exactly 50% variation in the data. The effect of correlation structure is also examined. For all experiments, $kmax$ is set to 8, and we use the original IC_1 estimator and its modified version, MIC_1. As shown in Section 2.6, IC_1 yields moderate overestimation probability bounds compared to other extreme cases: IC_2, and IC_3. Here we consider IC_1 as a representative case to check the performance improvement of our modified criteria. Moreover, the performances of IC_1 and MIC_1 are compared with those of other leading estimators: the ED estimator proposed by Onatski (2010), and the ER and GR estimators proposed by Ahn and Horenstein (2013). To analyze detection performances, we report root mean squared errors (RMSEs) of each estimator from $1,000$ simulated datasets. Without loss of generality, we only report results for $r = 3$.

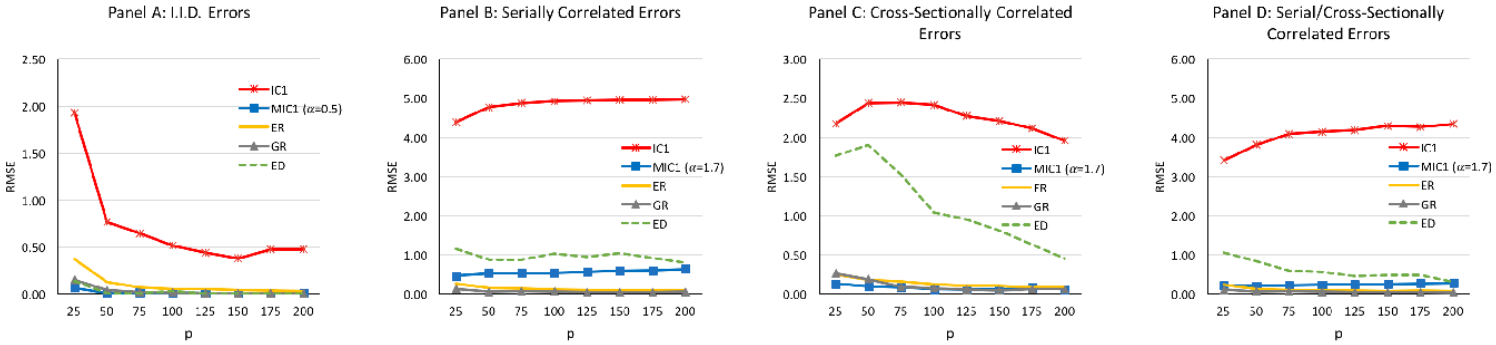
Figure 2.6 and 2.7 report the results from the first part of simulations. Three factors ($r = 3$) are drawn from $\mathcal{N}(0, 1)$, and θ is fixed at 0.2. Thus, all factors have SNRs equal to 5. First, Figure 2.6 depicts cases in which $p > n$; in particular, $n \in \{15, 25\}$ and $p \in \{25, 50, 75, 100, 125, 150, 175, 200\}$. For each n, Panel A shows the results from the data

generated with i.i.d. errors. Although the performance of the IC estimator is not too bad, the performance of the MIC_1 estimator ($\alpha = 0.5$) is much better than the IC_1 estimator; especially for the data with $p \geq 50$, the MIC_1 estimator shows perfect accuracy. Moreover, the MIC_1 estimator performs equally to or better than the ER, GR and ED estimators. For each n, Panels B, C and D report the results from the data with weakly serially or/and cross-sectionally correlated errors. Compared with Panel A for i.i.d. errors, here MIC_1 is more penalized by setting $\alpha = 1.7$ for the data with $n = 15$ while $\alpha = 3$ for the data with $n = 25$. We can see that correlation in the idiosyncratic errors reduces the precision of the IC_1 estimator, while the MIC_1 estimator remains very good in most cases. RMSEs of the MIC_1 estimator are much lower than those of the IC_1 estimator. In addition, the performance of MIC_1 is generally better than that of the ED estimator while being comparable to those of the ER and GR estimators.

Figure 2.7 considers cases where $n > p$; particularly, $n \in \{25, 50, 75, 100, 125, 150, 175, 200\}$ and $p \in \{15, 25\}$. Similarly to the case with $p > n$, the MIC_1 estimator ($\alpha = 2$) outperforms other estimators and shows perfect accuracy for the data with $n \geq 50$ when errors are i.i.d. (Panel A). Moreover, for the case with weak serial correlation (Panel B) as well as with both weak serial and cross-sectional correlation (Panel D), the MIC_1 estimator ($\alpha = 3$) outperforms the IC_1 and ED estimators while performing equally to or slightly less than the ER and GR estimators. For the case with cross-sectionally correlated errors (Panel C), however, RMSEs of the MIC_1 estimator are larger than for other cases in Panels A, B and D; particularly, the detection performance of MIC_1 gets worse as n grows. Such a tendency is also observed in the performance of the IC_1 and ED estimators, while RMSEs of the MIC_1 estimator are still lower than those of the IC_1 and ED estimators for the data with small n. Comparing Panels B and C, it seems that the performances of the MIC_1, IC_1 and ED estimators are more sensitive to cross-sectional correlation than serial correlation. Ahn and Horenstein (2013) reported the similar result from their simulation study.

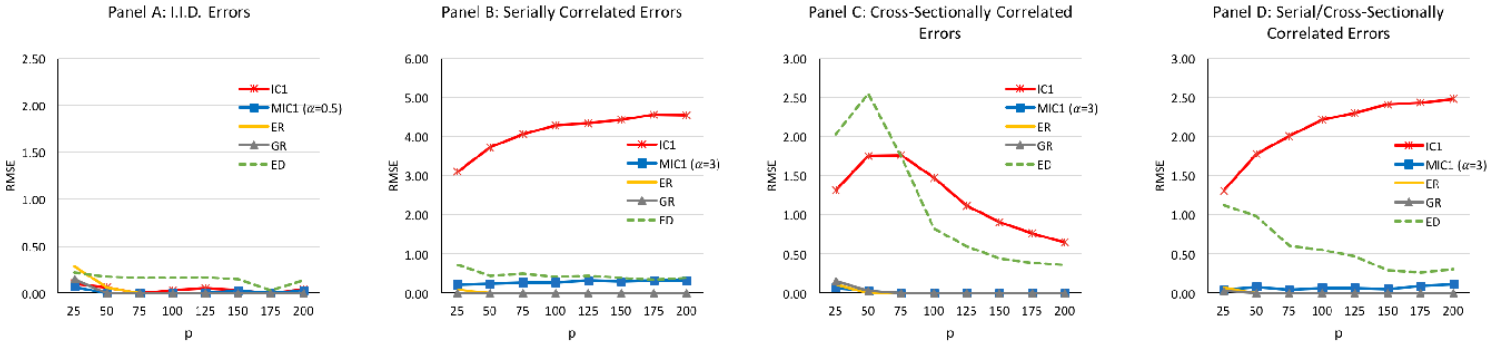
Figure 2.6: Effects of Error Covariance Structure (Three-factor Model, $\theta = 0.2,\ p > n$)

$(1)\ \theta = 0.2,\ p > n = 15$



r=3, θ=0.2, kmax=8, and ρ= β=0 | *r=3, θ=0.2, kmax=8, ρ=0.5,and β=0* | *r=3, θ=0.2, kmax=8, ρ=0, and β=0.2* | *r=3, θ=0.2, kmax=8, ρ=0.3, and β=0.1*

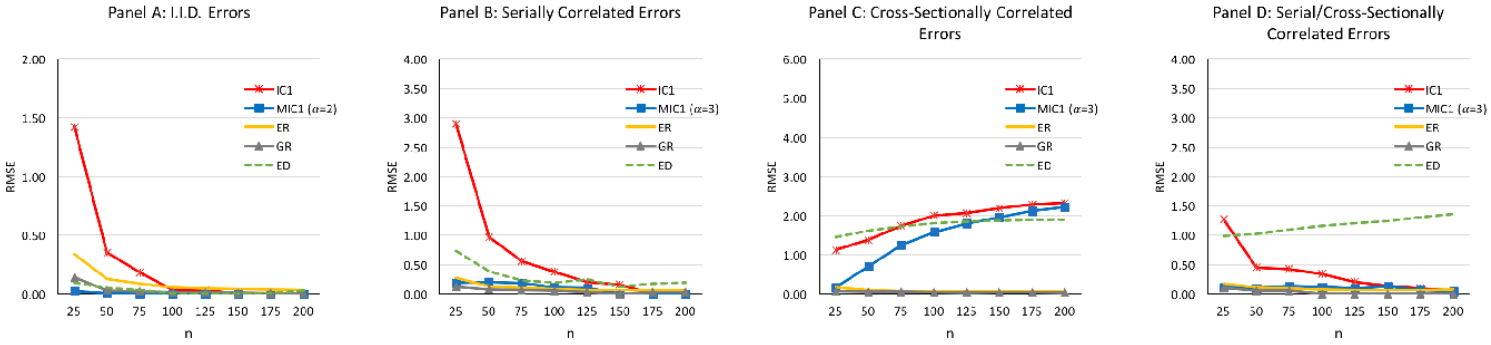
$(2)\ \theta = 0.2,\ p > n = 25$



r=3, θ=0.2, kmax=8, and ρ= β=0 | *r=3, θ=0.2, kmax=8, ρ=0.5,and β=0* | *r=3, θ=0.2, kmax=8, ρ=0, and β=0.2* | *r=3, θ=0.2, kmax=8, ρ=0.3, and β=0.1*

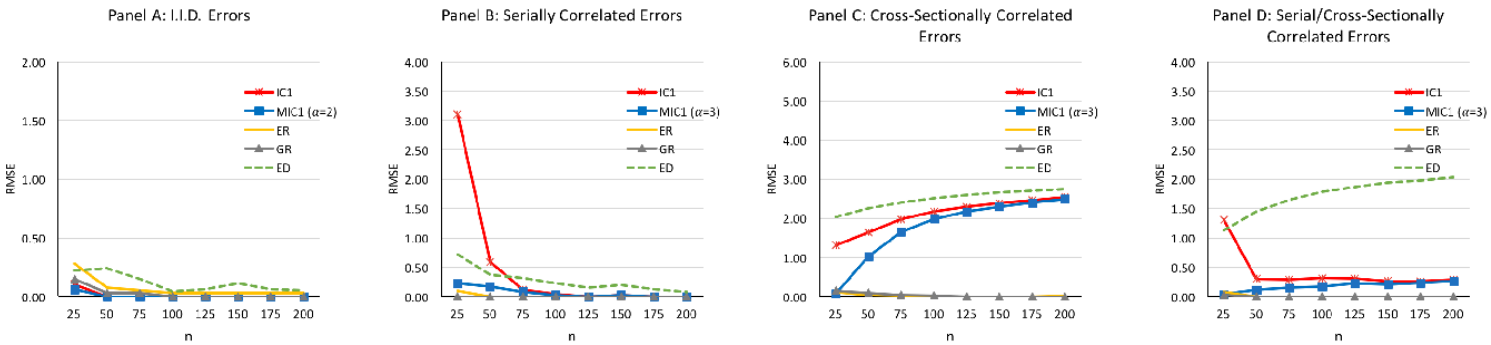
Figure 2.7: Effects of Error Covariance Structure (Three-factor Model, $\theta = 0.2,\ n > p$)

$(1)\ \theta = 0.2,\ n > p = 15$



r=3, θ=0.2, kmax=8, and ρ= β=0 | *r=3, θ=0.2, kmax=8, ρ=0.5,and β=0* | *r=3, θ=0.2, kmax=8, ρ=0, and β=0.2* | *r=3, θ=0.2, kmax=8, ρ=0.3, and β=0.1*

$(2)\ \theta = 0.2,\ n > p = 25$



r=3, θ=0.2, kmax=8, and ρ= β=0 | *r=3, θ=0.2, kmax=8, ρ=0.5,and β=0* | *r=3, θ=0.2, kmax=8, ρ=0, and β=0.2* | *r=3, θ=0.2, kmax=8, ρ=0.3, and β=0.1*

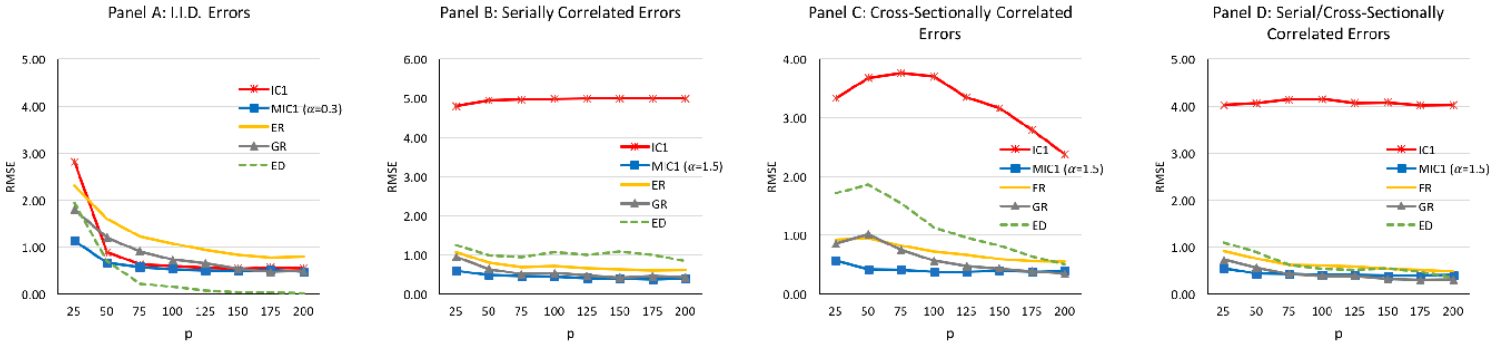
Figure 2.8 and 2.9 report the results from the second part of simulations. Here three factors ($r = 3$) are drawn from $\mathcal{N}(0, 1)$, and θ is fixed at 1 so that we consider lower SNRs equal to 1. As depicted in Figure 2.8 for the cases with $p > n$, the MIC_1 ($\alpha = 1.5$) estimator clearly outperforms the IC_1 and ED estimators when the idiosyncratic components are weakly correlated (Panels B, C and D). Comparing these cases to the case with i.i.d. errors (Panel A), we can see that correlation in the idiosyncratic terms substantially worsens the quality of the IC_1 estimator. The accuracy of the MIC_1 estimator remains very good, however. It is not much affected by the covariance structure of errors. Also, for each n the MIC_1 estimator generally performs equally to or better than the ER and GR estimators.

Further, the cases with $n > p$ are shown in Figure 2.9. We set $\alpha = 1$ for the case with i.i.d. errors (Panel A) while $\alpha = 3$ for the case with correlated errors (Panels B, C and D). For each p, the performance of MIC_1 estimator is comparable to, if not better than, those of the ED, ER and GR estimators. The only exception is the case with cross-sectionally correlated errors (Panel C) in which the MIC_1 estimator selects the number of factors with less precision and its RMSE gets larger as n increases. Like the cases with strong factors (Figure 2.7), it appears that the performances of the MIC_1, IC_1 and ED estimators are more sensitive to cross-sectional correlation than serial correlation. Even for this case, RMSEs of the MIC_1 are much lower than those of IC_1 when the sample size remains small.

Lastly, we consider an additional experiment where the temporal dimension of the data is comparable to their cross-sectional size; particularly, $n = p \in \{50, 75, 100, 125, 150, 175, 200\}$ (Figure 2.10). The results remain the same as those of previous experiments, regardless of values of $\theta \in \{1, 0.2\}$. Except for the case with cross-sectionally correlation (Panel C), the MIC_1 estimator clearly outperforms the IC_1 and ED estimators while its performance being comparable to, if not better than, those of the ER and GR estimators.

Figure 2.8: Effects of Error Covariance Structure (Three-factor Model, $\theta = 1,\ p > n$)

$(1)\ \theta = 1,\ p > n = 15$



$(2)\ \theta = 1,\ p > n = 25$

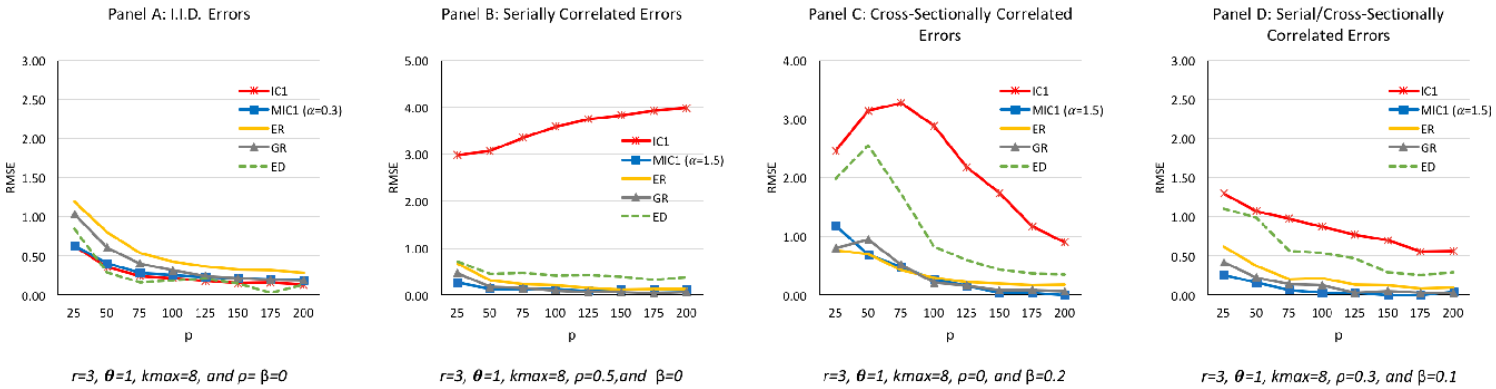
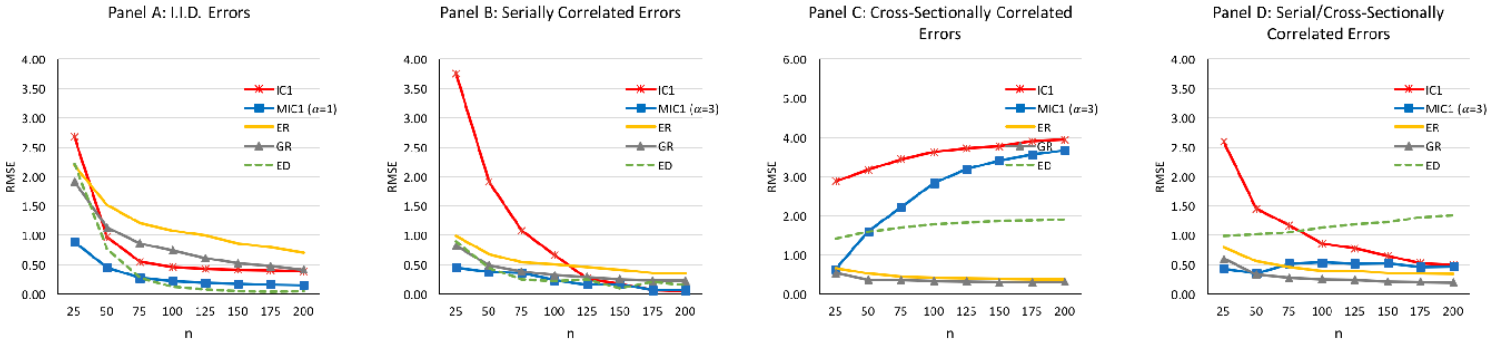
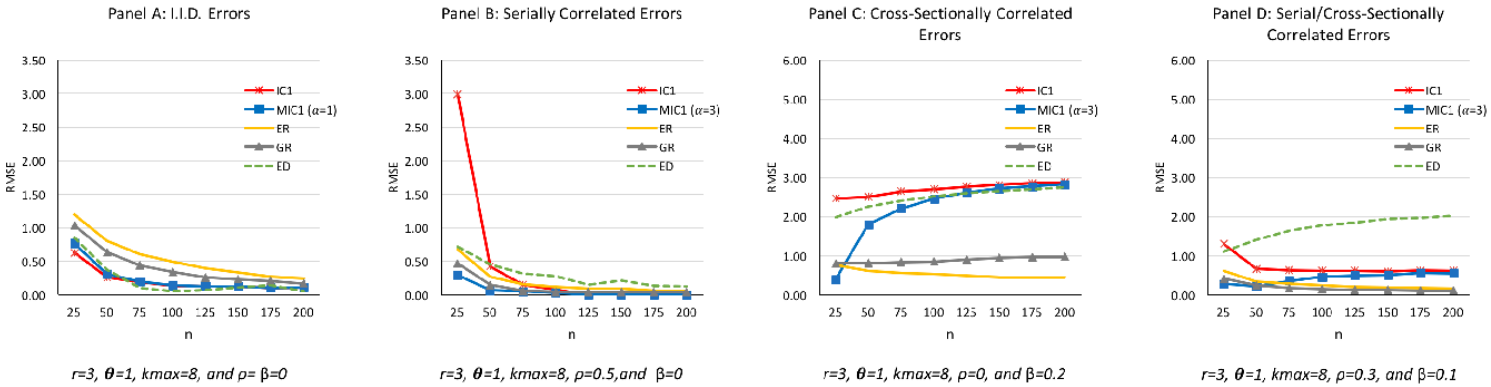


Figure 2.9: Effects of Error Covariance Structure (Three-factor Model, $\theta = 1,\ n > p$)

$(1)\ \theta = 1,\ n > p = 15$



$(2)\ \theta = 1,\ n > p = 25$



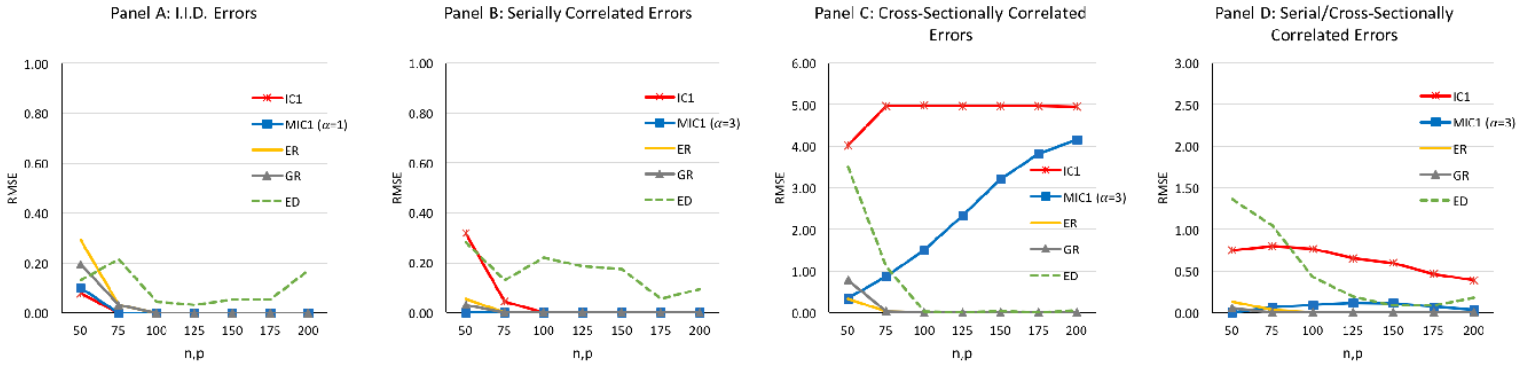
Obviously, when we use IC_2 and MIC_2, instead of IC_1 and MIC_1, for our simulations, we can obtain more precise estimates from both criteria; however, the main results remain the same and thus are not reported here. In particular, IC_2 still overestimates the number of factors when the sample size is small, and MIC_2 clearly outperforms IC_2 even when errors are weakly correlated. Further, the performance of MIC_2 gets closer to, if not better than, those of the ER and GR estimators, and it becomes less sensitive to the cross-sectional correlation of the error terms.

The Monte Carlo experiments from Bai and Ng (2002), Onatski (2010), and Ahn and Horenstein (2013) did not report the simulation results for the case with sufficiently strong factors. Rather, Bai and Ng (2002) especially noted that the IC estimator is expected to yield precise estimates of the true number of factors in such a case; however, our simulations show that the IC estimator does not perform well even for the case with strong factors when the sample size is small. Further, we see that weakly correlated errors significantly reduce the precision of the estimates. Overall, the results from our simulations show that our proposed criteria, MIC, improve the finite sample performance of the original IC estimator even for the weakly serially or/and cross-sectionally correlated error components, regardless of the relative size of n and p. Moreover, by adjusting α for the relative strength of signals to noise, the MIC estimator can yield comparable performance to, if not better than, those of the ED, ER and GR estimators, unless the idiosyncratic components are only cross-sectionally correlated with large population size.

Since this chapter focuses on the over-detection risk of IC, the main goal of the proposed MIC estimator is to reduce the upper probability bound of over-detection. But also, MIC is likely to worsen the under-detection of the number of factors at the same time because our modification leads to a higher penalty for over-fitting than the original criteria. When we consider simulation results showing a significant improvement in the overall performance of MIC, however, we can conjecture that deteriorating under-detection risk might be dominated by decreasing over-detection risk.

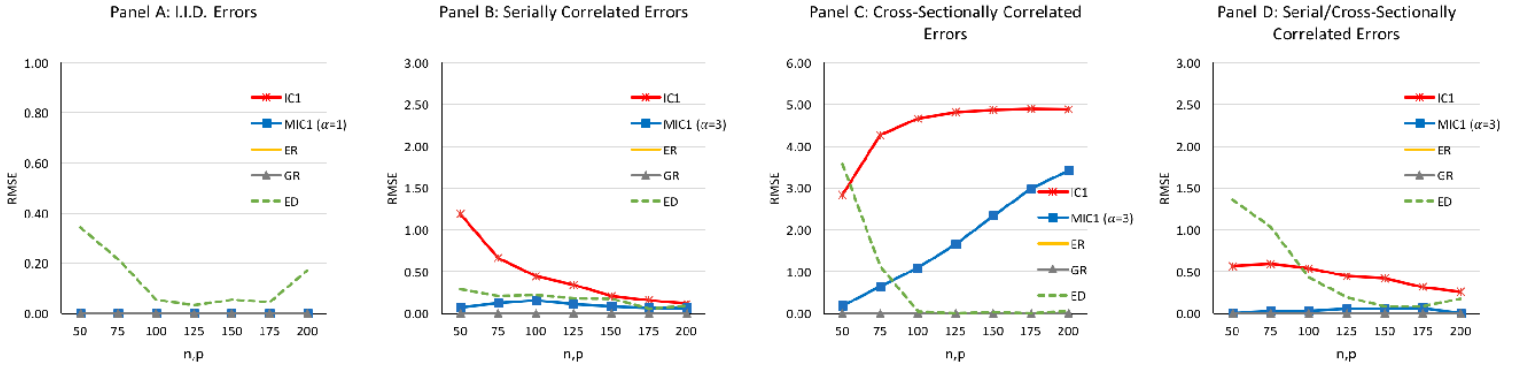
Figure 2.10: Effects of Error Covariance Structure (Three-factor Model, $n = p$)

(1) $\theta = 1$, $n = p$



r=3, θ=1, kmax=8, and ρ= β=0 | *r=3, θ=1, kmax=8, ρ=0.5,and β=0* | *r=3, θ=1, kmax=8, ρ=0, and β=0.2* | *r=3, θ=1, kmax=8, ρ=0.3, and β=0.1*

(2) $\theta = 0.2$, $n = p$



r=3, θ=0.2, kmax=8, and ρ= β=0 | *r=3, θ=0.2, kmax=8, ρ=0.5,and β=0* | *r=3, θ=0.2, kmax=8, ρ=0, and β=0.2* | *r=3, θ=0.2, kmax=8, ρ=0.3, and β=0.1*

2.8 Conclusion

The detection of the number of factors is a prerequisite for factor analysis. This chapter studies the detection performance of the *IC* estimator proposed by Bai and Ng (2002). First, we derive the computable formula for a theoretical upper bound on the over-detection probability of the *IC* estimator. More specifically, we pin down the expression for the overestimation condition of *IC* in terms of pure noise eigenvalues, and then we analyze a non-asymptotic bound on the overestimation probability by employing the results on the limiting behavior of the largest pure noise eigenvalue from random matrix theory.

Next, using this computable formula, we analyze the detection performance of the *IC* estimator. We compute overestimation probability bounds subject to various sample sizes and numbers of factors, and the choice of a penalty function. These numerical examples show that the *IC* estimator often overestimates the number of factors for the case with small sample sizes even when factors have strong explanatory power. Accordingly, this chapter provides a theoretical prediction for the overestimation probability of the *IC* estimator; specifically, users may use our computable formula as a diagnostic tool for misspecification.

Moreover, we show that the improved penalty for overfitting by degrees of freedom adjustment can reduce the overestimation probability of the *IC* estimator substantially in small samples. As a consequence, we propose a modified estimator, *MIC*, as a practical guide to improving the finite sample performance. Our performance analysis using the computable formula for the overestimation probability bound demonstrates that our *MIC* estimator improves the accuracy of the original *IC* estimator for the case with Gaussian i.i.d. errors. In addition, via Monte Carlo simulations, we show that the *MIC* estimator outperforms the *IC* estimator even for the case with the weakly serially or/and cross-sectionally correlated error terms. Furthermore, comparing the *MIC* estimator and other leading estimators such as the *ER* and *GR* estimators of Ahn and Horenstein (2013), and the *ED* estimator of Onatski (2010), we see that the *MIC* estimator generally performs well unless the error components are only cross-sectionally correlated.

Several interesting extensions are left for future research. One of them is to generalize our theoretical upper bound on the overestimation probability of the IC estimator to the cases with the more general covariance structure of errors. We have briefly sketched some ideas in this chapter, but it remains to be studied further. Another interesting topic is to study the large r asymptotics of the IC estimator in which the true number of factors can increase with the sample size and to examine its misspecification risk. Moreover, our analysis might be extended to general model selection criteria for factor models. For example, Choi and Jeong (2013) derived several criteria for large factor models based on the AIC and the BIC. So far as any criterion is represented by pure noise eigenvalues, our method might be applied.

Lastly, this chapter focused on the analysis of the over-detection risk based on random matrix theory. In a similar fashion, we will examine the overall misdetection risk of the IC estimator by extending our analysis to the case with under-detected factors and eventually discuss the optimal rule for detecting the number of factors in the second chapter.

Appendix

A.2.1. Proof of Lemma 2.1

Proof. Let us assume that the number of factors are known as k. f_t and λ_i can be estimated by the principal components method under the normalization of $\frac{\widetilde{\Lambda}'\widetilde{\Lambda}}{p} = I_k$ (for details, see Bai and Ng, 2002). That is, the principal components estimator $\widetilde{\Lambda} = \sqrt{p}B_n$, where B_n is the $p \times k$ matrix composed of the eigenvectors corresponding to k eigenvalues of S_n. And given $\widetilde{\Lambda}$, we get $\widetilde{f}_t = (\widetilde{\Lambda}'\widetilde{\Lambda})^{-1}\widetilde{\Lambda}'x_t$. Then, from (2.3.2),

$$\begin{aligned}
S(k) &= \frac{1}{pn}\sum_{t=1}^{n}(x_t - \widetilde{\Lambda}\widetilde{f}_t)'(x_t - \widetilde{\Lambda}\widetilde{f}_t) \\
&= \frac{1}{pn}\sum_{t=1}^{n} x_t'(I_p - P_{\tilde{\Lambda}})x_t \\
&= \frac{1}{p}tr\left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right) - \frac{1}{p^2}tr\left(\tilde{\Lambda}'\left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right)\tilde{\Lambda}\right) \\
&= \frac{1}{p}tr\left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right) - \frac{1}{p}tr\left(B_n'\left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right)B_n\right) \\
&= \frac{1}{p}\sum_{j=1}^{p}\ell_j - \frac{1}{p}\sum_{j=1}^{k}\ell_j \\
&= \frac{1}{p}\sum_{j=k+1}^{p}\ell_j,
\end{aligned}$$

where $P_{\tilde{\Lambda}} = \widetilde{\Lambda}(\widetilde{\Lambda}'\widetilde{\Lambda})^{-1}\widetilde{\Lambda}'$. □

A.2.2. Proof of Lemma 2.3

Proof. See Nadler (2010) for the proof of (2.4.2). Here, we prove (2.4.3).

Recall $\rho_{ii} = b_i'\left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right)b_i$, $\widetilde{\ell}_j = \tilde{d}_j'\left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right)\tilde{d}_j$, and $\eta_{ij} = \tilde{d}_j'\left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right)b_i$. By Remark 2.1 and 2.2, we can see that $\rho_{ii} = (\psi_i + 1)\left(1 + O_p\left(1/\sqrt{n}\right)\right)$ and $\widetilde{\ell}_j = 1 + O_p\left(1/\sqrt{n}\right)$. Also, we can write $\eta_{ij} = (\rho_{ii}\widetilde{\ell}_j)^{1/2}\frac{1}{n}\sum_{t=1}^{n}\alpha_t\beta_t$, where $\alpha_t = (b_i'x_t)/(\rho_{ii}^{1/2})$ and

$\beta_t = (x_t' \tilde{d}_j)/(\widetilde{\ell}_j^{1/2})$. Further, let us define $\kappa_{ij} = \frac{\sqrt{n}\,\eta_{ij}}{(\rho_{ii}\widetilde{\ell}_j)^{1/2}} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \alpha_t \beta_t$. Then, we get

$$\frac{1}{n} \sum_{i=1}^{r} \frac{(\sqrt{n}\,\eta_{ij})^2}{\rho_{ii} - \widetilde{\ell}_j} = \frac{1}{n} \sum_{i=1}^{r} \left(\frac{\rho_{ii}\widetilde{\ell}_j}{\rho_{ii} - \widetilde{\ell}_j} \right) \kappa_{ij}^2.$$

Note that α_t and β_t are independent of each other due to the orthogonality between b_i and $\tilde{d}_j$. And $E(\alpha_t) = 0$, $E(|\alpha_t|^2) = 1$, and $E(\alpha_t \alpha_s) = 0$ for $t \neq s$ since b_i, the i-th eigenvector of Σ, is fixed and independent of signals and noise random realizations. Similarly, $E(\beta_t) = 0$, $E(|\beta_t|^2) = 1$, and $E(\beta_t \beta_s) = 0$ for $t \neq s$. Also, by definition, $\frac{1}{n} \sum_{t=1}^{n} \alpha_t \beta_t$ is the sample correlation coefficient between the projection of the data onto a fixed direction b_i and its projection onto the orthogonal direction $\tilde{d}_j$. Thus, assuming i.i.d. Gaussian errors and factors, as $n \to \infty$, $\kappa_{ij} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \alpha_t \beta_t$ has the limiting distribution $\mathcal{N}(0,1)$ (see Anderson, 2003, Theorem 4.2.4)[1]. Hence, $\kappa_{ij} = O_p(1)$.

Consequently, for $\psi_i = O(1)$, as $n \to \infty$ we have

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{r} \frac{(\sqrt{n}\,\eta_{ij})^2}{\rho_{ii} - \widetilde{\ell}_j} &= \widetilde{\ell}_j \left(\frac{1}{n} \sum_{i=1}^{r} \left(\frac{\psi_i + 1}{\psi_i} + O_p\left(\frac{1}{\sqrt{n}} \right) \right) \kappa_{ij}^2 \right) \\
&= \widetilde{\ell}_j \left(\frac{1}{n} \sum_{i=1}^{r} \frac{\psi_i + 1}{\psi_i} + \frac{\sqrt{r}}{n} \frac{1}{\sqrt{r}} \sum_{i=1}^{r} \frac{\psi_i + 1}{\psi_i} \left(\kappa_{ij}^2 - 1 \right) \right) + O_p\left(\frac{1}{n^{3/2}} \right)
\end{aligned}$$

because $\frac{\rho_{ii}}{\rho_{ii} - \widetilde{\ell}_j} = \frac{(\psi_i + 1)\left(1 + O_p\left(1/\sqrt{n}\right)\right)}{\psi_i \left(1 + O_p\left(1/\sqrt{n}\right)\right)}$. □

[1] If $corr(n)$ is the sample correlation coefficient of a sample of n from a normal distribution with correlation ρ, then $\frac{\sqrt{n}(corr(n) - \rho)}{1 - \rho^2}$ has the limiting distribution $\mathcal{N}(0,1)$.

A.2.3. Proof of Theorem 2.1

Proof. From Lemma 2.3, we get (2.4.3) and (2.4.4). Now, if we insert them into (2.3.8), then we get

$$\begin{aligned}\triangle IC(1) &= \ln T_{p-r} - \ln T_{p-r-1} - G(p,n) \\ &= \ln \widetilde{T}_{p-r} - \ln \widetilde{T}_{p-r-1} - G(p,n) + \ln\left(1 - \frac{M_r}{n} - \frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r}}\sum_{j=r+1}^{p}\widetilde{\ell}_j Z_j\right) \\ &\quad - \ln\left(1 - \frac{M_r}{n} - \frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r-1}}\sum_{j=r+2}^{p}\widetilde{\ell}_j Z_j\right). \end{aligned} \tag{2.8.1}$$

First, $\frac{M(r)}{n}$ is negligible. Note that $M_r = O(r)$ since $M_r = \sum_{i=1}^{r}\frac{\psi_i+1}{\psi_i}$ and $\psi_i = O(1)$. Next, we show that $Z_j = O_p(1)$. As shown in the proof of Lemma 2.3, as $n \to \infty$, $\kappa_{ij} = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\alpha_t\beta_t$ has the limiting distribution $\mathcal{N}(0,1)$; hence, $\kappa_{ij} = O_p(1)$. Also, since $\kappa_{ij}^2 \sim \chi^2(1)$, $Var(\kappa_{ij}^2) = E\left((|\kappa_{ij}|^2-1)^2\right) = O(1)$. In Lemma 2.3, the zero mean random variable $Z_j = \frac{1}{\sqrt{r}}\sum_{i=1}^{r}\frac{\psi_i+1}{\psi_i}(\kappa_{ij}^2-1)$. Recall that vectors in $\widetilde{B} = (b_1,\ldots,b_r,\tilde{d}_{r+1},\ldots,\tilde{d}_p)$ are linearly independent of each other. Furthermore, from $\psi_i = O(1)$ and $E\left((|\kappa_{ij}|^2-1)^2\right) = O(1)$, we get

$$\begin{aligned} E\left|\frac{1}{\sqrt{r}}\sum_{i=1}^{r}\frac{\psi_i+1}{\psi_i}(\kappa_{ij}^2-1)\right|^2 &= \frac{1}{r}\sum_{i=1}^{r}\sum_{h=1}^{r}\frac{\psi_i+1}{\psi_i}\frac{\psi_h+1}{\psi_h}E\left((|\kappa_{ij}|^2-1)^2\right) \\ &\le \left(\frac{\psi_1+1}{\psi_1}\right)^2\frac{1}{r}\sum_{i=1}^{r}E\left((|\kappa_{ij}|^2-1)^2\right) = O(1), \end{aligned}$$

i.e., $E(|Z_j|^2) = O(1)$ so that $Z_j = O_p(1)$ (see Jiang, 2010, Theorem 3.1).

Now, we approximate (C.1) by the Taylor expansion. Note that $\frac{1}{p-r}\widetilde{T}_{p-r} = O_p(1)$, $\frac{1}{p-r}\widetilde{\ell}_{p-r} = O_p(\frac{1}{p-r})$, and $\frac{1}{p-r}\sum_{j=r+1}^{p}\widetilde{\ell}_j Z_j = \sum_{j=r+1}^{p}O_p(\frac{1}{p-r})O_p(1) = O_p(1)$ by Remark 2.2, and Jiang (2010, Lemma 3.12). Then, $\frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r}}\sum_{j=r+1}^{p}\widetilde{\ell}_j Z_j$ is sufficiently small for large n as

follows:

$$\frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r}}\sum_{j=r+1}^{p}\widetilde{\ell}_j Z_j = \frac{\sqrt{r}}{n}\left(\frac{p-r}{\widetilde{T}_{p-r}}\right)\frac{1}{p-r}\sum_{j=r+1}^{p}\widetilde{\ell}_j Z_j = \frac{\sqrt{r}}{n}O_p(1)O_p(1) = O_p\left(\frac{\sqrt{r}}{n}\right),$$

and similarly $\frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r-1}}\sum_{j=r+2}^{p}\widetilde{\ell}_j Z_j = O_p\left(\frac{\sqrt{r}}{n}\right)$. Therefore, by the Taylor expansion we can get

$$\begin{aligned}
&\ln\left(1-\frac{M_r}{n}-\frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r}}\sum_{j=r+1}^{p}\widetilde{\ell}_j Z_j\right) - \ln\left(1-\frac{M_r}{n}-\frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r-1}}\sum_{j=r+2}^{p}\widetilde{\ell}_j Z_j\right) \\
&\quad\approx -\frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r}}\sum_{j=r+1}^{p}\widetilde{\ell}_j Z_j + \frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r-1}}\sum_{j=r+2}^{p}\widetilde{\ell}_j Z_j. \qquad (2.8.2)
\end{aligned}$$

Finally, using (C.2), (C.1) can be rewritten as

$$\triangle IC(1) \approx \ln\widetilde{T}_{p-r} - \ln\widetilde{T}_{p-r-1} - \frac{\sqrt{r}}{n}Z - G(p,n), \qquad (2.8.3)$$

where

$$\frac{\sqrt{r}}{n}Z = \frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r}}\left(\widetilde{\ell}_{r+1}Z_{r+1}\right) - \frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r}}\frac{\widetilde{\ell}_{r+1}}{\widetilde{T}_{p-r-1}}\left(\sum_{j=r+2}^{p}\widetilde{\ell}_j Z_j\right) = I + II.$$

However, $\frac{\sqrt{r}}{n}Z$ is asymptotically negligible with respect to $G(p,n)$. More precisely, for I term,

$$\begin{aligned}
\frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r}}\left(\widetilde{\ell}_{r+1}Z_{r+1}\right) &= \frac{\sqrt{r}}{n}\left(\frac{p-r}{\widetilde{T}_{p-r}}\right)\left(\frac{\widetilde{\ell}_{r+1}}{p-r}\right)Z_{r+1} \\
&= \frac{\sqrt{r}}{n}O_p(1)O_p\left(\frac{1}{p-r}\right)O_p(1) = O_p\left(\frac{\sqrt{r}}{n(p-r)}\right),
\end{aligned}$$

and for II term,

$$\frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r}}\frac{\widetilde{\ell}_{r+1}}{\widetilde{T}_{p-r-1}}\left(\sum_{j=r+2}^{p}\widetilde{\ell}_j Z_j\right) = \frac{\sqrt{r}}{n}\left(\frac{p-r}{\widetilde{T}_{p-r}}\right)\left(\frac{\widetilde{\ell}_{r+1}}{p-r}\right)\left(\frac{p-r-1}{\widetilde{T}_{p-r-1}}\right)\frac{\sum_{j=k+2}^{p}\widetilde{\ell}_j Z_j}{p-r-1}$$
$$= \frac{\sqrt{r}}{n}O_p(1)O_p\left(\frac{1}{p-r}\right)O_p(1)O_p(1) = O_p\left(\frac{\sqrt{r}}{n(p-r)}\right),$$

while $G_3(p,n) = O\left(\ln p/p\right)$, for example. Thus, up to an $o_p\left(\frac{1}{n}\right)$ error term, (2.3.8) can be approximately equivalent to

$$\triangle\, IC(1) \approx \ln \widetilde{T}_{p-r} - \ln \widetilde{T}_{p-r-1} - G(p,n). \tag{2.8.4}$$

As a result, the approximate expression of $\triangle IC(1)$ does not include any signal contribution or interaction between signals and noise.

Now, let us define $\xi = \frac{\widetilde{\ell}_{r+1}}{\widetilde{T}_{p-r}}$ so that $\xi < 1$. Then, (C.4) can be written as

$$\begin{aligned}\triangle IC(1) &\approx \ln \widetilde{T}_{p-r} - \ln\left(\widetilde{T}_{p-r} - \widetilde{\ell}_{r+1}\right) - G(p,n)\\ &= \ln \widetilde{\ell}_{r+1} - \ln \xi - \ln\left(\frac{\widetilde{\ell}_{r+1}}{\xi} - \widetilde{\ell}_{r+1}\right) - G(p,n)\\ &= \ln \widetilde{\ell}_{r+1} - \ln \xi - \ln \widetilde{\ell}_{r+1} - \ln\frac{1-\xi}{\xi} - G(p,n)\\ &= -\ln(1-\xi) - G(p,n).\end{aligned} \tag{2.8.5}$$

Also, because $\xi < 1$, we can say

$$\triangle\, IC(1) \approx \xi + \frac{\xi^2}{2} - G(p,n) + o(\xi^3) \tag{2.8.6}$$

by the Taylor approximation.

Finally, the number of factors is overestimated by exactly one factor if $\triangle IC(1) > 0$ in

(C.6). Let $\xi_{n,p-r}$ denote a solution for

$$\xi + \frac{1}{2}\xi^2 - G(p,n) = 0;$$

that is, $\xi_{n,p-r} = -1 + \sqrt{1 + 2G(p,n)}$. As a result, if $\widetilde{\ell}_{r+1} > \widetilde{T}_{p-r} \cdot \xi_{n,p-r}$, then $\triangle IC(1) > 0$ because $\widetilde{\ell}_{r+1} = \widetilde{T}_{p-r} \cdot \xi$.

Note that $\widetilde{\ell}_{r+1}$, $\widetilde{T}_{p-r}$, and $\xi_{n,p-r}$ are the same as $\ell_1(W)$, $Tr(W)$, and $\xi_{n,p}$, respectively. □

A.2.4. Proof of Theorem 2.2

Part 1.

Proof. Consider the average of the sample eigenvalues of a $(p-r) \times (p-r)$ Wishart matrix, W. By Remark 2.1, $\frac{Tr(W)}{p-r} = \frac{\sum_{j=r+1}^{p} \widetilde{\ell}_j}{p-r} \sim \frac{\chi^2_{n(p-r)}}{n(p-r)}$. Let s be some positive number. Then we can write

$$\begin{aligned}\Pr(\triangle IC > 0) = \Pr\left(\triangle IC > 0 \bigcap \frac{Tr(W)}{p-r} \leq 1 - \frac{s}{\sqrt{n}}\right) \\ + \Pr\left(\triangle IC > 0 \bigcap \frac{Tr(W)}{p-r} > 1 - \frac{s}{\sqrt{n}}\right).\end{aligned}$$

Also, by Theorem 2.1, we obtain the following inequality:

$$\begin{aligned}\Pr(\triangle IC > 0) &\leq \Pr\left(\frac{\chi^2_{n(p-r)}}{n(p-r)} \leq 1 - \frac{s}{\sqrt{n}}\right) + \Pr\left(\frac{\ell_1(W)}{Tr(W)} > \xi_{n,p} \bigcap \frac{Tr(W)}{p-r} > 1 - \frac{s}{\sqrt{n}}\right) \\ &\leq \Pr\left(\frac{\chi^2_{n(p-r)}}{n(p-r)} \leq 1 - \frac{s}{\sqrt{n}}\right) + \Pr\left(\ell_1(W) > (p-r)\left(1 - \frac{s}{\sqrt{n}}\right)\xi_{n,p}\right) \\ &= I + II.\end{aligned}$$

□

Part 2.

Proof. Using the following lemma regarding a Chi-squared inequality (Johnstone and Lu, 2009, Appendix, A.2), the upper bound of I in part 1 can be obtained as follows.

Lemma 2.4. *(Johnstone and Lu, 2009)*

$$\Pr\left(\chi_v^2 \leq v(1-\epsilon)\right) \leq \exp\left(\frac{-v\epsilon^2}{4}\right), \qquad 0 \leq \epsilon < 1.$$

Thus, setting $v = n(p-r)$ and $\epsilon = \frac{s}{\sqrt{n}}$, we get

$$I = \Pr\left(\frac{\chi^2_{n(p-r)}}{n(p-r)} \leq 1 - \frac{s}{\sqrt{n}}\right) = \Pr\left(\chi^2_{n(p-r)} \leq n(p-r)\left(1 - \frac{s}{\sqrt{n}}\right)\right) \leq \exp\left(\frac{-(p-r)s^2}{4}\right).$$

□

Part 3.

Proof. Now, let us derive the upper bound of II in part 1. As already seen in Section 2.5 (2.5.8), Ledoux (2007)'s result can be applied to our model as follows.

Lemma 2.5. *By Ledoux (2007, Proposition 2.2), we get*

$$\Pr\left(\ell_1(W) \geq (1+\sqrt{\bar{c}})^2 + \varepsilon\right) \leq \exp\left(-nJ_{_{LAG}}(\varepsilon)\right), \tag{2.8.7}$$

where

$$J_{_{LAG}}(\varepsilon) = \int_1^x (x-y)\frac{(1+\bar{c})y + 2\sqrt{\bar{c}}}{(y+B)^2}\frac{dy}{\sqrt{y^2-1}}, \tag{2.8.8}$$

with $\bar{c} = \frac{p-r}{n}$, $x = 1 + \frac{\varepsilon}{2\sqrt{\bar{c}}}$, *and* $B = \frac{1+\bar{c}}{2\sqrt{\bar{c}}}$. *Then, by setting* $\varepsilon = (p-r)\left(1 - \frac{s}{\sqrt{n}}\right)\xi_{np} - \left(1+\sqrt{\bar{c}}\right)^2$, *the following inequality should hold:*

$$II = \Pr\left(\ell_1(W) > (p-r)\left(1 - \frac{s}{\sqrt{n}}\right)\xi_{n,p}\right) \leq \exp\left(-nJ_{_{LAG}}(\varepsilon)\right). \tag{2.8.9}$$

Now, let us derive the explicit expression of (D.3).

$$\begin{aligned} n \cdot J_{LAG}(\varepsilon) &= n \int_1^x (x-y) \frac{(1+\bar{c})y + 2\sqrt{\bar{c}}}{(y+B)^2} \frac{dy}{\sqrt{y^2-1}} \\ &= n \int_1^x \frac{(x-y)}{\sqrt{y^2-1}} \frac{(1+\bar{c})y + 2\sqrt{\bar{c}}}{\left(2y\sqrt{\bar{c}} + \bar{c} + 1\right)^2 (1/4\bar{c})} dy \\ &= 4(p-r) \int_1^x \frac{y(x-y)}{\sqrt{y^2-1}} \underbrace{\left(\frac{(1+\bar{c})y + 2\sqrt{\bar{c}}}{y(2y\sqrt{\bar{c}} + \bar{c} + 1)^2} \right)}_{III} \end{aligned}$$

and

$$\begin{aligned} III &= \frac{(1+\bar{c})\, y + 2\sqrt{\bar{c}}}{(1+\bar{c})^2 y + 4\sqrt{\bar{c}}(1+\bar{c})y^2 + 4\bar{c}y^3} \\ &= 1 - \frac{\bar{c}(1+\bar{c})y + 4\sqrt{\bar{c}}(1+\bar{c})y^2 + 4\bar{c}y^3 - 2\sqrt{\bar{c}}}{(1+\bar{c})^2 y + 4\sqrt{\bar{c}}(1+\bar{c})y^2 + 4\bar{c}y^3} \\ &= 1 - \sqrt{\bar{c}} \underbrace{\left(\frac{\sqrt{\bar{c}}(1+\bar{c})y + 4(1+\bar{c})y^2 + 4\sqrt{\bar{c}}y^3 - 2}{(1+\bar{c})^2 y + 4\sqrt{\bar{c}}(1+\bar{c})y^2 + 4\bar{c}y^3} \right)}_{IV}. \end{aligned}$$

By Jiang (2010, p. 54)'s Lemma 3.6 and Example 3.3,

$$IV = \frac{y(\bar{c})^{3/2} + 4y^2\bar{c} + (4y^3+y)(\bar{c})^{1/2} + 4y^2 - 2}{y(\bar{c})^2 + 4y^2(\bar{c})^{3/2} + (4y^3+2y)\bar{c} + 4y^2(\bar{c})^{1/2} + y} = O(1)$$

as $\bar{c} \to \infty$, while $IV = o(1)$ as $\bar{c} \to 0$. Thus, especially for large n, we get

$$n \cdot J_{LAG} = 4(p-k) \int_1^x \frac{y(x-y)}{\sqrt{y^2-1}} \left(1 + O\left(\sqrt{\frac{p-r}{n}} \right) \right) dy \tag{2.8.10}$$

and

$$\begin{aligned}\int_1^x \frac{y(x-y)}{\sqrt{y^2-1}}dy &= x\int_1^x \frac{y}{\sqrt{y^2-1}}dy - \int_1^x \frac{y^2}{\sqrt{y^2-1}}dy \\ &= \frac{1}{2}\left(1+\frac{\varepsilon}{2\sqrt{\bar{c}}}\right)\sqrt{\frac{\varepsilon^2}{4\bar{c}}+\frac{\varepsilon}{\sqrt{\bar{c}}}} - \frac{1}{2}\ln\left(1+\frac{\varepsilon}{2\sqrt{\bar{c}}}+\sqrt{\frac{\varepsilon^2}{4\bar{c}}+\frac{\varepsilon}{\sqrt{\bar{c}}}}\right) \\ &= \frac{1}{4}\left(1+\frac{\delta}{2}\right)\sqrt{\delta(\delta+4)} - \frac{1}{2}\underbrace{\ln\left(1+\frac{\delta}{2}+\frac{1}{2}\sqrt{\delta(\delta+4)}\right)}_{V},\end{aligned}$$

where $\varepsilon = \delta\sqrt{\bar{c}} = \delta\sqrt{\frac{p-r}{n}}$. Furthermore, let us define $q = \frac{\delta}{2}+\frac{1}{2}\sqrt{\delta(\delta+4)}$. Then, by the Taylor expansion,

$$V = \ln(1+q) = q - \frac{1}{2}q^2 + \frac{1}{3}q^3 - \frac{1}{4}q^4 + \frac{1}{5}q^5\cdots.$$

From this expansion, we get the following inequality; that is, for $q \geq 0$,

$$\ln(1+q) \leq q - \frac{1}{2}q^2 + \frac{1}{3}q^3. \tag{2.8.11}$$

These inequalities are quite intuitive. Let $g(q) = \ln(1+q) - q + \frac{1}{2}q^2 - \frac{1}{3}q^3$. Then $g(q)$ is a non-increasing function because $g(q)' = \frac{1}{1+q} - 1 + q - q^2 = -\frac{q^3}{1+q} \leq 0$, for all $q \geq 0$. Thus, $g(q) \leq g(0)$ so that $\ln(1+q) \leq q - \frac{q^2}{2} + \frac{q^3}{3}$.

As seen in (2.5.7) and (2.5.8), the non-asymptotic bound of the largest eigenvalue of W can be also defined as follows:

$$\Pr\left(\ell_1(W) \geq (1+\sqrt{\bar{c}})^2 + \varepsilon\right) \leq M\exp\left(-n\min\{\varepsilon, \varepsilon^{3/2}\}/M\right),$$

and also

$$\Pr\left(\ell_1(W) \geq (1+\sqrt{\bar{c}})^2 + \varepsilon\right) \leq \exp\left(-nJ_{_{LAG}}(\varepsilon)\right).$$

We can check that $J_{LAG(\varepsilon)} \geq \varepsilon^{3/2}/M$. From (D.5), we get

$$\begin{aligned} V &= \ln\left(1+\frac{\delta}{2}+\frac{1}{2}\sqrt{\delta(\delta+4)}\right) \\ &\leq \frac{\delta}{2}+\frac{1}{2}\sqrt{\delta(\delta+4)}-\frac{1}{2}\left(\frac{\delta}{2}+\frac{1}{2}\sqrt{\delta(\delta+4)}\right)^2+\frac{1}{3}\left(\frac{\delta}{2}+\frac{1}{2}\sqrt{\delta(\delta+4)}\right)^3. \end{aligned}$$

When ε is a sufficiently small, $\delta < 1$ can be a reasonable restriction because $\varepsilon = \delta\sqrt{\bar{c}}$. Therefore, we can obtain the following inequality:

$$\begin{aligned} \int_1^x \frac{y(x-y)}{\sqrt{y^2-1}}dy &\geq \frac{1}{4}\left(1+\frac{\delta}{2}\right)\sqrt{\delta(\delta+4)} \\ &\quad -\frac{1}{2}\left(\frac{\delta}{2}+\frac{1}{2}\sqrt{\delta(\delta+4)}-\frac{1}{2}\left(\frac{\delta}{2}+\frac{1}{2}\sqrt{\delta(\delta+4)}\right)^2+\frac{1}{3}\left(\frac{\delta}{2}+\frac{1}{2}\sqrt{\delta(\delta+4)}\right)^3\right) \\ &= \frac{1}{6}\delta\sqrt{\delta(\delta+4)}+o(\delta^2) \;\geq\; \frac{\delta^{3/2}}{3}. \end{aligned}$$

Using this result, we get

$$n \cdot J_{LAG} \geq \frac{4(p-r)}{3}\,\delta^{3/2} = \varepsilon^{3/2}/M. \tag{2.8.12}$$

Consequently, (D.6) is compatible with (2.5.7).

Finally, we get

$$II = \Pr\left(\ell_1(W) > (p-r)\left(1-\frac{s}{\sqrt{n}}\right)\xi_{n,p}\right) \leq \exp\left(-\frac{4(p-r)}{3}\,\delta^{3/2}\right). \tag{2.8.13}$$

Moreover, since $\varepsilon = (p-r)\left(1-\frac{s}{\sqrt{n}}\right)\xi_{n,p} - \left(1+\sqrt{\bar{c}}\right)^2$, we can get

$$\delta = \frac{\varepsilon}{\sqrt{\bar{c}}} = \sqrt{n(p-r)}\left(1-\frac{s}{\sqrt{n}}\right)\xi_{n,p} - \frac{1}{\sqrt{\bar{c}}} - 2 - \sqrt{\bar{c}}.$$

Then,

$$\frac{4(p-r)}{3}\,\delta^{3/2} = \frac{4n}{3}(\bar{c})^{1/4}\left(\xi_{n,p}(p-r)\left(1-\frac{s}{\sqrt{n}}\right)-(1+\sqrt{\bar{c}})^2\right)^{3/2}.$$

Thus, from (D.7), we finally obtain

$$II \;\le\; \exp\left(-\frac{4n}{3}(\bar{c})^{1/4}\left(\xi_{n,p}(p-r)\left(1-\frac{s}{\sqrt{n}}\right)-(1+\sqrt{\bar{c}})^2\right)^{3/2}\right). \tag{2.8.14}$$

□

Part 4.

Proof. In addition, the term in parenthesis in (D.8) should be positive. Thus,

$$s < \sqrt{n} - \frac{1}{\xi_{n,p}\sqrt{p-r}}\left(\sqrt{\frac{n}{p-r}}+2+\sqrt{\frac{p-r}{n}}\right) \tag{2.8.15}$$

should hold. Also, throughout this proof, we assume $\delta < 1$ in the sense that ε is small. That is,

$$\delta = \sqrt{p-r}\left(\sqrt{n}-s\right)\xi_{n,p} - \sqrt{\frac{n}{p-r}} - 2 - \sqrt{\frac{p-r}{n}} < 1.$$

Thus, the following inequality should hold:

$$s > \sqrt{n} - \frac{1}{\xi_{n,p}\sqrt{p-r}}\left(\sqrt{\frac{n}{p-r}}+3+\sqrt{\frac{p-r}{n}}\right). \tag{2.8.16}$$

Hence, from (D.9) and (D.10), Theorem 2.2 holds for any value of s such that

$$\sqrt{n} - \frac{1}{\xi_{n,p}\sqrt{p-r}}\left(3+\sqrt{\bar{c}}+\frac{1}{\sqrt{\bar{c}}}\right) < s < \sqrt{n} - \frac{1}{\xi_{n,p}\sqrt{p-r}}\left(2+\sqrt{\bar{c}}+\frac{1}{\sqrt{\bar{c}}}\right). \tag{2.8.17}$$

□

References

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

Anderson, T. W. (2003), An introduction to multivariate statistical analysis (3rd ed.). New York: Wiley.

Anderson, T. W., and S. D. Gupta (1963), "Some inequalities on characteristic roots of matrices," Biometrika, *50*, 522–524.

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

Bai, Z., and J. W. Silverstein (2010), Spectral analysis of large dimensional random matrices. New York: Springer.

Baik, J., G. Ben Arous, and S. Péché (2005), "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," Annals of Probability, *33*, 1643–1697.

Baik, J., and J. W. Silverstein (2006), "Eigenvalues of large sample covariance matrices of spiked population models," Journal of Multivariate Analysis, *97*, 1382–1408.

Choi, I., and H. Jeong (2013), "Model selection for factor analysis: Some new criteria and performance comparisons," Research Institute for Market Economy (RIME) Working Paper No.1209, Sogang University.

Geman, S. (1980), "A limit theorem for the norm of random matrices," Annals of Probability, *8*, 252–261.

Gray, R. M. (2006), Toeplitz and circulant matrices: A review. now publishers inc.

Greenaway-McGrevy, R., C. Han, and D. Sul (2012), "Estimating the number of common factors in serially dependent approximate factor models," Economics Letters, *116*, 531–534.

Harding, M. (2013), "Estimating the number of factors in large dimensional factor models," preprint.

Hallin, M., and R. Liška (2007), "Determining the number of factors in the general dynamic factor model," Journal of the American Statistical Association, *102*, 603–617.

Horn, R. A., and C. R. Johnson (1991), Matrix analysis. Cambridge university press.

Jiang, J. (2010), Large sample techniques for statistics, Springer Texts in Statistics. New

York: Springer.

Johnstone, I. M. (2001), "On the distribution of the largest eigenvalue in principal components analysis," Annals of Statistics, *29*, 295–327.

Johnstone, I. M., and A. Y. Lu (2009), "On consistency and sparsity for principal components analysis in high dimensions," Journal of the American Statistical Association, *104*, 682–693.

Karoui, N. E. (2008), "On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity," arXiv preprint math/0309355.

Kritchman, S., and B. Nadler (2009), "Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory," Signal Processing, IEEE Transactions on, *57*, 3930–3941.

Ledoux, M. (2007), "Deviation inequalities on largest eigenvalues, in geometric aspects of functional analysis," In Milman, V. D., and G. Schechtman (Eds.), Lecture Notes in Mathematics. New York: Springer, 2007, vol. 1910.

Ma, C. (2003), "Spatio-temporal stationary covariance models," Journal of Multivariate Analysis, *86*, 97–107.

Ma, Z. (2012), "Accuracy of the Tracy–Widom limits for the extreme eigenvalues in white Wishart matrices," Bernoulli, *18*, 322–359.

Moon, H. R., and M. Weidner (2015), "Linear regression for panel with unknown number of factors as interactive fixed effects," Econometrica, *83*, 1543–1579.

Nadler, B. (2008), "Finite sample approximation results for principal component analysis: A matrix perturbation approach," Annals of Statistics, *36*, 2791–2817.

Nadler, B. (2010), "Nonparametric detection of signals by information theoretic criteria: Performance analysis and an improved estimator," Signal Processing, IEEE Transactions on, *58*, 2746–2756.

Nadler, B. (2011), "On the distribution of the ratio of the largest eigenvalue to the trace of a wishart matrix," Journal of Multivariate Analysis, *102*, 363–371.

Ng, S., and P. Perron (2005), "A Note on the Selection of Time Series Models," Oxford Bulletin of Economics and Statistics, *67*, 115–134.

O'leary, D. P., and G. W. Stewart (1990),"Computing the eigenvalues and eigenvectors of symmetric arrowhead matrices, " Journal of Computational Physics, *90*, 497–505.

Onatski, A. (2005), "Determining the number of factors from empirical distribution of eigenvalues," Columbia University Discussion Paper No. 0405–19.

Onatski, A. (2007), "Asymptotics of the principal components estimator of large factor models with weak factors and i.i.d. Gaussian noise," Manuscript, University of Cambridge.

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

Onatski, A. (2012), "Asymptotics of the principal components estimator of large factor models with weakly influential factors," Journal of Econometrics, *168*, 244–258.

Onatski, A. (2015), "Asymptotic analysis of the squared estimation error in misspecified factor models," Journal of Econometrics, *186*, 388–406.

Paul, D. (2007), "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," Statistica Sinica, *17*, 1617–1642.

Rao, C. (1973), Linear Statistical Inference and Its Applications (2nd ed.). New York: Wiley.

Silverstein, J. W. (1985), "The smallest eigenvalue of a large dimensional Wishart matrix," Annals of Probability, *13*, 1364–1368.

Stein, M. L. (2005), "Space-time covariance functions," Journal of the American Statistical Association, *100*, 310–321.

Szegö, G., and U. Grenander (1958), Toeplitz forms and their applications. University of California Press.

Tracy, C. A., and H. Widom (1996), "On orthogonal and symplectic matrix ensembles," Communications in Mathematical Physics, *177*, 727–754.

Zhang, Q. T., K. M. Wong, P. C. Yip, and J. P. Reilly (1989), "Statistical analysis of the performance of information theoretic criteria in the detection of the number of signals in array processing," IEEE Transactions on Acoustics, Speech, and Signal Processing, *37*, 1557–1567.

Chapter 3

On the Misdetection Probability of the Number of Factors and the Optimized Penalization in Finite Samples

3.1 Introduction

This chapter analyzes the finite sample performance of the panel information criteria (IC) proposed by Bai and Ng (2002) for detecting the number of factors and proposes modified criteria to improve its performance. To do so, we derive the computable formula for a non-asymptotic upper bound on the misdetection probability of IC and determine the optimal penalty for overfitting which leads to the minimum upper bound of the misdetection probability.

The IC estimator is a leading estimation procedure to determine the number of strong factors in large dimensional factor models. It is well known, however, the IC estimator tends to over or under detect the number of factors in finite samples and especially its misdetection worsens when the explanatory power of the factors does not strongly dominate the explanatory power of the idiosyncratic components. A few Monte Carlo studies provided evidence for such misdetection (Bai and Ng, 2002; Onatski, 2010; Greenaway-McGrevy, Han and Sul, 2012; Ahn and Horenstein, 2013). Moreover, we have partly analyzed this misdetection risk by computing the theoretical probability bound of overdetection in the previous working paper (Kao and Oh, 2017), or Chapter 2 in this dissertation.

In large dimensional panel data analysis, the misdetected true number of factors causes serious problems. In particular, when the number of factors is overestimated, users suffer from the loss of degrees of freedom. In this regard, Onatski (2015) examines the consequences of the misspecified factors for the loss of asymptotic efficiency in the principal components estimator. The under-detection of factors is also critical. Moon and Weidner (2015) show that in a linear panel regression model with unknown number of factors, the limiting distribution of the least squares estimator for the regression coefficients is independent of the estimated number of factors only if it is not underestimated. Byun and Schmidt (2016) study the effects of misspecified factors in the Fama-French factor models. Their result implies that the underestimated number of factors may cause seemingly contradictory empirical asset pricing results from the literature, such as negative and statistically insignificant risk-return

trade-off.

A major issue that should be resolved to improve the finite sample performance of IC is a non-unique penalty function in the criteria (Hallin and Liška, 2007; Ahn and Horenstein, 2013). In particular, any scalar multiple of the penalty function prespecified by Bai and Ng (2002) is still asymptotically valid for consistent estimation for the number of factors and consequently, there are asymptotically many possible choices for the penalty for overfitting. Its finite sample performance, however, depends on the magnitude of such a multiplicative weight for the penalty. Hence, it is a crucial matter in finite samples to decide what the *optimal* penalty function is.

To provide an answer to the above question, we first derive the computable formula for an upper bound on the misdetection probability of IC by employing some results from random matrix theory, under certain conditions where such a bound exists. To do so, we revisit our initial work which presented a non-asymptotic upper bound on the over-detection probability of IC and showed that when the sample size is not sufficiently large, there exists a non-negligible overestimation risk even for the case with strong factors (see Chapter 2). The current chapter extends the previous results to the under-detection risk of IC. In the end, we can diagnose its comprehensive misdetection risk in finite samples by computing non-asymptotic upper bounds on the misdetection probability of IC. Our numerical examples show that the under-detection probability of IC is non-negligible if the eigenvalue corresponding to the least influential factor is not sufficiently larger than a certain threshold, which is known as the asymptotic limit of detection of factors in random matrix theory. It implies that a threshold for finite samples may be larger than the asymptotic limit of detection.

Next, in order to find the optimal penalty in finite samples, we consider the modified version of the original criteria whose penalty function is multiplicatively weighted by a positive constant. Let us call such modified criteria the weighted information criteria (WIC). Then, by computing the misdetection probability bounds of WIC subject to the choice of a weight,

we determine the optimal weight for the penalty which leads to the minimum probability bound of misdetection. Finally, we show that the misdetection risk of IC can be controlled by the user.

Random matrix theory plays a key role in this study. In our earlier study (Chapter 2 in this dissertation), we have already introduced some preliminary results regarding the limiting behavior of the largest eigenvalue of a pure noise matrix (e.g., Geman, 1980; Tracy and Widom, 1996; Johnstone, 2001; Baik, Arous, and Péché, 2005; Baik and Silverstein, 2006; Ledoux, 2007; Paul, 2007; Karoui, 2008; Ma, 2012). Besides, in this chapter we employ additional results concerning the phase transition behavior of the least influential factor. Our analysis is also inspired by signal detection analysis in the digital signal processing literature (e.g., Kritchman and Nadler, 2009; Nadler, 2008, 2010).

This chapter is organized as follows. In Section 3.2, we describe our factor model and assumptions. Section 3.3 introduces the panel information criteria (IC) of Bai and Ng (2002). Section 3.4 presents asymptotic expressions for the over- and under-detection probabilities of the IC estimator. As mathematical preliminaries, recent results from random matrix theory are reviewed in Section 3.5. Section 3.6 derives the computable formula for an upper bound on the misdetection probability of IC and analyzes its performance for finite values of p and n such that $n > p$. Section 3.7 proposes the optimal penalty in the panel information criteria and shows numerical examples which support the better finite sample performance of our proposed method. Concluding remarks are given in Section 3.8, and all the proofs are given in the Appendix.

A word on notation. Ordinary limits are denoted by $\rightarrow$ while convergence in distribution is denoted by $\xrightarrow{d}$. Orders of magnitude for a sequence converging in probability are denoted by O_p and o_p. The transpose operator is denoted by a prime symbol as in A'. I_p denotes the identity matrix of order p. An estimate of a parameter ϑ is denoted by $\hat{\vartheta}$. $x \sim D$ means that a random variable x has the probability distribution D. The Gaussian distribution with mean μ and covariance Σ is denoted by $\mathcal{N}(\mu, \Sigma)$ while the Chi-squared distribution with n

degrees of freedom is denoted by $\chi^2(n)$. *i.i.d.* means that a random variable is independent and identically distributed. *ln* denotes a natural logarithm. $Pr(X)$ is the probability of an event X.

3.2 Model

The current chapter studies the same large dimensional factor model as described in Chapter 2. Let x_{it} be the real-valued observed data for the i-th cross-section unit at time t, for $i = 1, \ldots, p$, and $t = 1, \ldots, n$. Note that we denote the cross-sectional and temporal dimensions of the data by p and n, respectively. Consider the factor representation of the data of the form

$$x_{it} = \lambda_i' f_t + e_{it}, \tag{3.2.1}$$

where f_t is an $r \times 1$ vector of the factors, λ_i is an $r \times 1$ vector of factor loadings, and r is the *true* number of factors. $\lambda_i' f_t$ is the common component and e_{it} is the idiosyncratic error. Factors, factor loadings and the idiosyncratic components are not observable. Moreover, the true number of factors is unknown beforehand.

In vector notation, (3.2.1) can be written as a p-dimension time series with n observations:

$$\underset{(p\times 1)}{x_t} = \underset{(p\times r)}{\Lambda}\ \underset{(r\times 1)}{f_t} + \underset{(p\times 1)}{e_t}, \tag{3.2.2}$$

where $x_t = (x_{it}, \ldots, x_{pt})'$ is a p-dimensional vector of real-valued cross-section observations at time t, $\Lambda = (\lambda_1, \ldots, \lambda_p)'$ is a $p \times r$ factor loading matrix composed of r linearly independent vectors, and $e_t = (e_{it}, \ldots, e_{pt})'$ is a p-dimensional real-valued vector. In matrix notation, the model is given by

$$\underset{(p\times n)}{X} = \underset{(p\times r)}{\Lambda}\ \underset{(r\times n)}{F'} + \underset{(p\times n)}{e}, \tag{3.2.3}$$

where $X = (x_1, \ldots, x_n)$, $F = (f_1, \ldots, f_n)'$, and $e = (e_1, \ldots, e_n)$.

As in the previous Chapter 2, the following assumptions are imposed on the model. First,

suppose that f_t is the zero mean random vector and independent of e_t. Both f_t and λ_i have positive definite covariance matrices Σ_F and Σ_Λ, respectively, so that each is of full rank, r. These assumptions imply that each factor has a nontrivial contribution to variance of x_t as in Bai and Ng (2002). For discussions related to random matrix theory, both the sample size and the dimension of the observations are allowed to approach infinity simultaneously with finite ratio. By this assumption, sample eigenvalues corresponding to errors remain bounded (Onatski, 2005). Moreover, the true number of factors is fixed regardless of n and p, as generally assumed in the literature (e.g., Bai and Ng, 2002; Onatski, 2010, 2012; Ahn and Horenstein, 2013; Choi and Jeong, 2013; Harding, 2013). Lastly, the idiosyncratic components are independently and identically normally distributed, where σ is the unknown noise variance. We set $\sigma = 1$ without loss of generality since an upper bound on the misdetection probability of IC is eventually given by the ratio of eigenvalues so that σ terms are cancelled out in this ratio.

In this chapter, we consider homogeneous uncorrelated errors for technical reasons; in particular, it enables us to employ some results from random matrix theory in order to derive the misdetection probability bound of IC. Of all theoretical results from random matrix theory, a result on the asymptotic behavior of the eigenvalues of a sample covariance matrix is necessary for our study; however, it has been established only for Gaussian i.i.d. errors. For a detailed discussion on the pertinence of the i.i.d. assumption to this chapter, see our prior study as well as a few papers on the signal detection analysis (e.g., Onatski, 2007; Moon and Weidner, 2015; Harding, 2013).

Concerning random matrix theory, we interpret our model with respect to a *spiked population covariance model* introduced by Johnstone (2001), where all the population eigenvalues are one except for a few eigenvalues which are larger than one. Under the assumptions mentioned above, the population covariance matrix can be written as $\Sigma = \Psi + \Omega$, where Ψ is the covariance matrix of the common component and Ω is the error covariance matrix. In line with the assumption that the common factors have non-trivial effects on data, consider

the j-th non-zero finite population eigenvalue of Ψ, denoted by ψ_j and sorted in a decreasing order $\psi_1 \geq \psi_2 \geq \ldots \geq \psi_r > 0$. Besides, p eigenvalues of Ω are each equal to one since $\sigma = 1$. Then, the population covariance matrix Σ can be diagonalized to have the form

$$B'\Sigma B = diag\ (\psi_1, \ldots, \psi_r, 0, \ldots, 0) + I_p, \tag{3.2.4}$$

where B is a p-dimensional orthogonal matrix composed of p eigenvectors corresponding to the eigenvalues of the population covariance matrix, Σ. Obviously, p population eigenvalues of Σ are

$$(\nu_1, \nu_2, \ldots, \nu_r, 1, 1, \ldots, 1), \tag{3.2.5}$$

where $\nu_j = \psi_j + 1$ for all $j = 1, \ldots, r$.

We denote by S_n the sample covariance matrix of the n observations x_t from the model (3.2.2),

$$S_n = \frac{1}{n} \sum_{t=1}^{n} x_t x_t', \tag{3.2.6}$$

which is a $p \times p$ matrix with n samples of p-dimensional mean zero vectors. We denote the eigenvalues of S_n by $\{\ell_j\}_{j=1}^{p}$ with a decreasing order $\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p$.

Note that while each factor has a nontrivial contribution to the data, the idiosyncratic term is an irrelevant disturbance so that it does not affect the data systematically. In this sense, f_t and e_t can be referred to as signals and noise, respectively, as in the literature on signal processing. Then, the eigenvalues of Ψ can be called *noise-free* population signal eigenvalues because Ψ is of rank r, while the eigenvalues of Ω are considered as *pure noise* eigenvalues. Accordingly, the first r sample eigenvalues are roughly considered to be associated with signals, while the remaining $p - r$ sample eigenvalues roughly correspond to noise.

3.3 Detection of the Number of Factors

3.3.1 *IC* estimator

Bai and Ng (2002) set up their estimation procedure for the number of factors as a model selection problem. They proposed the panel information criteria (IC) as follows:

$$IC(k) = \ln S(k) + k \cdot G(p, n), \tag{3.3.1}$$

where k is an arbitrary number such that $k < \min\{p, n\}$, $G(p, n)$ denotes the penalty function for overfitting, and $S(k)$ is the sum of squared residuals divided by pn such that

$$S\left(k\right) = \frac{1}{pn} \sum_{i=1}^{p} \sum_{t=1}^{n} (x_{it} - \widetilde{\lambda}_i^{'k} \widetilde{f}_t^{k})^2. \tag{3.3.2}$$

$\widetilde{f}_t^{k}$ and $\widetilde{\lambda}_i^{'k}$ denote estimated factors and loadings by the principal components method given the possible number of factors k, respectively. Then, the estimator for the true number of factors (IC estimator) is obtained by minimizing (3.3.1), namely that

$$\hat{k}_{IC} = \underset{0 \leq k \leq kmax}{\arg\min} IC(k), \tag{3.3.3}$$

where $kmax$ is a bounded integer which is a maximum possible number of factors prespecified by users such that $r \leq kmax$. The IC estimator was proven to be consistent, namely that

$$\lim_{n,p \to \infty} \Pr(\hat{k}_{IC} = r) = 1, \tag{3.3.4}$$

if (1) $G(p, n) \to 0$ and (2) $C_{pn}^2 G(p, n) \to \infty$ as $n, p \to \infty$, where $C_{pn} = \min\{\sqrt{p}, \sqrt{n}\}$. That is, in the joint limit $n, p \to \infty$, the probability limit with which this model selection criterion selects the true number of factors converges to one if the penalty factor asymptotically converges to zero at an appropriate rate. Also, Bai and Ng propose specific formulations of the

penalty factor to be used in practice: $G_1(p,n) = \left(\frac{p+n}{pn}\right)\ln\left(\frac{pn}{p+n}\right)$, $G_2(p,n) = \left(\frac{p+n}{pn}\right)\ln C_{pn}^2$, and $G_3(p,n) = \frac{\ln C_{pn}^2}{C_{pn}^2}$. Finally, they consider the following three criteria associated with three penalty terms:

$$IC_1(k) = \ln S(k) + k \cdot G_1(p,n) = \ln S(k) + k \cdot \left(\frac{p+n}{pn}\right)\ln\left(\frac{pn}{p+n}\right); \tag{3.3.5}$$

$$IC_2(k) = \ln S(k) + k \cdot G_2(p,n) = \ln S(k) + k \cdot \left(\frac{p+n}{pn}\right)\ln C_{pn}^2; \tag{3.3.6}$$

$$IC_3(k) = \ln S(k) + k \cdot G_3(p,n) = \ln S(k) + k \cdot \frac{\ln C_{pn}^2}{C_{pn}^2}. \tag{3.3.7}$$

As in our previous study, IC defined in (3.3.1) can be rewritten in terms of sample eigenvalues. It is the first step for applying random matrix theory to our research topic. Consider the following *eigenvalue representation* of IC:

$$IC(k) = \ln\left(\frac{1}{p}\sum_{j=k+1}^{p}\ell_j\right) + k \cdot G(p,n). \tag{3.3.8}$$

For a short proof, see Appendix A.2.1 of Chapter 2.

3.3.2 Misdetection of the *IC* estimator

In what follows, we specify a mathematical condition for the misdetection of IC and its misdetection probability in terms of sample eigenvalues based on the eigenvalue representation of IC, (3.3.8). The current chapter focuses on the situation when IC over or under detects the true number of factors by exactly *one* factor rather than multiple factors. Readers can check the detail of this premise in our prior study (see Chapter 2, Section 2.3.2). Here we simply assume that misdetection by one signal dominates the overall performance of the information criteria as in Nadler (2010).

First, for the case in which the IC estimator overselects the true number of factors, the result has already provided in Chapter 2 (Lemma 2.2). Suppose that the criterion (3.3.1) is minimized at $r_o + 1$, where r_o is the true number of factors. It means that the IC estimator

overdetects the true number of factors by one factor, namely that $\hat{k}_{IC} = r_o + 1$. Hence, a condition for overestimation by one factor is specified as

$$\triangle IC(1) = IC(r_o) - IC(r_o + 1) > 0. \tag{3.3.9}$$

Consequently, the overestimation probability of IC is specified as follows:

Lemma 3.1. ***(Overestimation of the IC estimator)*** *Consider the model (3.2.2). Suppose that IC (3.3.1) is minimized at $r_o + 1$, where r_o is the true number of factors. Let $\{\ell_j\}_{j=1}^{p}$ denote the eigenvalues of a sample covariance matrix, S_n defined in (3.2.6), which are decreasingly ordered, $\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p$. Also, we denote by T_{p-r_o} the sum of the last $p - r_o$ eigenvalues of S_n. Then, the IC estimator overestimates the true number of factors by exactly one factor if $\triangle IC(1) > 0$ with $\triangle IC(1)$ given by (3.3.9). Thus, the probability with which the number of factors would be overestimated by exactly one factor takes the form*

$$\Pr\left(\triangle IC(1) > 0\right) = \Pr\left(\ln \frac{T_{p-r_o}}{T_{p-r_o-1}} - G(p,n) > 0\right), \tag{3.3.10}$$

where $T_{p-r_o} = \sum_{j=r_o+1}^{p} \ell_j$, $T_{p-r_o-1} = \sum_{j=r_o+2}^{p} \ell_j$.

This chapter also specifies a condition for underdetection and a corresponding probability. Let r_u denote the true number of factors when underestimation occurs by exactly one factor. It implies that a criterion function (3.3.1) is minimized at $r_u - 1$, namely that $\hat{k}_{IC} = r_u - 1$. Thus, a condition for underestimation by one factor is described as

$$\triangle IC(-1) = IC(r_u - 1) - IC(r_u) < 0. \tag{3.3.11}$$

Then, a corresponding underdetection probability is specified as follows:

Lemma 3.2. ***(Underestimation of the IC estimator)*** *Consider the model (3.2.2). Suppose that IC (3.3.1) is minimized at $r_u - 1$, where r_u is the true number of factors. Let*

$\{\ell_j\}_{j=1}^{p}$ denote the eigenvalues of a sample covariance matrix, S_n defined in (3.2.6), which are decreasingly ordered, $\ell_1 \geq \ell_2 \geq \ldots \geq \ell_p$. Also, we denote by T_{p-r_u} the sum of the last $p-r_u$ eigenvalues of S_n. Then, the IC estimator underestimates the true number of factors by exactly one factor if $\triangle IC(-1) < 0$ with $\triangle IC(-1)$ given by (3.3.11). Thus, the probability with which the number of factors would be underestimated by exactly one factor takes the form

$$\Pr\left(\triangle IC(-1) < 0\right) = \Pr\left(\ln \frac{T_{p-r_u+1}}{T_{p-r_u}} - G(p,n) < 0\right), \tag{3.3.12}$$

where $T_{p-r_u+1} = \sum_{j=r_u}^{p} \ell_j$ and $T_{p-r_u} = \sum_{j=r_u+1}^{p} \ell_j$.

Moreover, an upper bound for (3.3.12) is obtained by using the log inequality; that is,

$$\Pr\left(\triangle IC(-1) < 0\right) \leq \Pr\left(\frac{\ell_{r_u}}{T_{p-r_u}} - \frac{G(p,n)}{1-G(p,n)} < 0\right). \tag{3.3.13}$$

A simple proof: Since $\ln(1+x) \geq \frac{x}{1+x}$ for all $x > -1$, $\ln \frac{T_{p-r_u+1}}{T_{p-r_u}} = \ln\left(1 + \frac{\ell_{r_u}}{T_{p-r_u}}\right) \geq \frac{\ell_{r_u}}{T_{p-r_u+1}}$. Moreover, we can show that $\Pr\left(\frac{\ell_{r_u}}{T_{p-r_u+1}} < G(p,n)\right) = \Pr\left(\frac{\ell_{r_u}}{T_{p-r_u}} < \frac{G(p,n)}{1-G(p,n)}\right)$.

Comment Let us simply denote by r the true number of factors throughout this chapter if it is not necessary to distinguish between r_o and r_u. As shown above, Lemma 3.1 implies that the over-detection probability is defined in terms of only the last $p-r$ eigenvalues of S_n. Likewise, in Lemma 3.2, the representation of the under-detection probability involves the last $p-r$ eigenvalues of S_n. The difference is that the expression for the overestimation probability is not a function of the first r sample eigenvalues, while the expressions for the underestimation probability and its upper bound, (3.3.13), contain the r-th sample eigenvalue. This ℓ_r corresponds to the least influential factor since $\{\ell_j\}_{j=1}^{p}$ are sorted in a decreasing order.

Accordingly, the limiting behaviors of the r-th sample eigenvalue related to a signal and the last $(p-r)$ sample eigenvalues related to noise are primary concerns to derive the probability limits of (3.3.10) and (3.3.13). Fortunately, random matrix theory provides

us with related results. Regretfully, such results are only obtained for the eigenvalues of a *pure* noise covariance matrix. It should be noted, however, that ℓ_{r+1} and T_{p-r} are not truly coming from pure noise. Since the space spanned by the signal–plus–noise subspace eigenvectors contains both signals and noise, ℓ_{r+1} contains not only contributions of noise but also those of signals and the interactions between signals and noise (for details, see Nadler, 2008, Theorem 2.1). Hence, both (3.3.10) and (3.3.13) are not good enough for our analysis based on random matrix theory.

As in our prior study, now we derive more suitable expressions for the overestimation and underestimation probabilities, which are written in terms of pure noise eigenvalues, to employ random matrix theory.

3.4 Misdetection Probability

Our approach motivated by Nadler (2008, 2010) has been already introduced in the previous study to rewrite (3.3.10) in terms of pure noise eigenvalues. For more details, see Lemma 2.3 and Theorem 2.1 in Chapter 2. In the current chapter, we will show that an upper bound for the under-detection probability, (3.3.13), can be asymptotically identified in terms of pure noise eigenvalues as well. First, let us clarify related terms and introduce preliminary results.

Definition 3.1. ***Wishart matrix*** *(Silverstein, 1985; Johnstone, 2001): Let A denote a $p \times n$ matrix whose A_t are i.i.d. $\mathcal{N}(0, \Sigma_A)$ random vectors, and let $H = \frac{1}{n}AA'$. Then, the random matrix H is commonly referred to as a Wishart matrix, and $nH = AA'$ is said to have the Wishart distribution, $W_p(n, \Sigma_A)$. For the null case in which $\Sigma_A = I_p$, H is especially referred to as a Wishart matrix with identity covariance matrix.*

Obviously, in the absence of signals, n times our sample covariance matrix, nS_n, follows the null case of the Wishart distribution with parameters n and p. Here we further consider our spiked covariance model with r signals in the context of a Wishart matrix. As seen

before, $B'\Sigma B = \text{diag}(\nu_1, \ldots, \nu_r, 1, \ldots, 1)$, where $B = (b_1, \ldots, b_p)$ is a p-dimensional orthogonal matrix whose each column b_j is the eigenvector corresponding to the j-th eigenvalue of the population covariance matrix, Σ. Now, let us consider a new p-dimensional matrix $\widetilde{B} = (b_1, \ldots, b_r, \tilde{d}_{r+1}, \ldots, \tilde{d}_p)$ whose vectors are linearly independent. In particular, $\{\tilde{d}_j\}_{j=r+1}^{p}$ are the last $p-r$ column vectors which diagonalize the lower right sub-matrix of $\widetilde{B}' S_n \widetilde{B}$. Then, in the basis $\widetilde{B}$, S_n has the following form:

$$\widetilde{B}' S_n \widetilde{B} = \left[\begin{array}{ccc|ccc} \rho_{11} & \cdots & \rho_{1r} & & & \\ \vdots & \ddots & \vdots & & L' & \\ \rho_{r1} & \cdots & \rho_{rr} & & & \\ \hline & & & \widetilde{\ell}_{r+1} & & \text{Ø} \\ & L & & & \ddots & \\ & & & \text{Ø} & & \widetilde{\ell}_p \end{array}\right]. \tag{3.4.1}$$

In matrix (3.4.1), ρ_{ii} is the i-th sample variance in the directions b_i corresponding to the i-th population eigenvalue, that is, $\rho_{ii} = b_i' \left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right) b_i$ such that $\rho_{ii} \sim \left(\frac{\psi_i+1}{n}\right)\chi^2(n)$.[1] Next, $\{\widetilde{\ell}_j\}_{j=r+1}^{p}$ are the $p-r$ diagonal elements of a lower right sub-matrix in matrix (3.4.1), that is, $\widetilde{\ell}_j = \tilde{d}_j' \left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right) \tilde{d}_j$. In the basis $\widetilde{B}$, this lower right sub-matrix is given by the projection of S_n onto the only noise subspace, which is independent of the projection of S_n onto the signal subspace; therefore, it does not contain any signal contributions. Accordingly, this $p-r$ dimensional sub-matrix is considered as the random realization of a Wishart matrix with identity covariance matrix, and its diagonal elements are considered as the sample eigenvalues of this Wishart matrix, that is, pure noise eigenvalues. Thus, $\widetilde{\ell}_j \sim \chi^2(n)/n$.[2] Meanwhile, another sub-matrix L contains the interaction terms between signals and noise. If we denote by η_{ij} each element of L, then $\eta_{ij} = \tilde{d}_j' \left(\frac{1}{n}\sum_{t=1}^{n} x_t x_t'\right) b_i$ for $i = 1, \ldots, r$ and $j = r+1, \ldots, p$.

[1] By Rao (1973, p. 534), let $nH \sim W_p(n, \Sigma_A)$ and Y denote any $p \times 1$ fixed vector such that $Y'A_t \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = Y'\Sigma_A Y$. Then, $nY'HY \sim \sigma^2\chi^2(n)$.

[2] Let a_j denote the j-th eigenvalue of H and Y denote the corresponding $p \times 1$ eigenvector such that $Y'A_t \sim \mathcal{N}(0,1)$. Then, $a_j \sim \chi^2(n)/n$ and $\sum_{j=1}^{p} a_j \sim \chi^2(np)/n$. Accordingly, $E(a_j) = 1$, $Var(a_j) = 2/n$, $E(\sum_{j=1}^{p} a_j) = p$, and $Var(\sum_{j=1}^{p} a_j) = 2p/n$. Finally, $a_j = 1 + O_p(1/\sqrt{n})$ and $\sum_{j=1}^{p} a_j = p + O_p(\sqrt{p/n})$.

So far, we have identified pure noise sample eigenvalues, $\{\widetilde{\ell}_j\}_{j=r+1}^{p}$. Now, we rewrite (3.3.10) and (3.3.13) in terms $\widetilde{\ell}_j$, based on the previous literature such as O'leary and Stewart (1990, Theorem 2.1), Nadler (2008, p. 2807) and Nadler (2010). First, Theorem 3.1 below identifies the asymptotic expression for the overestimation probability regarding only pure noise eigenvalues; consequently, it is asymptotically independent of the signal eigenvalue. For the detailed proof, see Chapter 2 (Appendix A.2.3).

Theorem 3.1. ***(Overestimation Probability of IC)*** *Let W be a $p - r_o$ dimensional Wishart matrix with identity covariance matrix. The largest eigenvalue of W is denoted by $\ell_1(W)$, and the sum of $p - r_o$ eigenvalues of W is denoted by $Tr(W)$. Assuming that IC (3.3.1) is minimized at $r_o + 1$, where r_o is the true number of factors, the IC estimator overestimates the true number of factors by exactly one factor. Then, asymptotically as $n \to \infty$, the overestimation probability of IC in the presence of r_o factors is given by*

$$\Pr(\triangle IC(1) > 0) = \Pr\left(\frac{\ell_1(W)}{Tr(W)} - \xi_{n,p} > 0\right) + O_p\left(n^{-1}\right), \tag{3.4.2}$$

where $\xi_{n,p} = -1 + \sqrt{1 + 2G(p,n)}$, and $G(p,n)$ is the penalty function of IC.

Note that since a $p - r_o$ dimensional lower right sub-matrix of (3.4.1) is considered as the random realization of W, the largest eigenvalue of W, $\ell_1(W)$, is equivalent to the first pure noise eigenvalue, $\widetilde{\ell}_{r_o+1}$. Also, $Tr(W)$ is equivalent to the sum of pure noise eigenvalues, $\widetilde{T}_{p-r_o}$.

In a similar fashion, we can present the asymptotic expression for the under-detection probability of *IC* in terms of (i) $p - r_u$ pure noise eigenvalues and (ii) the r_uth sample eigenvalue corresponding to the least influential signal.

Theorem 3.2. ***(Underestimation Probability of IC)*** *Consider a $p - r_u$ dimensional Wishart matrix with identity covariance matrix denoted by W . Its largest eigenvalue is denoted by $\ell_1(W)$, and the sum of eigenvalues is denoted by $Tr(W)$. Assuming that IC (3.3.1) is minimized at $r_u - 1$, where r_u is the true number of factors, the IC estimator*

underestimates the true number of factors by exactly one factor. Then, asymptotically as $n \to \infty$*, an upper bound for the underestimation probability of IC in the presence of* r_u *factors is given by*

$$\Pr(\triangle IC(-1) < 0) \leq \Pr\left(\frac{\ell_{r_u}}{Tr(W)} - \vartheta_{p,n} < 0\right) + O_p\left(n^{-1}\right), \tag{3.4.3}$$

where $\vartheta_{p,n} = G(p,n)/(1 - G(p,n))$*, and* $G(p,n)$ *is the penalty function of the IC estimator.*

Hitherto, we derived the asymptotic expressions for the overestimation and underestimation probabilities of *IC* in terms of pure noise eigenvalues and the least influential signal eigenvalue. In what follows, we determine a non-asymptotic upper bound on the misdetection probability in finite samples. This analysis is highly related to random matrix theory since the over-detection and under-detection probabilities as presented in Theorem 3.1 and 3.2 can be pinned down by using the limiting distributions of the sample eigenvalues of a Wishart matrix.

3.5 Mathematical Preliminaries: Random Matrix Theory

The main tools used in our analysis are recent results from random matrix theory regarding the asymptotic behaviors of the eigenvalues of the sample covariance matrix when both the sample size and the dimension of the observations approach infinity such that their ratio converges to a finite value. Some general results from random matrix theory were summarized in our initial paper. See Chapter 2, and for further details Geman (1980), Johnstone (2001), Karoui (2008), Nadler (2011) and Ma (2012). In this section, we mainly focus on relevant results to this chapter.

As in Definition 3.1, let $H = AA'/n$ denote a $p \times p$ Wishart matrix, where A is a $p \times n$ matrix with real valued Gaussian i.i.d. entries, and let a_j denote the j-th sample eigenvalue with a decreasing order, for $j = 1, \ldots, p$.

3.5.1 Null Case: Wishart matrix with identity covariance matrix

First, we consider the null case in which a $p \times p$ Wishart matrix H has identity covariance matrix. Let us consider the joint limit $n, p \to \infty$ with $\frac{p}{n} \to c \in [0, \infty)$. Here we introduce the almost sure limit of the largest eigenvalue of H, its limiting distribution, and its non-asymptotic bound for finite values of p and n.

Geman (1980), along with extensions of Baik and Silverstein (2006) and Paul (2007), showed that a_1 converges to $(1 + \sqrt{c})^2$ with a probability one. Regarding the limiting distribution of a_1, Johnstone (2001), Karoui (2008) and Ma (2012) suggested that the distribution of a_1 converges to a Tracy-Widom distribution with $O(\min\{n, p\}^{-2/3})$ errors. In particular, call

$$\mu_{n,p} = \frac{1}{n}\left(\sqrt{n - 1/2} + \sqrt{p - 1/2}\right)^2,$$
$$\sigma_{n,p} = \frac{1}{n}\left(\sqrt{n - 1/2} + \sqrt{p - 1/2}\right)\left(\frac{1}{\sqrt{n - 1/2}} + \frac{1}{\sqrt{p - 1/2}}\right)^{1/3},$$

and TW_1 is the Tracy-Widom distribution of order 1 for real-valued observations, it holds

$$\frac{a_1 - \mu_{n,p}}{\sigma_{n,p}} \xrightarrow{d} TW_1. \tag{3.5.1}$$

Also, for any real h, it can be written as

$$\left|\Pr\left(\frac{a_1 - \mu_{n,p}}{\sigma_{n,p}} \leq h\right) - TW_1(h)\right| = O(\min\{n, p\}^{-2/3}), \tag{3.5.2}$$

where $TW_1(h)$ is the Tracy-Widom CDF which is defined in terms of the Airy function (for details, see Tracy and Widom, 1996; Johnstone, 2001). The above result is applied for both situations in which $n \geq p$ as well as $n < p$. It is known that this Tracy-Widom approximation is reasonable even when one of the dimensions is small.

Next, for finite values of n and p, Ledoux (2007, Proposition 2.2), Kritchman and Nadler

(2009), and Nadler (2010) provided the following result:

Remark 3.1. *For some constant $M > 0$, $\varepsilon > 0$, and $n \geq 1$,*

$$\Pr\left(a_1 \geq (1+\sqrt{\bar{c}})^2 + \varepsilon\right) \leq M \exp\left(-n \min\{\varepsilon, \varepsilon^{3/2}\}/M\right), \tag{3.5.3}$$

where $\bar{c} = p/n$ for finite values n and p. As an extension of (3.5.3),

$$\Pr\left(a_1 \geq (1+\sqrt{\bar{c}})^2 + \varepsilon\right) \leq \exp\left(-n J_{LAG}(\varepsilon)\right), \tag{3.5.4}$$

where

$$J_{LAG}(\varepsilon) = \int_1^x (x-y) \frac{(1+\bar{c})y + 2\sqrt{\bar{c}}}{(y+B)^2} \frac{dy}{\sqrt{y^2-1}}$$

with $x = 1 + (\varepsilon/2\sqrt{\bar{c}})$, and $B = (1+\bar{c})/2\sqrt{\bar{c}}$.

This chapter strongly relies on the above result since we analyze the finite-sample property of *IC* by providing an explicit non-asymptotic bound on the misdetection probability rather than the approximate analysis by using (3.5.2).

Note that all the above results are stated for the case with no signal. Nonetheless, these results can be generalized to the case with r signals. In particular, the largest $(r+1)$th diagonal element of $\widetilde{B}' S_n \widetilde{B}$ defined in (3.4.1) or equivalently the largest eigenvalue of a $p-r$ dimensional matrix H (i.e., a_1) asymptotically converges to $(1+\sqrt{\bar{c}})^2$ almost surely, where $\bar{c} = (p-r)/n$, and a_1 asymptotically follows the TW distribution with parameters n and $p-r$ (Baik and Silverstein, 2006; Paul, 2007; Karoui, 2008; Kritchman and Nadler, 2009). Remark 3.1 can be also applied to a spiked covariance model with r signals (Kritchman and Nadler, 2009); in this case, $\bar{c}$ is adjusted to $(p-r)/n$ as well.

3.5.2 Non-null Case: Spiked covariance model with i.i.d. samples

Now we consider a Wishart matrix with the non-null population covariance matrix ($\Sigma_A \neq I_p$). This can be considered as a spiked model described in (3.2.4) in which the eigenvalues

of the population covariance matrix are all one except for a few eigenvalues which are larger than one. In line with random matrix theory, here we deal with n observations which are independently and identically distributed.

Baik et al. (2005), along with refinements done in Baik and Silverstein (2006) and Paul (2007), examine the almost sure limit of signal eigenvalues in the presence of noise and their asymptotic distribution when $n, p \to \infty$ simultaneously with finite ratio. First, the following result is about the almost sure limit of the j-th largest sample eigenvalue of a spiked covariance matrix.

Remark 3.2. *(Paul, 2007, Theorem 1 and 2) Consider i.i.d. observations $\{A_t\}_{t=1}^n$ from p variate real Gaussian distribution with zero mean and covariance $\Sigma_A = diag(\nu_1, \ldots, \nu_r, 1, \ldots, 1)$ so that the j-th population eigenvalue is denoted by ν_j. Suppose that $\{\nu_j\}_{j=1}^r$ are sorted in a decreasing order and ν_j has multiplicity one. Let a_j denote the j-th sample eigenvalue for $j = 1, \ldots, r$. In the joint limit $n, p \to \infty$ with $\frac{p}{n} \to c \in (0, 1)$, the j-th largest sample eigenvalue satisfies*

$$a_j \longrightarrow \begin{cases} (1+\sqrt{c})^2 & \text{if } \nu_j \leq 1+\sqrt{c}, \\ \nu_j \left(1 + \dfrac{c}{\nu_j - 1}\right) & \text{if } \nu_j > 1+\sqrt{c} \end{cases}$$

almost surely.

Note that Paul (2007) obtained the above result only for the case with real i.i.d. Gaussian samples and $c \in (0, 1)$. In addition, Paul assumed that the r-th population eigenvalue is simple. In contrast, Baik and Silverstein (2006, Theorem 1.2 and 1.3) extended the above result to a spiked model for a general class of i.i.d. samples which are either real or complex and are not necessarily Gaussian, as well as to the cases where $c \in [1, \infty)$ (i.e., $p \geq n$) and the r-th population eigenvalues are of higher multiplicity.

Consider the model (3.2.2) with r signals and n i.i.d. samples $\{x_t\}_{t=1}^n$. Our spiked model, (3.2.4), has the first r population eigenvalues, $\{\nu_j\}_{j=1}^r$, are larger than one (i.e., $\nu_j = 1+\psi_j$ for

$j = 1, \ldots, r$), while the remaining $p-r$ population eigenvalues each equal to one. By Remark 3.2, if $\psi_r \leq \sqrt{c}$, then the corresponding r-th sample eigenvalue, ℓ_r, converges to $(1+\sqrt{c})^2$ almost surely. Note that this limit is the same as the almost sure limit of the largest pure noise eigenvalue of a Wishart matrix with identity covariance matrix as shown in the null case before. In contrast, if $\psi_r > \sqrt{c}$, ℓ_r converges to a different limit. This result implies that in the joint limit $n, p \to \infty$, the r-th largest signal (i.e., the least influential signal) is detectable only if its explanatory power represented by the corresponding population eigenvalue must be larger than a threshold, $\sqrt{c}$. Hence, this threshold is deemed as the *asymptotic limit of detection* denoted by ψ_{DET} as in Kritchman and Nadler (2009). On the other hand, if the least influential signal is weak such that $\psi_r \leq \psi_{DET}$, then ℓ_r corresponding to this weak signal converges to the same limit of the last $p-r$ sample eigenvalues corresponding to noise; consequently, such a weak signal is not well separated from noise asymptotically.

Next, by following Paul (2007, Theorem 1 and 2) and Kritchman and Nadler (2009), we recap another result regarding the distributional limit of the r-th sample eigenvalue associated with the strong r-th signal whose population eigenvalue is larger than the asymptotic limit of detection.

Remark 3.3. *Consider i.i.d. observations $\{A_t\}_{t=1}^n$ from p variate real Gaussian distribution with zero mean and covariance $\Sigma_A = diag(\nu_1, \ldots, \nu_r, 1, \ldots, 1)$. Let ν_j and a_j denote the j-th population eigenvalue and the j-th sample eigenvalue sorted in a decreasing order for $j = 1, \ldots, r$, respectively. Suppose that $\nu_r > 1 + \sqrt{c}$ and that ν_r has multiplicity one. Then, in the joint limit $n, p \to \infty$ with $\frac{p}{n} \to c \in (0, 1)$, the limiting distribution of the r-th largest sample eigenvalue is Gaussian,*

$$\sqrt{n}(a_r - \pi(\nu_r)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\nu_r)), \tag{3.5.5}$$

where $\pi(\nu_r) = \nu_r \left(1 + \frac{\bar{c}}{\nu_r - 1}\right)$, $\sigma^2(\nu_r) = 2\nu_r^2 \left(1 - \frac{\bar{c}}{(\nu_r - 1)^2}\right)$, and $\bar{c} = \frac{p-r}{n}$.

Note that the above result has been also extended to complex i.i.d. Gaussian samples

and the higher multiplicity of v_j (e.g., Baik et al., 2005; Baik and Silverstein, 2006). Remark 3.3 says that if the r-th signal population eigenvalue is larger than the asymptotic limit of detection, $\psi_{DET} = \sqrt{c}$, or concisely if the r-th signal is sufficiently strong, then the corresponding sample eigenvalue satisfies an asymptotic normality. This result will be used directly to derive a non-asymptotic upper bound on the underestimation probability of *IC* in the presence of r strong signals. In contrast, Baik and Silverstein (2006) show that if $\psi_r \leq \psi_{DET}$, the r-th sample eigenvalue asymptotically follows the same Tracy-Widom distribution as the largest sample eigenvalue of a Wishart matrix with identity covariance matrix (i.e., the largest pure noise sample eigenvalue) as described in (3.5.1).

To sum up, these two remarks imply that if the non-unit eigenvalues of a Wishart matrix are close to one, their sample eigenvalues show a similar asymptotic behavior to pure noise eigenvalues as if the population covariance matrix is the identity matrix. On the contrary, if the non-unit eigenvalues are quite distinct from one (i.e., $\nu_j > 1 + \psi_{DET}$), corresponding sample eigenvalues have a different asymptotic property. Such asymptotic behaviors are referred to as a *phase transition phenomenon* in the literature.

3.6 Upper Bound on Misdetection Probability

This section finally examines a non-asymptotic bound on the misdetection probability of *IC*. We derive each bound for overdetection and underdetection separately.

3.6.1 Non–asymptotic Bound on Overestimation Probability

Here we recap a result regarding a non-asymptotic upper bound on the over-detection probability from our previous study. By applying Remark 3.1 to (3.3.10), Chapter 2 (Theorem 2.2) provided the following result.

Theorem 3.3. *Consider the model (3.2.2) in the presence of r_o signals and the panel information criteria (IC) defined in (3.3.1). Suppose that the IC estimator overestimates the*

true number of factors by exactly one factor, namely that IC is minimized at $r_o + 1$. *Then, a non-asymptotic upper bound on the overestimation probability of IC by exactly one factor is given by*

$$\Pr(\triangle IC(1) > 0) \leq \exp\left(\frac{-(p-r_o)s_o^2}{4}\right) + \exp\left(-\frac{4n}{3}(\bar{c}_o)^{1/4}\left((p-r_o)\left(1-\frac{s_o}{\sqrt{n}}\right)\xi_{n,p} - (1+\sqrt{\bar{c}_o})^2\right)^{3/2}\right), \tag{3.6.1}$$

for finite values of n *and* p. *This bound is appropriate for any positive value of* s_o *chosen by a user such that*

$$\sqrt{n} - \frac{1}{\xi_{n,p}\sqrt{p-r_o}}\left(3+\sqrt{\bar{c}_o}+\frac{1}{\sqrt{\bar{c}_o}}\right) < s_o < \sqrt{n} - \frac{1}{\xi_{n,p}\sqrt{p-r_o}}\left(2+\sqrt{\bar{c}_o}+\frac{1}{\sqrt{\bar{c}_o}}\right), \tag{3.6.2}$$

where $\bar{c}_o = \frac{p-r_o}{n}$ *and* $\xi_{n,p} = -1 + \sqrt{1+2G(p,n)}$. *Also, (3.6.1) holds for all the formulations of the penalty function* $G(p,n)$ *which are specified in (3.3.5), (3.3.6), and (3.3.7).*

Theorem 3.3 provides users with a simple diagnostic tool for the overdetection of the number of factors. It discloses numerically how maximally overestimation occurs so long as users know the temporal and cross-sectional size of the data. Recall that $\bar{c}$ and $\xi_{n,p}$ are functions of n and p. Also, the appropriate value of s_o depends on n and p. In practice, the user can choose the value of s_o such that it can minimize the upper bound defined in (3.6.1) as long as it satisfies (3.6.2).

Our prior work analyzed the over-detection performance of IC and provided numerical examples for practical users, by computing upper bounds on the over-detection probability according to finite values of n, p and $\hat{k}_{IC}$, and the choice of $G(p,n)$. Examples showed that when sample sizes are small, the over-detection risk is not negligible even in the presence of strong factors and the i.i.d. error components. Those findings were true for all the formulations of the penalty function; however, when we choose $G_2(p,n)$ as a penalty function, or equivalently when we use $IC_2(k)$, we obtain the lowest bounds. On the other hand, upper

bounds are particularly high when we employ $G_3(p,n)$. Such differences become negligible as the sample size grows.

Also, we saw that the overestimation probability given the sample size tends to increase as the estimated number of factors becomes larger. As the dimension of a noise subspace $(p - \hat{k}_{IC})$ shrinks, the effect of the idiosyncratic components weakens, whereas the relative explanatory power of signals is likely to be overly inflated. Obviously, when sample sizes are sufficiently large, we obtained nearly zero upper bounds. For more detailed results, see Table 2.1 and Figure 2.1 in Chapter 2.

3.6.2 Non–asymptotic Bound on Underestimation Probability

Now, we newly derive the computable formula for a non-asymptotic upper bound on the under-detection probability of *IC* by exactly one factor and also provide the numerical examples for practical users. The following theorem is derived from Remark 3.3.

Theorem 3.4. *Consider a dataset of n i.i.d. real Gaussian samples* $\{x_t\}_{t=1}^n$ *from the model (3.2.2) in the presence of* r_u *signals with a population covariance* $\Sigma = diag(\nu_1, \ldots, \nu_{r_u}, 1, \ldots, 1)$, *where* ν_j *is sorted in a decreasing order for* $j = 1, \ldots, r_u$. *Suppose that the IC estimator underestimates the true number of factors by exactly one factor, namely that IC is minimized at* $r_u - 1$. *Further, suppose that* $\nu_{r_u} > 1 + \sqrt{c}$ *and that* ν_{r_u} *has multiplicity one. Then, for any value of* $s_u \in [0, 2\sqrt{n})$, *a non-asymptotic upper bound on the underestimation probability of IC by exactly one factor is given by*

$$\Pr(\triangle IC(-1) < 0) \leq \exp\left(\frac{-3(p - r_u)s_u^2}{16}\right) + F_n(z), \tag{3.6.3}$$

where

$$F_n(z) = \begin{cases} 1 - \dfrac{2\phi(z)}{\sqrt{4 + z^2} + z} & \text{if } z \geq 0, \\ \\ \dfrac{2\phi(-z)}{\sqrt{2 + z^2} - z} & \text{if } z < 0 \end{cases} \tag{3.6.4}$$

by setting $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, *and*

$$z = \frac{\sqrt{n}}{\sigma(\nu_{r_u})} \left((p - r_u) \left(1 + \frac{s_u}{\sqrt{n}} \right) \vartheta_{n,p} - \pi(\nu_{r_u}) \right),$$

with $\pi(\nu_{r_u}) = \nu_{r_u} \left(1 + \frac{\bar{c}_u}{\nu_{r_u} - 1} \right)$, $\sigma^2(\nu_{r_u}) = 2\nu_{r_u}^2 \left(1 - \frac{\bar{c}_u}{(\nu_{r_u} - 1)^2} \right)$, $\bar{c}_u = \frac{p - r_u}{n}$, *and* $\vartheta_{n,p} = \frac{G(p,n)}{(1 - G(p,n))}$. *(3.6.3) holds for all the formulations of the penalty function* $G(p, n)$ *which are specified in (3.3.5), (3.3.6), and (3.3.7).*

Theorem 3.4 can be used to diagnose the underestimation risk of *IC* if users know sample sizes and the population eigenvalue of the least influential signal. The appropriate positive value of s_u can be chosen such that it can minimize (3.6.3) as long as $s_u \in [0, 2\sqrt{n})$.

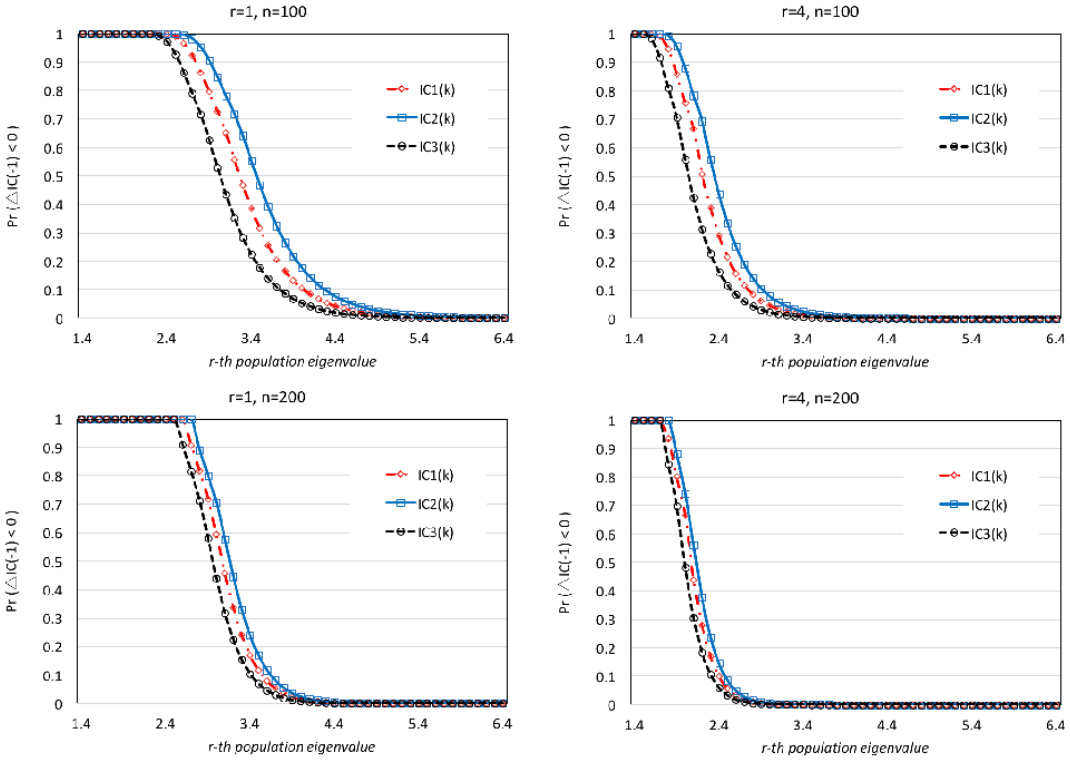
3.6.3 Numerical Examples of Under-detection Probability Bounds

This subsection analyzes the under-detection performance of the *IC* estimator in the presence of strong factors and provides its numerical examples for practical users. In particular, we use Theorem 3.4 to compute upper bounds on the under-detection probability subject to the sample size and the estimated number of factors, the choice of a penalty term, and the population eigenvalue corresponding to the least influential factors. In each case, s_u in (3.6.3) was chosen by minimizing an upper probability bound such that $s_u \in [0, 2\sqrt{n})$. Main results are illustrated in Figure 3.1 and 3.2.

Figure 3.1 shows how an upper bound on the underestimation probability of *IC* varies with the r_uth population eigenvalue, ν_{r_u}. First, we can see that even when factors have nontrivial contributions to variation in the data and the error components are i.i.d, the underestimation probability is not negligible for the case with small sample sizes. As ν_{r_u} becomes larger, however, an upper bound on the underestimation probability decreases.

These findings suggest the finite-sample implication of a phase transition phenomenon predicted by random matrix theory. Although the least influential signal is strong so that $\psi_{r_u} > \psi_{_{DET}}$, the underestimation risk of *IC* is still not negligible unless the r_uth eigenvalue

Figure 3.1: Under-detection of the IC estimator ($p = 10$)

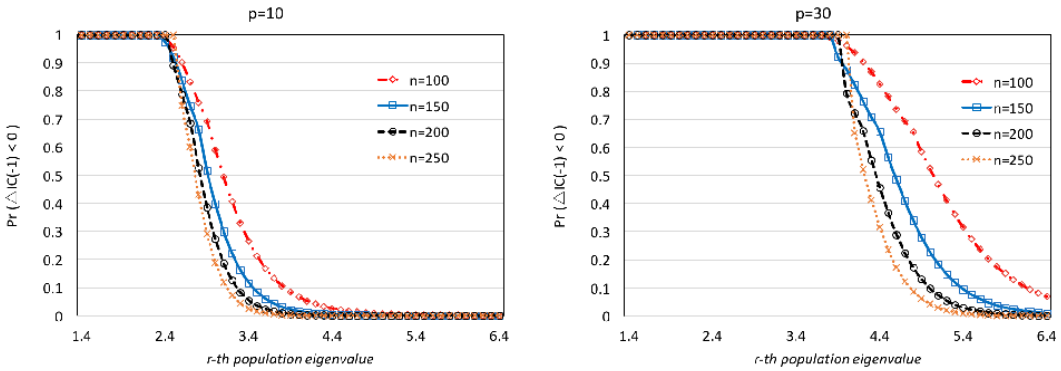


Note: This plots an upper bound on the underestimation probability of the IC estimator, $\Pr(\triangle IC(-1) < 0)$ defined in Theorem 3.4. A bound is computed by the formula (3.6.3). We consider the true number of factors $r_u \in \{1, 4\}$ such that $r_u = \hat{k}_{IC} + 1$. We only present the case with $(n, p) \in \{(100, 10), (200, 10)\}$ and the increasing r_uth population eigenvalue from 1.4 to 6.4. Each panel compares the under-detection probability bounds of three different panel information criteria, $IC_1(k)$, $IC_2(k)$ and $IC_3(k)$ which are defined in (3.3.5), (3.3.6) and (3.3.7), respectively.

is sufficiently larger than the asymptotic limit of detection. For example, in the left upper panel for $(r_u, n, p) = (1, 100, 10)$, the upper bound on the underestimation probability is still over 90% when $\psi_{r_u} = 0.4 > \psi_{DET} \approx 0.3$. An upper bound is under 50% only after $\psi_{r_u} \approx 2.0$. It implies that for finite samples, the r_uth factor might not be detected with high probability if the explanatory power of the signal does not sufficiently dominate the explanatory power of the error components. That is, even though Remark 3.2 and 3.3 is asymptotically true, we might need a much larger threshold for small sample sizes in order that the signal can be clearly separated from noise and consequently well detected.

The above interpretation is consistent with theoretical results in the previous literature.

Figure 3.2: Under-detection of the IC estimator (IC_2)



Note: This plots an upper bound on the underestimation probability of the IC estimator, $\Pr(\triangle IC(-1) < 0)$ defined in Theorem 3.4. A bound is computed by the formula (3.6.3). We only present the case with the true number of factors $r_u = 2$ such that $r_u = \hat{k}_{IC} + 1$. We consider $p \in \{10, 30\}$ and the increasing r_uth population eigenvalue from 1.4 to 6.4. Each panel compares the under-detection probability bounds of $IC_2(k)$, defined in (3.3.6), according to different sample sizes, $n \in \{100, 150, 200, 250\}$.

Ahn and Horenstein (2013), Onatski (2010) and Harding (2013) studied the limiting behavior of sample eigenvalues when signals are not sufficiently strong. They argued if the explanatory power of r_uth signal does not strongly dominate that of noise, it is difficult to separate eigenvalues into signals and noise in small sample sizes.

Next, comparing the left panels and the right panels in Figure 3.1, we can see that the under-detection probability falls as the estimated number of factors increases, given the sample size and ν_{r_u}. This can be explained by the shrinkage of a noise subspace $p - \hat{k}_{IC}$ which leads to the decreasing effect of the idiosyncratic components.

Moreover, the above findings hold for all the formulations of the penalty function. However, when we choose $G_3(p, n)$ as a penalty function, or equivalently when we use $IC_3(k)$, upper bounds on the underestimation probability are lower than any other cases. On the other hand, the IC_2 estimator yields a higher underestimation probability bound. This is the opposite of a result for over-detection.

Figure 3.2 describes how the underestimation probability varies with n and p. First, as the sample size (n) increases, an upper bound on the underestimation probability falls given ν_{r_u}. Second, for the data with larger population size (p), we obtain a higher upper

bound given ν_{r_u}. It is closely related to a phase transition phenomenon. Obviously, since ψ_{DET} $(= \sqrt{c})$ increases with p, a larger ν_{r_u} is required for the r_uth signal to be detected as p grows. More precisely, since the cumulative effect of $p - r_u$ noise components grows with p, the r_uth signal may not be clearly distinguished from noise components as p increases. Monte Carlo studies in the literature support our finding as well. For example, Harding (2013, Table 1) reports the finite sample performance of the *IC* estimator under Gaussian i.i.d. factors and errors, and it shows that even when factors are strong, the true number of factors is more likely to be underestimated with larger p.

3.7 Optimized Penalization for Detecting the Number of Factors

So far, we have identified non-asymptotic bounds on the over- and under-detection probabilities of the *IC* estimator. In this section, we will address the second question about the optimal penalty. To do so, we first present a non-asymptotic upper bound on the overall misdetection probability of the *IC* estimator by merging Theorem 3.3 and 3.4. Then, we can find the optimal weight for the penalty function which leads to the minimum bound of the misdetection probability. Before proceeding, we briefly introduce our idea for the optimal penalty.

3.7.1 Optimal Penalty for overfitting

As shown in the previous section, the *IC* estimator has a non-negligible over-detection probability in small sample sizes, and it also has a non-negligible under-detection probability especially when signals are not sufficiently strong. These results raise an interesting question of how to reduce or, more rigorously, how to minimize the misdetection probability of the *IC* estimator preserving its consistency.

Here is a clue to the answer to this question. As Hallin and Liška (2007) and Ahn and Horenstein (2013) pointed out, the penalty function defined by Bai and Ng (2002) is not unique since it is only required to satisfy certain asymptotic conditions for the consistency

of the IC estimator. For example, we can consider any positive constant (w) and refer to $w \cdot G(p,n)$ as a *weighted* penalty function. Then, this weighted penalty still satisfies the asymptotic conditions: (i) $w \cdot G(p,n) \to 0$ and ii) $C_{pn}^2 \cdot w \cdot G(p,n) \to \infty$ as $n, p \to \infty$ because w is fixed regardless of n and p. However, the finite sample performance of the panel information criteria with this weighted penalty is affected by the magnitude of w so that it will be different from the performance of the original IC. Nadler (2010) employed a similar idea and modified the Akaike information criterion (AIC) by multiplying its original penalty term by an arbitrary constant; however, Nadler focused on only the overestimation probability of AIC and did not provide a theoretical guidance on how to choose this constant.

This chapter develops the above idea so that we can deal with both over- and under-detection risk and finally propose the *optimal weight* for the penalty which minimizes the *overall* misdetection risk. In particular, if $w > 1$, a weighted penalty function yields a higher penalty for overfitting; consequently, the overestimation probability reduces in finite samples, whereas the underestimation probability worsens to some extent. On the other hand, if $w < 1$, a weighted factor lessens the penalty for overfitting; hence, it would mitigate the underestimation risk, while it is likely to aggravate the overestimation risk. As a consequence, a change in w leads to a trade-off between the over- and under-detection risk of the information criteria. By using this trade-off, we can determine the optimal weight (w^*) for the penalty factor such that it minimizes the sum of non-asymptotic upper bounds on the over-detection and under-detection probabilities.

3.7.2 Weighted Information Criteria and Misdetection Probability

Now, we present the computable formula for a non-asymptotic upper bound on the misdetection probability of the original IC estimator by one factor. Recall that the true number of factors is denoted by r_o for overestimation cases while r_u for underestimation cases. As mentioned before, each case is defined for the situation when the IC estimator over or under detects by only one factor; that is, $\hat{k}_{IC} = r_o + 1$ or $r_u - 1$.

Here we denote by $\Pr(\triangle IC \neq 0)$ the probability that the true number of factors is misdetected by one factor. Then, it is the sum of (3.3.10) and (3.3.12) since these two events are mutually exclusive:

$$\Pr(\triangle IC \neq 0) = \Pr(\triangle IC(1) > 0) + \Pr(\triangle IC(-1) < 0) \ . \tag{3.7.1}$$

Combining Theorem 3.3 and 3.4, we can accordingly formulate a non-asymptotic upper bound on the misdetection probability of IC as follows:

Corollary 3.5. ***(Non-asymptotic bound on the misdetection probability of IC)*** *Consider a dataset of n i.i.d real Gaussian samples $\{x_t\}_{t=1}^{n}$ from the model (3.2.2) in the presence of r signals with a spiked population covariance matrix defined in (3.2.4); that is, $\Sigma = diag(\nu_1, \ldots, \nu_r, 1, \ldots, 1)$, where ν_j is sorted in a decreasing order for $j = 1, \ldots, r$. Suppose that the IC estimator over or under estimates the true number of factors by exactly one factor. Let r_o denote the true number of factors for the case of overestimation by one factor and r_u denote the true number of factors for the case of underestimation by one factor. Further, for the case of underestimation, suppose that $\nu_{r_u} > 1 + \sqrt{c}$ and that ν_{r_u} has multiplicity one. Then, a non-asymptotic upper bound on the misdetection probability of IC by exactly one factor is given by*

$$\begin{aligned}\Pr(\triangle IC \neq 0) \leq & \exp\left(-\frac{4n}{3}(\bar{c}_o)^{1/4}\left((p-r_o)\left(1-\frac{s_o}{\sqrt{n}}\right)\xi_{n,p} - (1+\sqrt{\bar{c}_o})^2\right)^{3/2}\right) + \\ & \exp\left(\frac{-(p-r_o)s_o^2}{4}\right) + \exp\left(\frac{-3(p-r_u)s_u^2}{16}\right) + F_n(z),\end{aligned} \tag{3.7.2}$$

where $\bar{c}_o$, $\xi_{n,p}$ and $F(z)$ are those defined in Theorem 3.3 and 3.4. This non-asymptotic bound is appropriate for any positive value of $s_u \in [0, 2\sqrt{n})$ and s_o which satisfies (3.6.2).

Next, we define modified criteria by considering a weighted penalty factor, $w \cdot G(p, n)$. Let us call this modified version of IC the *weighted* panel information criteria and denote it

by *WIC*. Then, *WIC* has the form

$$WIC(k,w) = \ln S(k) + kw \cdot G(p,n), \tag{3.7.3}$$

where k is an arbitrary number ($k < \min\{p,n\}$), w is a fixed positive scalar and $S(k)$ is the sum of squared residuals is divided by pn. $G(p,n)$ is the penalty function which has three different forms: $G_1(p,n) = \left(\frac{p+n}{pn}\right)\ln\left(\frac{pn}{p+n}\right)$; $G_2(p,n) = \left(\frac{p+n}{pn}\right)\ln C_{pn}^2$; and $G_3(p,n) = \frac{\ln C_{pn}^2}{C_{pn}^2}$, where $C_{pn} = \min\{\sqrt{p},\sqrt{n}\}$. In relation to three formulations of the penalty factor, we consider three criteria:

$$WIC_1(k,w) = \ln S(k) + kw \cdot G_1(p,n); \tag{3.7.4}$$

$$WIC_2(k,w) = \ln S(k) + kw \cdot G_2(p,n); \tag{3.7.5}$$

$$WIC_3(k,w) = \ln S(k) + kw \cdot G_3(p,n). \tag{3.7.6}$$

Since the only difference between *IC* and *WIC* is a weight for $G(p,n)$, a non-asymptotic upper bound on the misdetection probability of *WIC* can be directly obtained from Corollary 3.5.

Corollary 3.6. ***(Non-asymptotic bound on the misdetection probability of WIC)*** *Consider the weighted panel information criteria (3.7.3), denoted by WIC. Under the conditions and notations in Corollary 3.5, a non-asymptotic upper bound on the misdetection probability of WIC by exactly one factor is given by*

$$\begin{aligned}\Pr(\triangle WIC \neq 0) \leq & \exp\left(-\frac{4n}{3}(\bar{c}_o)^{1/4}\left((p-r_o)\left(1-\frac{s_o}{\sqrt{n}}\right)\ddot{\xi}_{n,p} - (1+\sqrt{\bar{c}_o})^2\right)^{3/2}\right) + \\ & \exp\left(\frac{-(p-r_o)s_o^2}{4}\right) + \exp\left(\frac{-3(p-r_u)s_u^2}{16}\right) + F_n(\ddot{z}) \\ = & P_{ub}(\triangle WIC \neq 0),\end{aligned} \tag{3.7.7}$$

where $\ddot{\xi}_{n,p} = \sqrt{1+2w\cdot G(p,n)} - 1$ *and* $\ddot{z} = \frac{\sqrt{n}}{\sigma(\nu_{r_u})}\left((p-r_u)\left(1+\frac{s_u}{\sqrt{n}}\right)\ddot{\vartheta}_{n,p} - \pi(\nu_{r_u})\right)$, *with*

$\ddot{\vartheta}_{n,p} = \frac{w \cdot G(p,n)}{1 - w \cdot G(p,n)}$. *Also,* $\bar{c}_o$, $\pi(\cdot)$, $\sigma^2(\cdot)$ *and* $F(\cdot)$ *are those defined in Theorem 3.3 and 3.4. This bound is appropriate for any positive value of* $s_u \in [0, 2\sqrt{n})$ *and* s_o *which satisfies*

$$\sqrt{n} - \frac{1}{\ddot{\xi}_{n,p}\sqrt{p - r_o}} \left(3 + \sqrt{\bar{c}_o} + \frac{1}{\sqrt{\bar{c}_o}} \right) < s_o < \sqrt{n} - \frac{1}{\ddot{\xi}_{n,p}\sqrt{p - r_o}} \left(2 + \sqrt{\bar{c}_o} + \frac{1}{\sqrt{\bar{c}_o}} \right) . \quad (3.7.8)$$

Note that Corollary 3.6 is the same as Corollary 3.5 except for $\ddot{\xi}_{n,p}$ and $\ddot{z}$ which are defined in terms of $w \cdot G(p,n)$, not $G(p,n)$. Finally, we can find the optimal weight (w^*) for the penalty for overfitting by minimizing an upper bound on the misdetection probability of *WIC* presented in Corollary 3.6, given sample sizes, the least influential population eigenvalue ν_{r_u}, and the choice of a penalty function. That is,

$$w^* = \arg\min_{w>0} P_{ub}(\triangle WIC \neq 0). \quad (3.7.9)$$

Let us conclude this subsection by considering a signal detection procedure in line with (3.7.9), which leads to the minimum upper bound of the misdetection probability of the number of factors. That is,

$$\hat{k}_{_{WIC}} = \arg\min_{0 \leq k \leq kmax} WIC(k, w^*), \quad (3.7.10)$$

where $kmax$ is a bounded integer which is a maximum possible number of factors prespecified by users and w^* is the optimal weight for the penalty for overfitting defined in (3.7.9). A possible algorithm for this estimation procedure is conjectured as follows:

1. Estimate the number of factors $\hat{k}_{_{IC}}$ by the *IC* estimator. Set $r_u = \hat{k}_{_{IC}} + 1$ and $r_o = \hat{k}_{_{IC}} - 1$.

2. Given r_u and r_o, find w^* which minimizes (3.7.7).

3. Given w^*, estimate the number of factors $\hat{k}_{_{WIC}}$ based on (3.7.10).

The empirical validity of this estimation procedure is left for a future study.

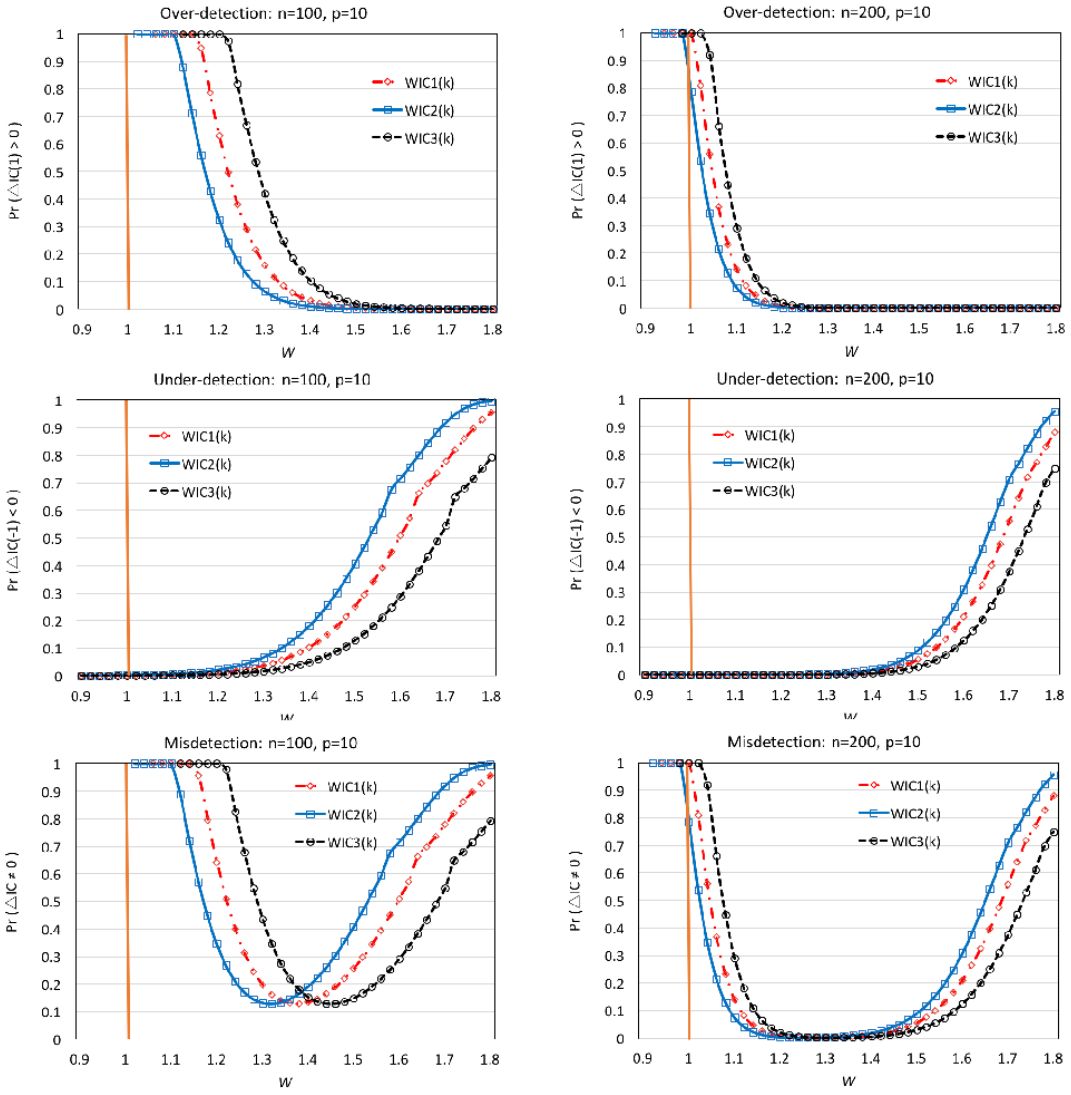
3.7.3 Numerical Examples of the Optimized Penalization

As a counterpart to the performance analysis of *IC* in Section 3.6, here we examine the finite sample performance of *WIC* by computing its non-asymptotic bound on the misdetection probability given by (3.7.7). Moreover, we can see how the optimal weight for the penalty is determined given ν_{r_u}, n, p, and the choice of a penalty term.

First, Figure 3.3 illustrates the detection performance of *WIC* for different weights and the choice of a penalty term. Without loss of generality, we report results for the data with $\hat{k}_{IC} = 4$, $\nu_{r_u} = 3.8$ and $(n, p) \in \{(100, 10), (200, 10)\}$. Obviously, when $w = 1$, *WIC* is the same as *IC*. As predicted theoretically, we see that as w becomes larger, the over-detection probability bound of *WIC* falls, whereas the under-detection probability bound of *WIC* increases. Due to this trade-off, we can achieve the minimum upper bound of the misdetection probability by adjusting w. Comparing this minimum bound with the upper bound for the original *IC* when $w = 1$, we can see that detection performance is substantially improved. For example, the left panels show that when we use WIC_1, an upper bound on the misdetection probability is minimized at $w^* \approx 1.4$, and consequently it decreases from 100% at $w = 1$ to around 10%. Obviously, comparing the left and right panels, we can see that as the sample size (n) increases, the misdetection probability decreases given w, and a smaller weight is needed to achieve the minimum bound.

Figure 3.4 considers the cases with a lower r_uth population eigenvalue ($\nu_{r_u} = 2.8$). Comparing the left panels in this figure to the right panels of Figure 3.3, we can see that as the strength of a signal becomes weaker, under-detection risk worsens so that an upper bound on the misdetection probability increases given w. The previous findings in Figure 3.3 are still supported, however. By adjusting a weight for the penalty, we can obtain a minimum bound so that an upper bound on the misdetection probability decreases substantially from at least 80% for IC_2 to less than 10% (Left panels). In addition, comparing the left and right panels, we can see that as the population size (p) increases, the overestimation probability decreases while the underestimation probability increases as discussed before.

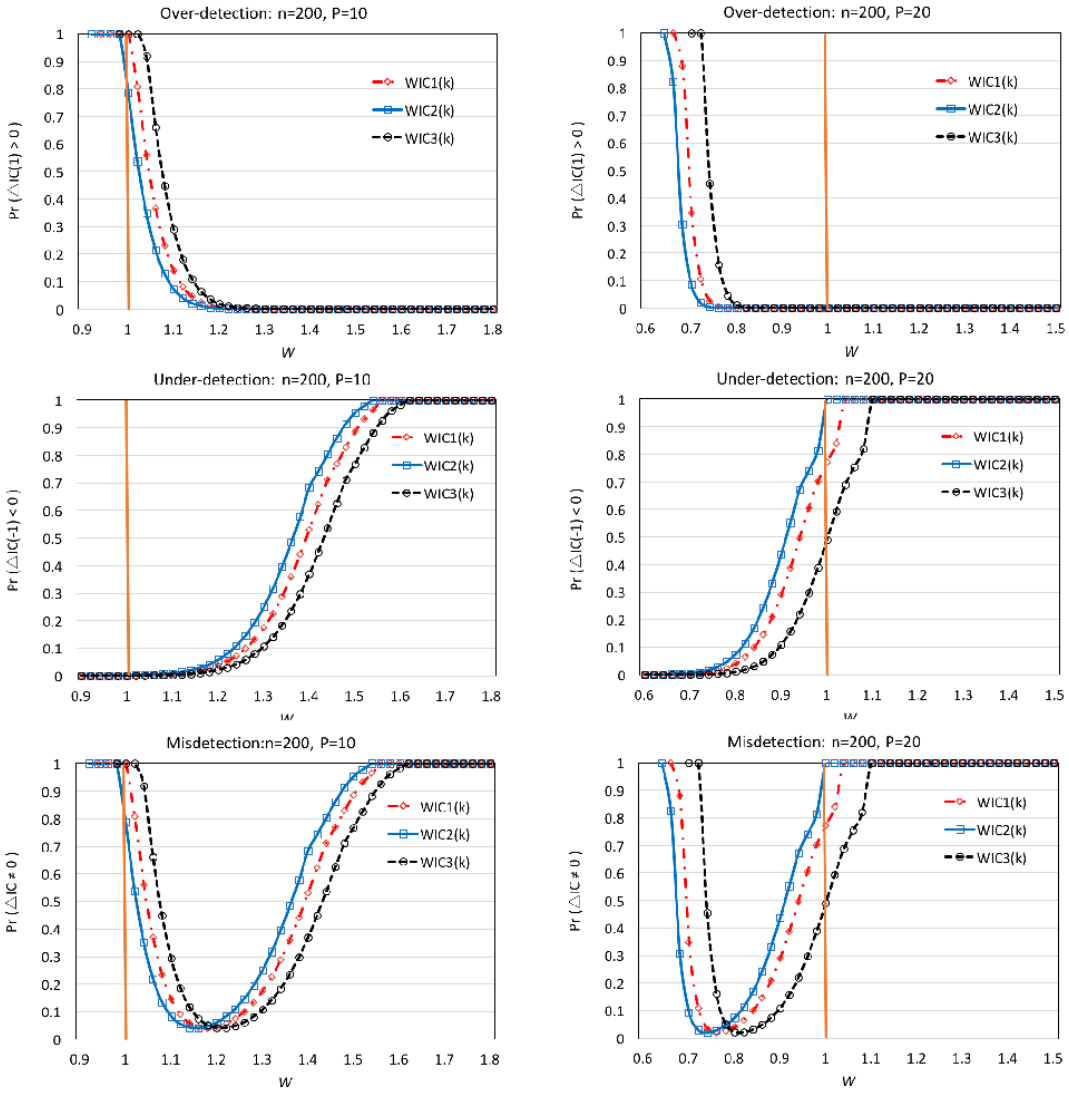
Figure 3.3: Performance of the *WIC* estimator and Optimal Weight ($\hat{k}_{IC} = 4, \nu_{r_u} = 3.8$)



Note: This plots upper bounds on the over- (top panels), under- (middle panels), and overall mis-detection (bottom panels) probabilities of the *WIC* estimator, $\Pr(\triangle WIC \neq 0)$ defined in Corollary 3.6. A bound is computed by the formula (3.7.7). We consider the true number of factors $r_u = 5$ such that $r_u = \hat{k}_{IC} + 1$ and $r_o = 3$ such that $r_o = \hat{k}_{IC} - 1$, respectively. We only present the case with $(n, p) \in \{(100, 10), (200, 10)\}$ and the r_uth population eigenvalue $\nu_{r_u} = 3.8$. Each panel compares the misdetection probability bounds of three different panel information criteria, $WIC_1(k)$, $WIC_2(k)$ and $WIC_3(k)$ which are defined in (3.7.4), (3.7.5) and (3.7.6), respectively.

To explore in more detail the effect of the signal strength to misdetection risk and the optimal weight, Figure 3.5 depicts the cases with much lower eigenvalues of the least influential signal ($\psi_{r_u} = 0.8$ in the left panels and $\psi_{r_u} = 1.0$ in the right panels). Although

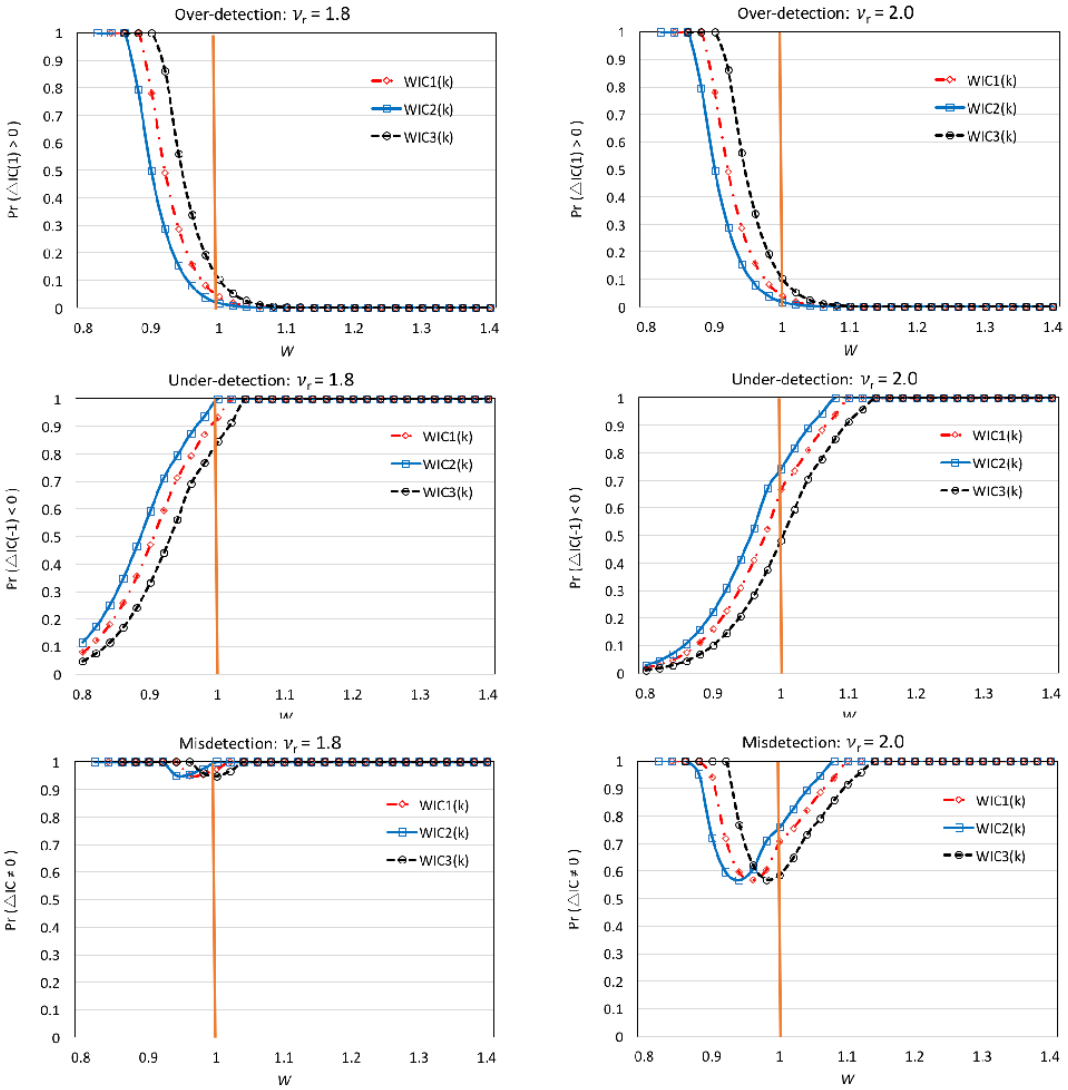
Figure 3.4: Performance of the WIC estimator and Optimal Weight ($\hat{k}_{IC} = 4, \nu_{r_u} = 2.8$)



Note: This plots upper bounds on the over- (top panels), under- (middle panels), and overall mis-detection (bottom panels) probabilities of the WIC estimator, $\Pr(\triangle WIC \neq 0)$ defined in Corollary 3.6. A bound is computed by the formula (3.7.7). We consider the true number of factors $r_u = 5$ such that $r_u = \hat{k}_{IC} + 1$ and $r_o = 3$ such that $r_o = \hat{k}_{IC} - 1$, respectively. We only present the case with $(n, p) \in \{(200, 10), (200, 20)\}$ and the r_uth population eigenvalue $\nu_{r_u} = 2.8$. Each panel compares the misdetection probability bounds of three different panel information criteria, $WIC_1(k)$, $WIC_2(k)$ and $WIC_3(k)$ which are defined in (3.7.4), (3.7.5) and (3.7.6), respectively.

r_uth signal eigenvalues are larger than the asymptotic limit of detection in both cases, the under-detection probability bound of IC (i.e., the upper bound at $w = 1$) is still high. This is consistent with our previous finding in Section 3.6; that is, unless the strength of a signal

Figure 3.5: Effect of Signal Strength to Detection Performance and Optimal Weight



Note: This plots upper bounds on the over- (top panels), under- (middle panels), and overall mis-detection (bottom panels) probabilities of the WIC estimator, $\Pr(\triangle WIC \neq 0)$ defined in Corollary 3.6. A bound is computed by the formula (3.7.7). We only present the case with $(n, p) = (200, 10)$ and $\hat{k}_{IC} = 3$. We consider two cases: (i) $\nu_{r_u} = 1.8$ (left three panels) and (ii) $\nu_{r_u} = 2.0$. (right three panels). Each panel compares the misdetection probability bounds of three different panel information criteria, $WIC_1(k)$, $WIC_2(k)$ and $WIC_3(k)$ which are defined in (3.7.4), (3.7.5) and (3.7.6), respectively.

strongly dominates that of noise, the under-detection risk of IC would not be negligible. Besides, in this case, even after we adjust a weight for the penalty, a resulting performance may not be significantly improved. For example, when we achieve a minimum bound at the optimal weight, an upper bound is still over 60% in the right panels and over 90% in the left

panels.

3.8 Concluding Remarks

This study builds on our earlier work, Kao and Oh (2017) or Chapter 2 in this dissertation, which studied the over-detection risk of the IC estimator proposed by Bai and Ng (2002) and proposed a practical method to reduce its over-detection probability in finite samples. In this chapter, we extend the previous results to the under-detection risk of the IC estimator so that we formulate an upper bound on the *overall* misdetection probability and finally find the optimal penalty function of the information criteria to minimize a misdetection probability bound in finite samples.

Recent results from random matrix theory still play a key role in this chapter. For this reason, our theoretical results hold under certain (somewhat idealistic) conditions which are required to apply random matrix theory to this chapter. Regretfully, a phase transition phenomenon concerning the limiting distribution of the least influential signal eigenvalue is currently available only for the i.i.d. samples and the case of $n > p$. Also, the limiting behavior of the largest pure noise eigenvalue is only known for the case with homogeneous uncorrelated noise.

In this regard, there remain interesting extensions for future research. Obviously, one of topics is to extend our result to more general settings such as heterogeneous factors and unknown noise structure, and to the data with $p > n$. Another interesting topic is to study our topics regarding the situation when the true number of factors increases with the sample size. Lastly, we remark that our approach introduced in this chapter can also be applied to general model selection criteria for detecting the number of factors models.

Appendix

A.3.1. Proof of Theorem 3.1

Proof. See Appendix A.2.3 of Chapter 2. □

A.3.2. Proof of Theorem 3.2

Proof. Recall that the true number of factors for the case with under-detection by one factor is denoted by r_u. For simplicity, here we omit the subscript u. In Chapter 2, the proof of Theorem 2.1 (Appendix A.2.3) shows that $\frac{1}{p-r}\widetilde{T}_{p-r} = O_p(1)$, $\frac{M(r)}{n}$ is negligible and $\frac{\sqrt{r}}{n}\frac{1}{\widetilde{T}_{p-r}}\sum_{j=r+1}^{p}\widetilde{\ell}_j Z_j$ is sufficiently small for large n; that is, $O_p\left(\frac{\sqrt{r}}{n}\right)$. Moreover, from Lemma 2.3 in Chapter 2, we get $T_{p-r} = \widetilde{T}_{p-r}\left(1 + O_p\left(\frac{\sqrt{r}}{n}\right)\right)$. This result shows T_{p-r} approximates the trace of a Wishart matrix with identity covariance matrix, $Tr(W)$, up to $o_p(1/n)$ error term. □

A.3.3. Proof of Theorem 3.3

Proof. See Appendix A.2.4 of Chapter 2. □

A.3.4. Proof of Theorem 3.4

Part 1.

Proof. For simplicity, here we omit the subscript u in r_u. Consider the average of the sample eigenvalues of a $(p-r)\times(p-r)$ Wishart matrix, W. Then, $\frac{Tr(W)}{p-r} = \frac{\sum_{j=r+1}^{p}\widetilde{\ell}_j}{p-r} \sim \frac{\chi^2_{n(p-r)}}{n(p-r)}$ (see Footnote 2). Let s be some positive number. Then we can write

$$\Pr(\triangle IC(-1) < 0) = \Pr\left(\triangle IC(-1) < 0 \bigcap \frac{Tr(W)}{p-r} < 1 + \frac{s}{\sqrt{n}}\right) + \Pr\left(\triangle IC(-1) < 0 \bigcap \frac{Tr(W)}{p-r} \geq 1 + \frac{s}{\sqrt{n}}\right) .$$

Also, by Theorem 3.2, we obtain the following inequality:

$$\begin{aligned}\Pr\left(\triangle IC(-1)<0\right) &\leq \Pr\left(\frac{\ell_r}{\widetilde{T}_{p-r}} < \vartheta_{n,p} \bigcap \frac{\widetilde{T}_{p-r}}{p-r} < 1+\frac{s}{\sqrt{n}}\right) + \Pr\left(\frac{\chi^2_{n(p-r)}}{n(p-r)} \geq 1+\frac{s}{\sqrt{n}}\right) \\ &\leq \Pr\left(\ell_r < (p-r)\left(1+\frac{s}{\sqrt{n}}\right)\vartheta_{n,p}\right) + \Pr\left(\frac{\chi^2_{n(p-r)}}{n(p-r)} \geq 1+\frac{s}{\sqrt{n}}\right) \\ &= I + II\ .\end{aligned}$$

□

Part 2.

Proof. Using the following lemma regarding a Chi-squared inequality (Johnstone and Lu, 2009, Appendix, A.2), the upper bound of II in part 1 can be obtained as follows.

Lemma 3.3. *(Johnstone and Lu, 2009)*

$$\Pr\left(\chi^2_v \geq v(1+\epsilon)\right) \leq \exp\left(-\frac{3v\epsilon^2}{16}\right), \quad 0 \leq \epsilon < 1/2.$$

Thus, setting $v = n(p-k)$ and $\varepsilon = \frac{s}{\sqrt{n}}$, we get

$$II = \Pr\left(\chi^2_{n(p-r)} \geq n(p-r)\left(1+\frac{s}{\sqrt{n}}\right)\right) \leq \exp\left(\frac{-3(p-r)s^2}{16}\right),$$

for $s \in [0, 2\sqrt{n}\ )$ since Lemma 3.3 holds when $\epsilon \in [0, 1/2)$. □

Part 3.

Proof. Now, let us derive the upper bound of I in part 1. By Remark 3.3, the ℓ_r asymptotically follows the Gaussian distribution as $n, p \to \infty$. that is, $\frac{\sqrt{n}(\ell_r - \pi(\nu_r))}{\sigma(\nu_r)} \xrightarrow{d} \mathcal{N}(0,1)$.

Thus,

$$\begin{aligned} I &= \Pr\left(\ell_r < (p-r)\left(1+\frac{s}{\sqrt{n}}\right)\vartheta_{n,p}\right) \\ &= \Pr\left(N(0,1) < \frac{\sqrt{n}}{\sigma(\nu_r)}\left((p-r)\left(1+\frac{s}{\sqrt{n}}\right)\vartheta_{n,p} - \pi(\nu_r)\right)\right) \\ &= \Pr\left(N(0,1) < z\right) = \Phi(z), \end{aligned}$$

where Φ denotes the standard Gaussian density function.

Consider the following result regarding inequalities for Mills' ratio $(1-\Phi)/\phi$, where ϕ denote the standard Gaussian distribution function, $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$.

Lemma 3.4. *(Birnbaum, 1942; Komatu, 1955)*

$$\frac{2\phi(x)}{\sqrt{4+x^2}+x} \leq 1-\Phi(x) \leq \frac{2\phi(x)}{\sqrt{2+x^2}+x} \quad \textit{for} \quad x \geq 0.$$

By the above lemma, if $z \geq 0$, then

$$\Phi(z) \leq 1 - \frac{2\phi(z)}{\sqrt{4+z^2}+z}\,.$$

On the other hand, when $z < 0$,

$$\Phi(z) \leq \frac{2\phi(-z)}{\sqrt{2+z^2}-z}\,.$$

□

References

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

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

Baik, J., G. Ben Arous, and S. Péché (2005), "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," Annals of Probability, *33*, 1643–1697.

Baik, J., and J. W. Silverstein (2006), "Eigenvalues of large sample covariance matrices of spiked population models," Journal of Multivariate Analysis, *97*, 1382–1408.

Birnbaum, Z. W. (1942), "An inequality for Mills ratio," Annals of Mathematical Statistics, *13*, 245–246.

Byun, S. J., and L. Schmidt (2016), "Real risk or paper risk: Mis-measured factors, granular measurement errors, and empirical asset pricing tests," Unpublished manuscript.

Choi, I., and H. Jeong (2013), "Model selection for factor analysis: Some new criteria and performance comparisons," Research Institute for Market Economy (RIME) Working Paper No.1209, Sogang University.

Geman, S. (1980), "A limit theorem for the norm of random matrices," Annals of Probability, *8*, 252–261.

Greenaway-McGrevy, R., C. Han, and D. Sul (2012), "Estimating the number of common factors in serially dependent approximate factor models," Economics Letters, *116*, 531–534.

Harding, M. (2013), "Estimating the number of factors in large dimensional factor models," preprint.

Hallin, M., and R. Liška (2007), "Determining the number of factors in the general dynamic factor model," Journal of the American Statistical Association, *102*, 603–617.

Johnstone, I. M. (2001), "On the distribution of the largest eigenvalue in principal components analysis," Annals of Statistics, *29*, 295–327.

Johnstone, I. M., and A. Y. Lu (2009), "On consistency and sparsity for principal components analysis in high dimensions," Journal of the American Statistical Association, *104*, 682–693.

Karoui, N. E. (2008), "On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity," arXiv preprint math/0309355.

Kao, C., and J. Oh (2017), "On the Over-detection Probability of the Number of Factors," Unpublished manuscript.

Komatu, Y. (1955), "Elementary inequalities for Mill's ratio," Reports of Statistical Application Research (Union of Japanese Scientists and Engineers), *4*, 69–70.

Kritchman, S., and B. Nadler (2009), "Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory," Signal Processing, IEEE Transactions on, *57*, 3930–3941.

Ledoux, M. (2007), "Deviation inequalities on largest eigenvalues, in geometric aspects of functional analysis," In Milman, V. D., and G. Schechtman (Eds.), Lecture Notes in Mathematics. New York: Springer, 2007, vol. 1910.

Ma, Z. (2012), "Accuracy of the Tracy–Widom limits for the extreme eigenvalues in white Wishart matrices," Bernoulli, *18*, 322–359.

Moon, H. R., and M. Weidner (2015), "Linear regression for panel with unknown number of factors as interactive fixed effects," Econometrica, *83*, 1543–1579.

Nadler, B. (2008), "Finite sample approximation results for principal component analysis: A matrix perturbation approach," Annals of Statistics, *36*, 2791–2817.

Nadler, B. (2010), "Nonparametric detection of signals by information theoretic criteria: Performance analysis and an improved estimator," Signal Processing, IEEE Transactions on, *58*, 2746–2756.

Nadler, B. (2011), "On the distribution of the ratio of the largest eigenvalue to the trace of a wishart matrix," Journal of Multivariate Analysis, *102*, 363–371.

O'leary, D. P., and G. W. Stewart (1990),"Computing the eigenvalues and eigenvectors of symmetric arrowhead matrices, " Journal of Computational Physics, *90*, 497–505.

Onatski, A. (2005), "Determining the number of factors from empirical distribution of eigenvalues," Columbia University Discussion Paper No. 0405–19.

Onatski, A. (2007), "Asymptotics of the principal components estimator of large factor models with weak factors and i.i.d. Gaussian noise," Manuscript, University of Cambridge.

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

Onatski, A. (2012), "Asymptotics of the principal components estimator of large factor models with weakly influential factors," Journal of Econometrics, *168*, 244–258.

Onatski, A. (2015), "Asymptotic analysis of the squared estimation error in misspecified factor models," Journal of Econometrics, *186*, 388–406.

Paul, D. (2007), "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," Statistica Sinica, *17*, 1617–1642.

Rao, C. (1973), Linear Statistical Inference and Its Applications (2nd ed.). New York: Wiley.

Silverstein, J. W. (1985), "The smallest eigenvalue of a large dimensional Wishart matrix," Annals of Probability, *13*, 1364–1368.

Tracy, C. A., and H. Widom (1996), "On orthogonal and symplectic matrix ensembles," Communications in Mathematical Physics, *177*, 727–754.

Chapter 4

Misspecified Recovery and Recovery of the Long-term Risk: Evidence from the Gaussian Affine Term Structure

4.1 Introduction

This paper examines the applicability of the Recovery theorem proposed by Ross (2015) to fixed-income markets in the framework of an affine Gaussian dynamic term structure model, and further explores the issue of what the Recovery theorem actually recovers. The Recovery theorem claimed that the investors' true expectations (or equivalently, the physical probability distribution of stock returns) can be recovered from only state prices without pre-specifying any parameters for risk aversion, and consequently the stochastic discount factor (SDF), which captures an agent's risk aversion, can be identified simultaneously.

Ross's claim has been followed by numerous theoretical extensions and empirical applications to equity markets (e.g., Carr and Yu, 2012; Tsui, 2013; Spears, 2013; Martin and Ross, 2013; Tran and Xia, 2014; Audrino, Huitema, and Ludwig, 2015; Walden, 2017). To the best of our knowledge, however, there are only a few studies on its application to fixed-income markets (Aydin and Yildirim, 2015; Qin, Linetsky, and Nie, 2016).

For equity markets, the Recovery theorem is appealing. As Ross (2015) mentioned, there has been a theoretical hurdle to using market prices to forecast future asset returns. To identify the physical probability distribution of future returns from asset prices, we need to specify investors' risk aversion embedded in the SDF since any asset is priced by the risk-neutral probability measure which absorbs risk aversion; however, the agent's risk aversion is not directly observable. For this reason, existing studies have specified the physical probability distribution by imposing parameter-restrictions on risk aversion, or they have forecasted asset returns by using historical market returns or survey data. In contrast, Ross (2015) develops a theory of how to infer the physical probability from the risk-neutral probability, without placing restrictions on risk aversion.

For fixed-income markets, on the other hand, there exists a large literature on estimation for investors' interest rates expectations under the physical probability measure from zero-coupon bond prices. Especially when we consider an affine term structure model, various estimation methods have been provided relying on model specifications. For example,

Kalman filter estimation is available when the state variables are unobservable. Also, simulated maximum likelihood or quasi-maximum likelihood can be employed when the likelihood function is unknown (Piazzesi, 2010; Duffee and Stanton, 2012). In a Gaussian framework, a standard maximum likelihood estimation is feasible (Joslin, Singleton, and Zhu, 2011; Wright, 2011; Bauer, Rudebusch, and Wu, 2012, 2014).

In the above estimation procedures, it is well known that highly persistent interest rates lead to a critical identification issue, *small-sample bias* (Kim and Orphanides, 2012; Bauer et al., 2012; Bauer, 2016). When the sample size is small, the mean reversion coefficient in the state dynamics under the physical probability measure tends to be over estimated. Much of the literature has dealt with this issue. For example, Kim and Orphanides (2012) used survey data, whereas Joslin, Priebsch, and Singleton (2014) imposed parameter restrictions on risk aversion. Also, Bauer et al. (2012) proposed a statistical method for correcting bias. But still, how to precisely estimate the physical probability in affine term structures is an ongoing issue. Hence, it is worth considering the Recovery theorem as a different identification approach for fixed-income markets.

The results of Ross (2015), if true, could be attributed to the future information contained in state prices; that is, investors' expectations on future interest rates across different possible states. The state price is the price in the current state of the Arrow-Debreu security that pays off a dollar for sure if a certain state is realized in the next period. In this sense, we may hypothesize (as Ross did) that if the state prices are fully identified even for unrealized states, such additional future (and also cross-sectional) information helps identify the investors' true beliefs.

Another group of articles, however, argues that Ross recovered something different from the physical probability measure (Borovička, Hansen, and Scheinkman, 2015; Bakshi, Chabi-Yo, and Gao, 2015; Qin and Linetsky, 2016). This claim is based on theoretical results from the literature on the SDF decomposition (e.g., Alvarez and Jermann, 2005; Hansen and Scheinkman, 2009; Hansen, 2012; Bakshi and Chabi-Yo, 2012). By extracting a martingale

component, which represents risk aversion to permanent shocks, from the SDF, the authors found that the Recovery theorem can recover the physical probability only when a martingale component is one. They also showed, however, that such a degenerating martingale is implausible both theoretically and empirically. In particular, Borovička, Hansen, and Scheinkman (2015, hereafter *BHS*) referred to this claim as "misspecified recovery." Also, BHS (2015) identified the probability measure recovered by Ross (2015) as another risk-adjusted probability measure which absorbs risk compensation for exposure to only permanent shocks, and referred to it as the *long-term risk-neutral probability measure*.

The contributions of this paper are as follows. First, we show how to implement the Recovery theorem in an affine Gaussian dynamic term structure model (hereafter *GDTSM*). We use a finite-state Markov-chain approximation method developed by Gospodinov and Lkhagvasuren (2014) to construct state prices and the risk-neutral state transition probabilities. We then recover a certain probability measure (called the *recovered probability measure*) by the Perron-Frobenius theorem. In addition, we estimate a GDTSM and further decompose forward rates into interest rate expectations and term premia under the recovered probability measure. Note that while this paper was being prepared, we were aware that Aydin and Yildirim (2015) had applied the Recovery theorem to a GDTSM with the US data; however, this paper uses an international panel dataset (10 countries) and our procedure is robust to the highly persistent factors and the number of states.

Second, we find empirical evidence that the Recovery theorem infers the long-term risk-neutral probability while misspecifying the physical probability as claimed in BHS (2015). Our approach is distinguished from the previous research such as Alvarez and Jermann (2005, hereafter *AJ*) which studied the variance bound on the martingale component of the SDF. This paper instead formulates a condition for equality between the physical and recovered probabilities in terms of forward term premia as well as the market prices of risk. In detail, by using the SDF decomposition and the change of measure, we specify the connection of term premia under the physical probability measure and the recovered probability measure, and

estimate term premia (and risk prices as well) under each probability measure so that we can directly compare those estimates. Consequently, we find that "misspecified recovery" can be rejected only if term premia regarding permanent shocks are zero so that term premia under the physical measure equal those under the long-term risk-neutral measure. Our empirical results showed, however, term premia corresponding to permanent shocks (referred to as long-term risk premia) are substantially different from zero.

There are additional findings. Term premia and interest rates expectations under the recovered long-term risk-neutral probability measure are very similar to those under the risk-neutral measure. This empirical similarity supports theoretical predictions in BHS (2015) and Qin et al. (2016). Next, by using the decomposition of forward rates under each probability measure, we finally decompose overall term premia into nearly constant short-term risk premia corresponding to transitory shocks and highly volatile long-term risk premia associated with permanent shocks. Correspondingly, we find that the secular downward trend and volatility of forward rates are mostly attributed to investors' interest rate expectations under the long-term risk-neutral probability measure, and all important variations in overall term premia are captured by long-term risk premia. Concisely, long-term risk matters for asset pricing.

The rest of the paper is organized as follows. Section 4.2 delineates our GDTSM and summarizes how standard GDTSM analysis identifies the physical and risk-neutral probability measures. By following a conventional method, we analyze our GDTSM to provide a benchmark against which the reliability of the Recovery theorem can be tested. In Section 4.3, after reviewing the Recovery theorem of Ross (2015), we show how to apply it to a GDTSM. Section 4.4 investigates the misspecification issue of the Recovery theorem. In Section 4.5, we conduct empirical studies to recover the probability measure and to analyze our GDTSM under the recovered probability measure. As a result, we provide empirical evidence on "misspecified recovery" and decompose term premia into the long-term and short-term components. The implications of long-term risk premia are also examined. Section 4.6 is

summary and discussion.

A word on notation. The transpose operator is denoted by a prime symbol as in A'. $x \sim D$ means that a random variable x has the probability distribution D. The Gaussian distribution with mean μ and covariance σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2)$. *i.i.d.* means that a random variable is independent and identically distributed. The remaining notations and symbols are defined in the body of the paper.

4.2 Term Structure Model and Estimation

4.2.1 Model Specification

This paper studies an affine Gaussian dynamic term structure model in a discrete-time framework developed by Ang and Piazzesi (2003) which is subsumed under the admissible affine class (Duffie and Kan, 1996; Dai and Singleton, 2000). We consider only yields as the state variables, whereas Ang and Piazzesi (2003) combined macro-economic variables with yield factors. Eventually, bond prices, yields and forward rates are all affine in yield factors, and the prices of risk are time-varying.

4.2.1.1 Affine Gaussian Dynamic Term Structure

First, we set up the state dynamics. Let X_t denote an N-dimensional vector of unobservable state variables; $X_t = (X_{t,1}, \ldots, X_{t,N})'$. Suppose that X_t follows a Gaussian VAR(1) process under the physical probability measure denoted by $\mathbb{P}$. We then write the $\mathbb{P}$ state dynamics as follows:

$$X_{t+1} = \mu + \Phi X_t + \Sigma \epsilon_{t+1}, \tag{4.2.1}$$

where μ is an $N \times 1$ vector, Φ is an $N \times N$ matrix, an $N \times 1$ vector $\epsilon_t \sim \mathcal{N}(0, I_N)$, and Σ is an $N \times N$ lower triangular matrix such that $\Sigma\Sigma' = V$.

Next, one-period interest rates denoted by r_t are assumed to be affine in all latent state

variables; hence, a short rate equation is defined as

$$r_t = \delta_0 + \delta_1^{'} X_t, \tag{4.2.2}$$

where δ_0 is a scalar and δ_1 is an N-vector. An observable short interest rate r_t is thought of as the one-period yield denoted by $y_t^{(1)}$.

Third, as the standard results from much of the literature, the SDF is defined as

$$\frac{S_{t+1}}{S_t} = \exp\left(-r_t - \frac{1}{2}\lambda_t^{'}\lambda_t - \lambda_t^{'}\epsilon_{t+1}\right), \tag{4.2.3}$$

where an $N \times 1$ vector λ_t denotes the market prices of risk that measure the additional expected return required per unit of risk from each of the shocks in ϵ_t. λ_t is parametrized as the affine process of latent state variables:

$$\lambda_t = \Sigma^{-1}(\lambda_0 + \lambda_1 X_t), \tag{4.2.4}$$

for an $N \times 1$ vector λ_0 and an $N \times N$ matrix λ_1. Our GDTSM assumes a constant Σ. Considering that term premia are the product of the prices of risk (λ_t) and the quantities of risk (Σ), a non-zero matrix λ_1 causes the market prices of risk and term premia to be time-varying. As Piazzesi (2010) pointed out, such a risk-price specification is a special case of the essentially affine class defined by Duffee (2002) which allows *maximal flexibility* to the prices of risk (i.e., no restriction on λ_0 and λ_1) so that a risk price varies independently of a factor volatility.

A key restriction behind the SDF is the no-arbitrage assumption that guarantees the existence of an equivalent martingale measure (or equivalently the risk-neutral measure) denoted by $\mathbb{Q}$. Suppose that we have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a sample space and $\mathcal{F}$ is a set of events, and a filtration $\mathcal{F}_t$ defined for $0 \leq t \leq T$, where T is a fixed final time. Further, consider a nonnegative random variable ξ satisfying $E(\xi) = 1$, where E

denotes the expectation under the $\mathbb{P}$ measure. We then define the $\mathbb{Q}$ measure as

$$\mathbb{Q}(A) = \int_A \xi(\alpha) d\mathbb{P}(\alpha) \quad \text{for all} \quad A \in \mathcal{F}. \tag{4.2.5}$$

Here ξ converts the $\mathbb{P}$ measure to the $\mathbb{Q}$ measure such that $E^{\mathbb{Q}}(Z) = E(Z\xi)$ for any random variable Z, where $E^{\mathbb{Q}}$ denotes the expectation under the $\mathbb{Q}$ measure. In the literature, ξ is referred to as the *Radon-Nikodym derivative* of $\mathbb{Q}$ with respect to $\mathbb{P}$ and written as $\xi = d\mathbb{Q}/d\mathbb{P}$. Also, the Radon-Nikodym derivative process is defined as $\xi_t = E(\xi|\mathcal{F}_t)$ which is a martingale, or simply $\xi_t = E_t(\xi)$, where E_t denotes the conditional expectation under the $\mathbb{P}$ measure. Correspondingly, we have $E_t^{\mathbb{Q}}(Z_{t+1}) = E_t(\xi_{t+1}Z_{t+1})/\xi_t$, where $E_t^{\mathbb{Q}}$ denotes the conditional expectation under the $\mathbb{Q}$ measure.

Under the $\mathbb{Q}$ measure, the price of any asset (V_t), which does not pay any dividends at time $t+1$, satisfies $V_t = E_t^{\mathbb{Q}}(\exp(-r_t)V_{t+1})$; that is, asset prices are the expected values of their payoffs discounted at the riskless rate, where the conditional expectation is computed by the $\mathbb{Q}$ measure. Suppose that ξ_{t+1} follows the log-normal process:

$$\log \xi_{t+1} = \log \xi_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t'\epsilon_{t+1}, \tag{4.2.6}$$

and define the SDF as $S_{t+1}/S_t = \exp(-r_t)\xi_{t+1}/\xi_t$. Substituting (4.2.2) for r_t, we can obtain (4.2.3). Under the $\mathbb{Q}$ measure, the price of a τ-period zero-coupon bond at time t is

$$P_t^{(\tau)} = E_t\left(\frac{S_{t+1}}{S_t} \cdot P_{t+1}^{(\tau-1)}\right) = E_t\left(e^{-r_t} \cdot \frac{\xi_{t+1}}{\xi_t} \cdot P_{t+1}^{(\tau-1)}\right) = E_t^{\mathbb{Q}}\left(e^{-\sum_{h=0}^{\tau-1} r_{t+h}}\right). \tag{4.2.7}$$

By our risk-price specification (4.2.4), the dynamics of latent state variables under the $\mathbb{Q}$ measure (referred to as the $\mathbb{Q}$ *dynamics*) also follows a Gaussian VAR(1) process:

$$X_{t+1} = \mu^{\mathbb{Q}} + \Phi^{\mathbb{Q}} X_t + \Sigma\epsilon_{t+1}^{\mathbb{Q}}, \tag{4.2.8}$$

where $\mu^{\mathbb{Q}} = \mu - \lambda_0$, $\Phi^{\mathbb{Q}} = \Phi - \lambda_1$, and $\epsilon_t^{\mathbb{Q}} \sim \mathcal{N}(0, I_N)$ (for details, see Appendix A.4.2). Obviously, when λ_t is a vector of zeros for all t, the $\mathbb{P}$ and $\mathbb{Q}$ measures are identical. Note that the state vector X_t follows a time-homogeneous stationary Markov process under the $\mathbb{Q}$ measure. The stationarity assumption corresponds well with empirical properties of the yield curve (Bauer et al., 2012, p. 457).

As Duffie and Kan (1996) showed, the state dynamics (4.2.1) with a risk-price specification (4.2.4), a short rate equation (4.2.2), and the Radon-Nikodym derivative (4.2.5) together form an affine Gaussian dynamic term structure with N latent factors, and consequently model-implied bond prices are exponential affine functions of the state variables:

$$P_t^{(\tau)} = \exp\left(\bar{A}_\tau + \bar{B}_\tau' X_t\right), \tag{4.2.9}$$

where loadings (a constant $\bar{A}_\tau$ and an $N \times 1$ vector $\bar{B}_\tau$) follow the difference equations:

$$\bar{A}_{\tau+1} = \bar{A}_\tau + \bar{B}_\tau' \mu^{\mathbb{Q}} + \frac{1}{2}\bar{B}_\tau' \Sigma\Sigma' \bar{B}_\tau - \delta_0, \qquad \bar{B}_{\tau+1}' = \bar{B}_\tau' \Phi^{\mathbb{Q}} - \delta_1', \tag{4.2.10}$$

with $\bar{A}_0 = 0$ and $\bar{B}_0 = 0$ so that $\bar{A}_\tau = \bar{A}_\tau(\mu^{\mathbb{Q}}, \Phi^{\mathbb{Q}}, \delta_0, \delta_1, \Sigma)$ and $\bar{B}_\tau = \bar{B}_\tau(\Phi^{\mathbb{Q}}, \delta_1)$. This implies that for determining loadings and bond pricing, only the $\mathbb{Q}$ dynamics matters. For the derivation of difference equations, see Cochrane and Piazzesi (2005). Similarly, the continuously compounded yield on a τ-period zero-coupon bond at time t is also affine in X_t:

$$y_t^{(\tau)} = -\frac{1}{\tau}\log P_t^{(\tau)} = A_\tau + B_\tau' X_t, \tag{4.2.11}$$

where $A_\tau = -\bar{A}_\tau/\tau$ and $B_\tau = -\bar{B}_\tau/\tau$ so that $A_\tau = A_\tau(\mu^{\mathbb{Q}}, \Phi^{\mathbb{Q}}, \delta_0, \delta_1, \Sigma)$ and $B_\tau = B_\tau(\Phi^{\mathbb{Q}}, \delta_1)$. Again, loadings only depend on parameter estimates and the error covariance V in the $\mathbb{Q}$ dynamics of the state variables. We also write yield equations (4.2.11) for n different maturities as the following n-dimensional vector form. Letting $(\tau_1, \tau_2, \ldots, \tau_n)$ denote the set of fixed maturities such that $N < n$ and $y_t = (y_t^{(\tau_1)}, \ldots, y_t^{(\tau_n)})'$ denote the corresponding set of

yields, we have

$$y_t = A + BX_t, \tag{4.2.12}$$

where an $n \times 1$ vector $A = (A_{\tau_1}, \ldots, A_{\tau_n})'$, and an $n \times N$ matrix $B = (B_{\tau_1}, \ldots, B_{\tau_n})'$. Moreover, the log forward rates at time t for loans starting at $t + \tau_j$ and maturing at $t + \tau_k$ is given by

$$f_t^{(\tau_j,\tau_k)} = -\frac{1}{\tau_k - \tau_j}\left(\log P_t^{(\tau_j)} - \log P_t^{(\tau_k)}\right) = \frac{1}{\tau_k - \tau_j}\left(\tau_k \cdot y_t^{(\tau_k)} - \tau_j \cdot y_t^{(\tau_j)}\right). \tag{4.2.13}$$

As long as we are not living in a risk-neutral world, λ_t is not a zero vector and $\mathbb{P} \neq \mathbb{Q}$ so that bond yields should include premia to compensate risk-averse investors for exposure to risk such as uncertainty about future inflation which may erode the value of nominal bonds. Such term premia (ytp_t) are hence defined as the difference between the risk-adjusted yields (y_t) and the hypothetical yields ($\widetilde{y}_t$) that would prevail if investors were risk-neutral. That is,

$$ytp_t^{(\tau)} = y_t^{(\tau)} - \widetilde{y}_t^{(\tau)}. \tag{4.2.14}$$

As in (4.2.11), $y_t^{(\tau)}$ is measured by the risk-neutral probability measure. In the literature, $\widetilde{y}_t^{(\tau)}$ is often referred to as *risk-neutral rates* as if $\mathbb{P} = \mathbb{Q}$. Following Bauer et al. (2012) and Bauer (2016), risk-neutral rates can be calculated by using parameter estimates for the $\mathbb{P}$ state dynamics:

$$\widetilde{y}_t^{(\tau)} = \widetilde{A}_\tau + \widetilde{B}_\tau' X_t, \quad \widetilde{A}_\tau = -\frac{1}{\tau}A_\tau(\mu, \Phi, \delta_0, \delta_1, \Sigma), \quad \widetilde{B}_\tau = -\frac{1}{\tau}B_\tau(\Phi, \delta_1). \tag{4.2.15}$$

Put differently, $ytp_t^{(\tau)} = y_t^{(\tau)} - \frac{1}{\tau}\sum_{h=0}^{\tau-1} E_t y_{t+h}^{(1)} - \textit{Jensen's inequality term}$ (Cochrane, 2009). Since the Jensen's term is modest at maturities of ten years or less, risk-neutral rates can be closely approximated by the average of short-term interest rate expectations over the life of the bond; that is, $\frac{1}{\tau}\sum_{h=0}^{\tau-1} E_t y_{t+h}^{(1)}$, where the expectation is computed by the $\mathbb{P}$ measure (Piazzesi, 2010; Gürkaynak and Wright, 2012). In this sense, $\widetilde{y}_t^{(\tau)}$ is also referred to as the

short-rate expectations under $\mathbb{P}$. It reflects investors' expectations about real interest rates and inflation (Wright, 2011).

Similarly, the τ_j- to τ_k-year forward term premia (ftp_t) are defined as differences between far-ahead forward rates (f_t) and risk-neutral forward rates ($\widetilde{f}_t$):

$$ftp_t^{(\tau_j,\tau_k)} = f_t^{(\tau_j,\tau_k)} - \widetilde{f}_t^{(\tau_j,\tau_k)}, \tag{4.2.16}$$

where $\widetilde{f}_t^{(\tau_j,\tau_k)} = \frac{1}{\tau_k-\tau_j}\left(\tau_k\widetilde{y}_t^{(\tau_k)} - \tau_j\widetilde{y}_t^{(\tau_j)}\right)$.

4.2.1.2 GDTSM with observable yield factors: JSZ representation

The state variables (X_t) are not directly observed; however, they can be inferred from observable yields. For example, as Duffie and Kan (1996) proposed, we can take yields themselves as latent factors by simply inverting the linear relationship (4.2.12). We adopt a different approach to this paper as in Joslin et al. (2011). They developed the JSZ representation of a canonical GDTSM where factors are represented as the first N principal components of yields such that $N < n$. These observable yield factors are denoted by $\mathcal{P}_t$ and follow a VAR(1) process. First, recall the dynamics of the latent state variables. In mean-reverting process forms, we can rewrite (4.2.1) and (4.2.8) as

$$\Delta X_{t+1} = \mu + KX_t + \Sigma\epsilon_{t+1}, \tag{4.2.17}$$

$$\Delta X_{t+1} = \mu^{\mathbb{Q}} + K^{\mathbb{Q}}X_t + \Sigma\epsilon_{t+1}^{\mathbb{Q}}, \tag{4.2.18}$$

$$r_t = \delta_0 + \delta_1^{'}X_t, \tag{4.2.19}$$

where $K = \Phi - I_N$, $K^{\mathbb{Q}} = \Phi^{\mathbb{Q}} - I_N$ and the model is stationary under the $\mathbb{Q}$ measure. By allowing measurement errors, the observed yields take the following form as

$$y_t^{(\tau)^o} = y_t^{(\tau)} + \varepsilon_t = A_\tau(\Theta^{\mathbb{Q}}) + B_\tau(\Theta^{\mathbb{Q}})^{'}X_t + \varepsilon_t, \tag{4.2.20}$$

where $y_t^{(\tau)^o}$ with the superscript 'o' denotes observed yields, $y_t^{(\tau)}$ denotes model implied yields, $\Theta^{\mathbb{Q}} = (\mu^{\mathbb{Q}}, K^{\mathbb{Q}}, \Sigma, \delta_0, \delta_1)$ is the set of parameters relevant for a bond pricing, and ε_t denotes measurement errors with the conditional normal distribution P^{θ_τ} for some $\theta_\tau \in \Theta_\tau$ and independent of X_t.

Now, we replace latent factors (X_t) by observed yield factors ($\mathcal{P}_t$). Suppose that $\mathcal{P}_t \equiv W' y_t$ for an $n \times N$ matrix W with full rank N. Denote by W_n an $n \times n$ orthogonal matrix whose columns are standardized eigenvectors of the matrix $\text{Var}(y_t)$. W becomes its submatrix with the first N eigenvectors, and $\mathcal{P}_t$ is the first N principal components of yields. As long as $\mathcal{P}_t$ is measured without error, the JSZ representation has the following unique and observationally equivalent representation to (4.2.17), (4.2.18) and (4.2.19) (Joslin et al., 2011, Theorem 1):

$$\Delta \mathcal{P}_{t+1} = \mu_{\mathcal{P}} + K_{\mathcal{P}} \mathcal{P}_t + \Sigma_{\mathcal{P}} \epsilon_{t+1}, \tag{4.2.21}$$

$$\Delta \mathcal{P}_{t+1} = \mu_{\mathcal{P}}^{\mathbb{Q}} + K_{\mathcal{P}}^{\mathbb{Q}} \mathcal{P}_t + \Sigma_{\mathcal{P}} \epsilon_{t+1}^{\mathbb{Q}}, \tag{4.2.22}$$

$$r_t = \rho_0 + \rho_1^{'} \mathcal{P}_t, \tag{4.2.23}$$

where $\Sigma_{\mathcal{P}} = (W' B \Sigma \Sigma' B' W)^{1/2}$. The parameter space of the $\mathbb{P}$ dynamics is $\Theta_{\mathcal{P}}^{\mathbb{P}} \equiv (\mu_{\mathcal{P}}, K_{\mathcal{P}}, \Sigma_{\mathcal{P}})$. Meanwhile, Joslin et al. (2011, Proposition 2) showed that $(\mu_{\mathcal{P}}^{\mathbb{Q}}, K_{\mathcal{P}}^{\mathbb{Q}}, \rho_0, \rho_1)$ are functions of the following $\mathbb{Q}$ parameters: (i) $r_\infty^{\mathbb{Q}}$, the long-run mean of short rates, (ii) $\phi^{\mathbb{Q}}$, the eigenvalues of $\Phi^{\mathbb{Q}} = K^{\mathbb{Q}} + I_N$. Thus, the parameters of the $\mathbb{Q}$ dynamics of $\mathcal{P}_t$ are fully characterized by $\Theta_{\mathcal{P}}^{\mathbb{Q}} \equiv (\phi^{\mathbb{Q}}, r_\infty^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$. To sum up, the JSZ representation is parametrized by $\Theta_{\mathcal{P}} \equiv (\phi^{\mathbb{Q}}, r_\infty^{\mathbb{Q}}, \mu_{\mathcal{P}}, K_{\mathcal{P}}, \Sigma_{\mathcal{P}})$.

From $\mathcal{P}_t \equiv W' y_t$ and (4.2.12), we have

$$\mathcal{P}_t = A_W(\Theta^{\mathbb{Q}}) + B_W(\Theta^{\mathbb{Q}}) X_t, \tag{4.2.24}$$

where an $N \times 1$ vector $A_W = W^{'}(A_{\tau_1}, \ldots, A_{\tau_n})^{'}$ and an $N \times N$ matrix $B_W = W^{'}(B_{\tau_1}, \ldots, B_{\tau_n})^{'}$. Assume that B_W is invertible so that $\mathcal{P}_t$ contains the same information as X_t. Even after

the change of variables, a short rate and a bond price are unchanged. This is called the *invariant transform* by Dai and Singleton (2000). Now, we can express yields as an affine function of $\mathcal{P}_t$ as

$$y_t = A_{\mathcal{P}}(\Theta^{\mathbb{Q}}, W) + B_{\mathcal{P}}(\Theta^{\mathbb{Q}}, W)\mathcal{P}_t, \tag{4.2.25}$$

where $A_{\mathcal{P}} = (I_N - B(W'B)^{-1}W')A$ and $B_{\mathcal{P}} = B(W'B)^{-1}$. These loadings satisfy $W'A_{\mathcal{P}} = 0$ and $W'B_{\mathcal{P}} = I_N$ so that the yields coming out of the model are identical to those going into the model as the state variables. This is called the *internal consistency* by Duffee (2011).

Lastly, we impose normalizations for econometric identification as in Joslin et al. (2011). Under the $\mathbb{Q}$ stationary process of X_t, (i) $K^{\mathbb{Q}}$ is invertible so that there is no zero eigenvalue and its eigenvalues are real and distinct, and (ii) $\mu^{\mathbb{Q}} = 0$, $\delta_0 = r_\infty^{\mathbb{Q}}$ and $\delta_1 = \iota$ where ι is a vector of ones.

4.2.2 Estimation

4.2.2.1 MLE under the separation property

Since risk factors inferred from yields are now observable and the density of yields is known to be Gaussian, a maximum likelihood (ML) is feasible to estimate the state dynamics and a system of yield equations. As long as yield factors ($\mathcal{P}_t$) are observed without error, the conditional density of observed yields would be factorized into the product of the conditional density of the measurement error of (4.2.20) and the conditional density of $\mathcal{P}_t$ as follows:

$$f(y_t^o|y_{t-1}^o;\Theta) = f(y_t^o|\mathcal{P}_t;\ \phi^{\mathbb{Q}}, r_\infty^{\mathbb{Q}}, \Sigma_{\mathcal{P}}, P^{\theta_n}) \times f(\mathcal{P}_t|\mathcal{P}_{t-1};\ K_{\mathcal{P}}, \mu_{\mathcal{P}}, \Sigma_{\mathcal{P}}). \tag{4.2.26}$$

The first term (referred to as $\mathbb{Q}$ *likelihood*) corresponds to the cross-sectional dependence between yields and yield factors in (4.2.25), while the second term (referred to as $\mathbb{P}$ *likelihood*) is associated with the time series of yield factors in (4.2.21).

Joslin et al. (2011) showed that ordinary least squares (OLS) recovers the ML estimates of the $\mathbb{P}$ likelihood, and the conditional covariance matrix of yield factors ($\Sigma_{\mathcal{P}}$) is independent

of the OLS estimates of $(K_{\mathcal{P}}, \mu_{\mathcal{P}})$. Note that the $\mathbb{P}$ parameters, $(K_{\mathcal{P}}, \mu_{\mathcal{P}})$, are not involved in the $\mathbb{Q}$ likelihood. The $\mathbb{Q}$ likelihood, on the other hand, is determined by the no-arbitrage restriction on cross-sectional relationships among yields. Given N yield factors, hence, the yield curve can be constructed by specifying $(r_\infty^{\mathbb{Q}}, \phi^{\mathbb{Q}}, \Sigma_{\mathcal{P}})$ which are estimated independently of the OLS $\mathbb{P}$ estimates. Joslin et al. (2011) referred to this result as the *separation property*. Such a complete separation between the $\mathbb{P}$ and $\mathbb{Q}$ likelihoods is feasible since the maximally flexible GDTSM does not impose any restriction on the market prices of risk.

This separation property makes estimations much easier. By OLS, we first estimate time series $\mathbb{P}$ parameters $(\mu_{\mathcal{P}}, K_{\mathcal{P}})$, independently of the $\mathbb{Q}$ likelihood. By using these $\mathbb{P}$ estimates as starting values, we obtain the ML estimates of $\mathbb{Q}$ parameters from cross-sectional relationships. Since the $\mathbb{Q}$ likelihood is characterized by a low-dimensional parameter space $(\phi^{\mathbb{Q}}, r_\infty^{\mathbb{Q}})$, the estimation speed in the exact ML can be greatly improved. This estimation method is referred to as *JSZ two-step procedures* hereafter.

4.2.2.2 Bias correction

Although ML estimation is feasible, it suffers from a small-sample bias due to the high persistence of factors, which leads to an upward bias in the estimated mean-reversion process (Bauer et al., 2012, 2014). Actually, much of the literature showed that the first principal component, which is called the level factor, is very persistent. Further, in conventional term structure analyses, a sample length is relatively short due to the concern about structural changes and the zero lower bound of interest rates (Wright, 2011; Bauer, 2016).

As seen in (4.2.15), short-rate expectations are computed by using the parameter estimates of the $\mathbb{P}$ dynamics and therefore inaccurate estimates for the $\mathbb{P}$ parameters falsify the decomposition of forward rates. In small samples, the estimated persistence is much lower than it should be. Short-rate expectations under $\mathbb{P}$ (i.e., risk-neutral rates) quickly revert to mean and hence are too stable over time. Consequently, a secular decline in yields is affected by the behavior of term premia much more than by the behavior of short-rate ex-

pectations. To address this issue, additional information can be considered as a supplement to small samples; for example, the survey forecasts of short-term interest rates as in Kim and Orphanides (2012). Such information, however, is neither always available nor reliable (Bauer et al., 2014). An alternative is to impose restrictions on risk-price parameters so that cross-sectional information can help specify time series of the factor dynamics (Cochrane and Piazzesi, 2009; Joslin et al., 2014). As Bauer (2016) pointed out, however, there is the model uncertainty of how to choose restrictions. Further, Bauer et al. (2012, p. 455) argued that bias is still large even with restrictions on risk prices.

This paper instead considers a statistical method proposed by Bauer et al. (2012) for correcting a small-sample bias. Their method, which is called an indirect inference estimator, can be conducted consistently with *JSZ* two-step procedures. First, they correct bias in the OLS estimates of time series $\mathbb{P}$ parameters. To correct bias in the $\mathbb{P}$ parameters, they find data-generating VAR parameters from repeated simulations which give a mean of the OLS estimator equal to the actual OLS estimates obtained from the data. After that, they obtain the ML estimates of cross-sectional $\mathbb{Q}$ parameters in the normal fashion. It is referred to as *BC two-step procedures* hereafter.

4.2.3 Empirical Study

For later use, we analyze our GDTSM by ML estimation as described above. Observable factors ($\mathcal{P}_t$) are the first three principal components of yields ($N = 3$) and priced without error. Such a three-factor (yields only) GDTSM is common in the literature because the first three principal components explain almost all of the total variation in yields (Litterman and Scheinkman, 1991). As in Joslin et al. (2011), the measurement errors of yields are taken to be an i.i.d. process, and normalizations for identification are also imposed. Our estimations are implemented on two tracks. First, we conduct *JSZ* two-step procedures. Resulting estimates are called *JSZ estimates*. Next, we implement *BC* two-step procedures by using an indirect inference estimator as described in Bauer et al. (2012). Resulting estimates are

called *BC estimates*.

4.2.3.1 Previous Studies

Joslin et al. (2011) estimated the three-factor GDTSM of the US zero-coupon bond yields based on their JSZ representation. They changed latent state variables into observable yield factors which are the linear combinations of yields (yields-only model).[1] In contrast, assuming that the state variables are directly observable, Wright (2011) estimated GDTSMs across ten countries by adding two macro factors: inflation and output growth (macro-factor model). Wright also decomposed forward rates into term premia and risk-neutral rates by using his international panel dataset.

Further, Bauer et al. (2012, 2014) revisited those studies. After correcting bias, they observed that the estimated risk-neutral rates are highly volatile and show distinct downward trends, while Wright (2011) obtained nearly flat risk-neutral rates so that corresponding term premia parallel the fitted forward rates. Decreasing risk-neutral rates corresponds with empirical evidence showing downward trends in the expectations of inflation and the survey forecast of short-rates over time (Wright, 2011; Kim and Orphanides, 2012). In this sense, Bauer et al. (2012, 2014) argued that bias correction yields more plausible implications on the decomposition of forward rates.

4.2.3.2 Data

We use the international panel dataset constructed by Wright (2011). It consists of continuously compounded nominal yields on zero-coupon bonds at maturities from 3 months to 10 years in increments of a quarter across 10 countries: Australia, Canada, Germany, Japan, New Zealand, Norway, Sweden, Switzerland, the United Kingdom, and the United States. Although the original dataset was constructed at a monthly frequency over the

[1] Joslin et al. (2011) considered various specifications depending on how to model the linear combinations of yields in empirical studies (see Ch.5). The first three principal components of yields are one of their specifications.

Table 4.1: Three-factor GDTSM Estimation (US data)

	$\mathbb{P}$					$\mathbb{Q}$				
	$\mu_{\mathcal{P}}$	$\Phi_{\mathcal{P}}$			$eig(\Phi_{\mathcal{P}})$	$r_\infty^{\mathbb{Q}}$	$\phi^{\mathbb{Q}}$	$\Sigma_{\mathcal{P}}$		
	0.0177	0.9402	-0.0194	-0.9163	0.9155	0.0917	0.9710	0.0202	0	0
	(0.0118)	(0.0377)	(0.1348)	(0.7055)		(0.0058)	(0.0077)	(0.0018)		
JSZ	-0.0078	-0.0061	0.9087	1.2268	0.8190		0.9238	0.0031	0.0056	0
	(0.0034)	(0.0111)	(0.0395)	(0.2068)			(0.0158)	(0.0009)	(0.0004)	
	0.0022	0.0099	-0.0088	0.6479	0.7624		0.4347	-0.0016	-0.0003	0.0016
	(0.0014)	(0.0045)	(0.0160)	(0.0837)			(0.0938)	(0.0002)	(0.0002)	(0.0001)
	0.0115	0.9975	-0.0044	-1.0369	0.9862	0.0924	0.9711	0.0209	0	0
	(0.0121)	(0.0389)	(0.1389)	(0.7271)		(0.0060)	(0.0077)	(0.0018)		
BC	-0.0099	0.0011	0.9388	1.1823	0.9094		0.9237	0.0032	0.0056	0
	(0.0035)	(0.0111)	(0.0389)	(0.2083)			(0.0158)	(0.0010)	(0.0005)	
	0.0027	0.0041	-0.0045	0.6881	0.7288		0.4349	-0.0017	-0.0004	0.0016
	(0.0014)	(0.0045)	(0.0162)	(0.0848)			(0.0937)	(0.0002)	(0.0003)	(0.0001)

Note: $\mu_{\mathcal{P}}$, $r_\infty^{\mathbb{Q}}$, and $\Sigma_{\mathcal{P}}$ are reported on an annual basis (by multiplying 4). $\Phi_{\mathcal{P}}$ is $(I_3 + K_{\mathcal{P}})$, where $K_{\mathcal{P}}$ is the mean-reversion coefficient matrix in (4.2.21). $\phi^{\mathbb{Q}}$ here is reported by one plus the ordered eigenvalues of the mean-reversion coefficient matrix; that is $eig(I_3 + K^{\mathbb{Q}})$ in (4.2.18). Asymptotic standard errors for parameters are reported in parentheses on an annual basis (by multiplying 4).

period from January 1990 to May 2009, we use aggregated data at the quarterly frequency from 1Q, 1990 to 1Q, 2009 as in Bauer et al. (2014). Thus, time t is measured in quarters, and short rates are defined as three-month interest rates.

4.2.3.3 Results

For comparison purposes, we replicate most analyses in Wright (2011) and Bauer et al. (2014) on a five-factor (macro-factor) GDTSM with observable state variables. Note that their model specifications are different from our three-factor GDTSM (yields-only model with unobservable state variables); however, our main findings are consistent with their empirical results.

First, Table 4.1 presents the parameter estimates of the state dynamics under both the $\mathbb{P}$ and $\mathbb{Q}$ measures for the US data. The coefficient estimates show that there is a very persistent factor, a less persistent but still highly persistent factor, and the last mean-reverting factor. In the $\mathbb{Q}$ dynamics, there is not a large gap between *JSZ* and *BC estimates*. Under the $\mathbb{P}$ measure, however, *BC* two-step procedures yield much higher persistence than *JSZ* two-step

Table 4.2: Estimation – Summary Statistics for Ten Countries

			Max Eig($\Phi_{\mathcal{P}}$)		IRF		Half-life		Volatility			△(90-91/08-09)		
		RMSE(%)	$\mathbb{P}$	$\mathbb{Q}$	$\mathbb{P}$	$\mathbb{Q}$	$\mathbb{P}$	$\mathbb{Q}$	*forw*	*frn*	*ftp*	*forw*	*frn*	*ftp*
US	JSZ	0.052	0.9155	0.9710	0.16	0.60	9.00	27.00	1.25	0.03	1.24	-387	-4	-382
	BC	0.052	0.9862	0.9711	0.76	0.60	49.00	27.00	1.25	0.84	0.94	-387	-239	-147
Japan	JSZ	0.025	0.9262	0.9873	0.27	0.83	12.00	63.00	1.77	0.28	1.61	-473	-81	-393
	BC	0.025	0.9844	0.9873	0.72	0.83	35.00	63.00	1.77	1.47	1.64	-473	-398	-76
Germany	JSZ	0.043	0.9715	0.9737	0.54	0.66	23.00	32.00	1.48	0.93	0.88	-407	-279	-127
	BC	0.043	0.9997	0.9737	0.96	0.66	-	32.00	1.48	2.09	1.35	-407	-614	207
UK	JSZ	0.074	0.9528	0.9929	0.27	0.86	9.00	98.00	2.12	0.40	1.72	-568	-110	-458
	BC	0.074	0.9959	0.9929	0.67	0.86	97.00	98.00	2.12	1.43	0.69	-568	-386	-182
Canada	JSZ	0.043	0.9546	0.9973	0.34	0.82	12.00	200.00	1.98	0.41	1.59	-628	-137	-491
	BC	0.043	0.9883	0.9973	0.72	0.82	51.00	200.00	1.98	1.26	0.88	-628	-424	-204
Norway	JSZ	0.031	0.7668	0.9999	0.00	0.30	5.00	9.00	0.59	0.00	0.60	*n.a*	*n.a*	*n.a*
	BC	0.031	0.9157	0.9999	0.25	0.30	12.00	9.00	0.59	0.30	0.48	*n.a*	*n.a*	*n.a*
Sweden	JSZ	0.040	0.9504	0.9999	0.33	0.99	12.00	-	2.18	0.47	1.71	*n.a*	*n.a*	*n.a*
	BC	0.040	0.9982	0.9999	0.90	0.99	-	-	2.18	1.91	0.28	*n.a*	*n.a*	*n.a*
Switzerland	JSZ	0.055	0.9262	0.9892	0.30	0.55	13.00	30.00	1.14	0.40	0.96	-259	-97	-162
	BC	0.055	0.9917	0.9892	0.86	0.55	52.00	30.00	1.14	2.09	1.81	-259	-477	218
Australia	JSZ	0.040	0.9252	1.0000	0.17	1.00	7.00	-	2.27	0.15	2.12	-623	-42	-581
	BC	0.040	0.9804	1.0000	0.55	1.00	26.00	-	2.27	0.75	1.52	-623	-211	-412
NZ	JSZ	0.022	0.8911	0.9997	0.10	0.80	6.00	-	1.55	0.08	1.48	-476	-24	-452
	BC	0.022	0.9587	0.9998	0.40	0.80	16.00	-	1.55	0.51	1.08	-476	-161	-315

Note: (i) RMSE is the root mean squared of fitting errors computed by the square root of the average squared difference between the actual forward rates and the fitted rates from *JSZ* and *BC* two-step procedures. It is averaged across all quarters and all maturities. It is measured in annualized percentage points (4×100). (ii) IRF is the impulse-response function at horizon of five years of the first yield factor to a level shock. (iii) Half-life is the horizon (quarters) at which the IRF falls first below 0.5. If a computed half-life is larger than 50 years, we do not report it. (iv) The "Volatility" columns report the standard deviations of the fitted five- to ten-year forward rates denoted by *forw*, those of risk-neutral rates (three-month interest rate expectations under the $\mathbb{P}$ measure) denoted by *frn*, and those of corresponding term premia denoted by *ftp*. (v) The last three columns show changes in *forw*, *frn*, and *ftp* computed by the difference between the mean of observations from 1990:III to 1991:III (the early part of the sample) and from 2008:I to 2009:I (the late part of the sample.) We report in basis points. We do neither report Sweden whose observations starts from Dec. 1992. nor Norway whose observations starts from Jan. 1998.

procedures (for the remaining countries, see Table A.4.1 in Appendix).

Table 4.2 also reports summary statistics for all countries. Both *JSZ* and *BC estimates* produce the same cross-sectional fits and fitting errors are small. The persistence of yield factors is variously measured by the maximum eigenvalue of the coefficient matrix of the factor dynamics, the impulse response function and the half-life. The high persistence of factors shown in Table 4.2 corresponds well with empirical evidence that interest rates have a large permanent component (Cochrane and Piazzesi, 2009; Piazzesi, 2010). Under the $\mathbb{Q}$ measure, *JSZ* and *BC estimates* yield almost identical statistics. In fact, bias correction does

not affect the parameter estimates of the $\mathbb{Q}$ dynamics because the second-step ML estimation of the $\mathbb{Q}$ parameters is separated from the first-step estimation of the $\mathbb{P}$ parameters while bias correction is conducted only in the first-step estimation. Under the $\mathbb{P}$ measure, contrarily, all of the statistics for persistence estimated by *BC* two-step procedures are much higher than those estimated by *JSZ* two-step procedures. This means that, after bias correction, the persistence of the $\mathbb{P}$ dynamics sharply increases and moves toward the persistence of the $\mathbb{Q}$ dynamics across all countries so that *BC estimates* reflect the actual persistence of the $\mathbb{P}$ dynamics more reasonably.

Next, as described in (4.2.15), we decompose five- to ten-year forward rates into risk-neutral rates and term premia for ten countries. The "Volatility" columns in Table 4.2 report the volatilities of three components measured by the standard deviations. Comparing estimates from *JSZ* and *BC* two-step procedures, we can see that the volatility of the fitted forward rates does not change after bias correction since they are priced by the $\mathbb{Q}$ measure. On the other hand, the volatility of risk-neutral rates varies substantially across *JSZ* and *BC* two-step procedures since they are computed by the $\mathbb{P}$ parameter estimates. The increasing persistence of the $\mathbb{P}$ dynamics after bias correction renders risk-neutral rates more volatile. Moreover, the last three columns in Table 4.2 report the change from the early sample period to the late sample period of each component. Before correcting bias, the decline in risk-neutral rates can explain only a small portion of the decline in forward rates, and consequently term premia contribute to most of the secular trend in forward rates (excepting Germany). After bias correction, however, the majority of a secular decline in forward rates can be attributed to decreasing risk-neutral yields, rather than to term premia.

In Appendix, we depict the historical evolutions of risk-neutral rates and term premia. Figure A.4.1 shows the fitted five- to ten-year forward rates and the estimated risk-neutral rates. For all countries, the fitted forward rates exhibit a secular decline over the sample period. For the case of risk-neutral rates, *BC estimates* yield a distinct downward trend, whereas *JSZ estimates* produce a stable process. Figure A.4.2 illustrates corresponding

forward term premia as well. Due to changes in risk-neutral rates after bias correction, the movement of term premia also changes. Specifically, term premia from *BC estimates* no longer parallel the fitted forward rates but reveal a more counter-cyclical behavior: rising during recessions while falling during expansions.

To sum up, our results reproduce nearly all of the empirical findings in Wright (2011) and Bauer et al. (2012, 2014). We conclude this section by introducing one concern about *BC estimates*. As Bauer et al. (2012, 2014) pointed out, *BC* two-step procedures suffer from estimation uncertainty which is shown by the wide confidence intervals around *BC* risk-neutral rates in Figure A.4.1.

4.3 Recovery Theorem in the Gaussian Affine Term Structure

Now, we review the Recovery theorem for equity markets proposed by Ross (2015) and examine its applicability to fixed-income markets in the context of a GDTSM.

4.3.1 Recovery Theorem (Ross, 2015)

In Section 4.2, we delineated how standard GDTSM analyses identify the $\mathbb{P}$ and $\mathbb{Q}$ measures. Due to a separation property, the $\mathbb{P}$ measure is estimated by using time series data while the $\mathbb{Q}$ measure is estimated by using cross-sectional observations. They are only linked by the market prices of risk a posteriori.

On the other hand, Ross (2015) claimed that the $\mathbb{P}$ measure and the corresponding SDF can be recovered simultaneously from only the state prices. Note that Ross referred to his recovered probability as the *subjective probability* under the assumption of the existence of a representative agent and further equated it with the physical probability. A few papers argued that, however, the Recovery theorem does not necessarily recover the investors' expectations of future interest rates under the $\mathbb{P}$ measure. We will investigate what Ross really recovered in Section 4.4. For the moment, we set this issue aside. Instead, we refer to it as the *recovered probability measure* denoted by $\mathbb{L}$. Also, letting $\widehat{P}$ denote the *recovered*

transition probability matrix, we distinguish it from the *physical transition probability matrix* denoted by P and the *risk-neutral transition probability matrix* denoted by $\widetilde{P}$.

As in Ross (2015), we consider discrete-time and finite states that follow a time-homogeneous Markov process. Ross assumed the no-arbitrage restriction and a complete market as well. Let θ_i denote the current state and θ_j a state in the future. For one period, the state price is priced by

$$q(\theta_i, \theta_j) = e^{-r(\theta_i)}\, \widetilde{p}(\theta_i, \theta_j), \tag{4.3.1}$$

where $q(\theta_i, \theta_j)$ is the state price and $r(\theta_i)$ is the one-period interest rate in state θ_i. Also $\widetilde{p}(\theta_i, \theta_j)$ is each element of $\widetilde{P}$ which is the state transition probability from θ_i to θ_j under the $\mathbb{Q}$ measure. For multi-periods, the forward risk-neutral transition probability for going from state θ_i to θ_j in $T - t_1$ periods can be defined as

$$\widetilde{p}(\theta_i, \theta_j, T - t_1) = \sum_{\theta} \widetilde{p}(\theta_i, \theta, t_2 - t_1)\, \widetilde{p}(\theta, \theta_j, T - t_2), \tag{4.3.2}$$

where the summation is over all the possible intermediate states (θ) at time t_2 for $t_1 \leq t_2 \leq T$. This transition is time-homogeneous so that it does not depend on calendar time but the time interval. Then, we have the state price for the transition from θ_i at any time t to θ_j at T such that

$$q(\theta_i, \theta_j, t, T) = e^{-r(\theta_i)(T-t)}\, \widetilde{p}(\theta_i, \theta_j, T - t). \tag{4.3.3}$$

For simplicity, we let $\widetilde{p}(\theta_i, \theta_j) = \widetilde{p}_{ij}$, $r(\theta_i) = r_i$, and $\widetilde{q}(\theta_i, \theta_j) = \widetilde{q}_{ij}$, where i, j denote current and future states, respectively. To consider the change of measure from $\mathbb{Q}$ to $\mathbb{L}$, we define the Radon-Nikodym derivative of $\mathbb{L}$ with respect to $\mathbb{Q}$ as $\zeta_{ij} = \widehat{p}_{ij}/\widetilde{p}_{ij}$, where $\widehat{p}_{ij}$ is each element of $\widehat{P}$. We then find that

$$q_{ij} = e^{-r_i(T-t)}\, \widetilde{p}_{ij} = e^{-r_i(T-t)}\, \widehat{p}_{ij}/\zeta_{ij} = \widehat{s}_{ij}\, \widehat{p}_{ij}, \tag{4.3.4}$$

where $\widehat{s}_{ij}$ is the SDF associated with the $\mathbb{L}$ measure such that $\widehat{s}_{ij} = q_{ij}/\widehat{p}_{ij} = e^{-r_i(T-t)}/\zeta_{ij}$.

Ross (2015) imposed several restrictions to identify both $\widehat{s}_{ij}$ and $\widehat{p}_{ij}$ simultaneously from q_{ij} in (4.3.4). The transition-independent SDF is one of them. In particular, Ross considered an example of an inter-temporal model with an additively time-separable preference of a representative agent and derived the SDF as its equilibrium solution. The resulting SDF is the product of a constant time-discount rate (ς) and the marginal rate of substitution between future and current consumption:

$$\widehat{s}(\theta_i, \theta_j) = \varsigma \, \frac{h(\theta_j)}{h(\theta_i)}, \tag{4.3.5}$$

where $h(\theta_i)$ is the marginal utility of consumption in state θ_i (or equivalently, a pricing kernel). Thus, the above SDF does not depend on the intermediate path between initial and final states. Obviously, the state price is expressed as

$$q_{ij} = \widehat{s}_{ij} \, \widehat{p}_{ij} = \varsigma \, \frac{h_j}{h_i} \, \widehat{p}_{ij}, \tag{4.3.6}$$

where $\widehat{s}_{ij} = \widehat{s}(\theta_i, \theta_j)$ and $h_i = h(\theta_i)$. In matrix notation, (4.3.6) can be written as

$$DQ = \varsigma \widehat{P} D \quad \text{and} \quad \widehat{P} = \varsigma^{-1} D Q D^{-1}, \tag{4.3.7}$$

where Q is the state-price matrix and D is a diagonal matrix whose each diagonal element is h_i. Note that the sum of each row in Q is the current value in each current state of a dollar for sure in the future; that is, $\sum_j q_{ij} = e^{-r_i}$.

Generally, Q alone is not enough to identify the recovered transition probability ($\widehat{P}$) and the SDF separately. Let m^* denote the total number of states. In (4.3.7), we have only m^{*2} equations with $(m^{*2} + m^* + 1)$ unknowns: m^{*2} probabilities, m^* pricing kernels, and a constant discount rate. Under the assumption of the irreducible transition matrix, however, Ross (2015) could solve the above system of equations by using the Perron-Frobenius theorem (hereafter *PF theorem*). Let us consider a characteristic function for a given square matrix

A such that $AV = \Gamma V$, where Γ is a diagonal matrix whose non-zero elements are the eigenvalues of A, and V is a matrix composed of corresponding eigenvectors. The PF theorem says that if A is non-negative and irreducible, there exists a unique positive real eigenvalue which is referred to as the *perron root*, and all other eigenvalues are smaller in absolute value. Moreover, a corresponding unique positive eigenvector is called the *perron vector*, and there are no strictly positive eigenvectors except for positive multiples of the perron vector (Meyer, 2000, p. 673). In layman's terms, A is irreducible if there is always at least one path such that any state j can be attainable from any state i in finite steps. For a formal definition, see Jiang (2010, p. 325).

Since $\widehat{P}$ is a stochastic matrix, $\widehat{P}e = e$ where e is a vector of ones. Consequently, we have

$$\widehat{P}e = e = \varsigma^{-1} DQD^{-1}e \quad \text{and} \quad QD^{-1}e = \varsigma D^{-1}e. \tag{4.3.8}$$

Equivalently,

$$Qv = \varsigma v, \tag{4.3.9}$$

where $v = D^{-1}e$. If Q is irreducible, the discount rate ς is the same as the perron root of Q, and v is the perron vector whose elements $v_i = 1/h_i$. Thus, if state prices are known, we can recover a certain probability density ($\widehat{p}_{ij}$) and a corresponding SDF ($\widehat{s}_{ij}$) from the following equations:

$$\widehat{s}_{ij} = \varsigma(v_i/v_j) \quad \text{and} \quad \widehat{p}_{ij} = q_{ij}/\widehat{s}_{ij}. \tag{4.3.10}$$

It is worth highlighting that the state-price matrix should be fully specified over all parallel universes to solve (4.3.10). Obviously, most states are neither realized nor observable however. To address this issue in equity markets, Ross (2015, Section V) described how to compute state prices for unrealized states by using option prices. Regretfully, this method is not always feasible and it is very complicated for the case with a multi-dimensional state space (Ross and Martin, 2013, p. 14). Alternatively, Ross and Martin (2013) sidestepped this issue by connecting the perron root and the perron vector to the yield and the return

on the long bond with an infinite maturity, respectively; however, there still remains the question of how well a long but finite bond can approximate the infinitely long bond.

4.3.2 Application of the Recovery Theorem in GDTSMs

Our GDTSM is in line with the framework of Ross (2015). Under no arbitrage, the state dynamics is described as a time-homogeneous stationary Markov chain. Also, a complete market assumption is acceptable since the fixed-income derivatives market is one of the most developed derivatives markets (Ross and Martin, 2013).

To apply the Recovery theorem to GDTSMs, we start from the $\mathbb{Q}$ state dynamics rather than consider utility maximization as in Ross (2015). Due to the specific structure of the $\mathbb{Q}$ dynamics, we can obtain the risk-neutral state transition probability matrix ($\widetilde{P}$) from the true data-generating process under the $\mathbb{Q}$ measure by using Markov-chain approximations. Then, the state-price matrix (Q) can be constructed from a risk-neutral pricing equation (4.3.1). Our method for specifying Q is different from those proposed by Ross and Martin (2013) and Ross (2015) for equity market applications. Lastly, if Q is non-negative and irreducible, we can recover the transition probability matrix ($\widehat{P}$) and the corresponding SDF by using the PF theorem.

4.3.2.1 Step 1: Construction of the risk-neutral probability transition matrix

In this subsection, we introduce a finite-state Markov approximation method to obtain the risk-neutral transition probability matrix ($\widetilde{P}$) from the estimated $\mathbb{Q}$ state dynamics. Since our results are significantly affected by the accuracy of approximation, we choose an appropriate method for our model specification carefully.

A finite-state Markov-chain approximation method Tauchen (1986a) proposed a finite-state Markov-chain approximation to univariate (AR) and vector autoregressions (VAR) with a diagonal error covariance matrix such that a generated discrete state-space Markov process can closely replicate the underlying stationary dynamics of the state vari-

ables. The method needs to select discrete values which each state variable can take (also called grid points) and constructs time-homogeneous state transition probabilities based on the distribution of the underlying process. The accuracy of this method is very sensitive to the number of grid points (m^*). Tauchen argued that the method can yield better approximations as m^* becomes larger so that the state space becomes finer. Note that Aydin and Yildirim (2015) employed the method of Terry and Knotek (2011) which extends Tauchen's method to a VAR with a non-diagonal error covariance matrix.

Follow-up studies show that Tauchen's method and its extension do not perform well when a VAR process is highly persistent; in particular, the accuracy remains poor even though the number of grid points increases sharply (Floden, 2008; Kopecky and Suen, 2010; Farmer and Toda, 2015). This might be ascribed to the fact that Tauchen targeted only the first conditional moment of the underlying process (Gospodinov and Lkhagvasuren, 2014, p. 846). Considering highly persistent factors in GDTSMs, Tauchen's method seems inappropriate for our study. Besides, it is infeasible in practice. When we set a state space much finer to improve the accuracy of approximation in the presence of highly persistent factors, the process is very time consuming and computer memory may be insufficient to deal with a large-dimensional transition matrix.

As a response, Rouwenhorst (1995) developed an alternative method that approximates both the conditional mean and variance of the underlying AR process. Gospodinov and Lkhagvasuren (2014) extended it to a VAR process with a diagonal error covariance matrix. In a highly persistent VAR process, Rouwenhorst's method and its extension (hereafter *GL method*) outperform Tauchen's method even without increasing the number of grid points. For example, when the largest eigenvalue of the coefficient matrix of the $\mathbb{Q}$ state dynamics is close to unity, the Tauchen's method needs at least 25 grid points for each dimension in order to be comparable to the GL method with 5 grid points in terms of approximation quality (Kopecky and Suen, 2010; Galindev and Lkhagvasuren, 2010; Farmer and Toda,

2015). The GL method also reduces the computing time substantially.[2] More importantly, the GL method produces irreducible state-price matrices for all countries in our empirical study. In contrast, when we use the method of Terry and Knotek (2011), we fail to obtain irreducible matrices for most countries except for the UK and the US. For these reasons, we employ the GL method in our empirical study.

Application of the GL method Recall our trivariate VAR(1) process of yield factors under the $\mathbb{Q}$ measure. For simplicity, we suppress the $\mathcal{P}$ subscripts here:

$$\mathcal{P}_{t+1} = \mu^{\mathbb{Q}} + \Phi^{\mathbb{Q}}\mathcal{P}_t + \Sigma\epsilon^{\mathbb{Q}}_{t+1}, \tag{4.3.11}$$

where $\mathcal{P}_t = (\mathcal{P}_{1,t}, \mathcal{P}_{2,t}, \mathcal{P}_{3,t})'$, $\mu^{\mathbb{Q}}$ is a 3×1 vector, $\Phi^{\mathbb{Q}}$ is a 3×3 matrix, a 3×1 vector $\epsilon_t \sim \mathcal{N}(0, I_3)$ and Σ is a 3×3 lower triangular matrix such that $\Sigma\Sigma' = V$. Under stationarity, the largest eigenvalue of $\Phi^{\mathbb{Q}}$ is less than 1. Since V is not necessarily diagonal, we transform (4.3.11) to a VAR with a diagonal error covariance matrix by a linear transformation described in Tauchen (1986b). In detail, letting $\mathcal{Y}_t = C^{-1}(\mathcal{P}_t - (I - \Phi^{\mathbb{Q}})^{-1}\mu^{\mathbb{Q}})$, $A = C^{-1}\Phi^{\mathbb{Q}}C$, and $\eta_t = C^{-1}\Sigma\epsilon^{\mathbb{Q}}_t$, we have

$$\mathcal{Y}_{t+1} = A\mathcal{Y}_t + \eta_{t+1}, \tag{4.3.12}$$

where $\eta_t \sim$ i.i.d. $\mathcal{N}(0, \Omega)$, C is a 3×3 lower triangular matrix, and Ω is a 3×3 diagonal matrix such that $V = C\Omega C'$.

Let $\widetilde{\mathcal{Y}}_t$ denote the approximate discrete-valued vector of $\mathcal{Y}_t$. Now, we construct grid points for each element of $\widetilde{\mathcal{Y}}_t$. We denote each element by $\widetilde{\mathcal{Y}}_{k,t}$ and its grid points by $\bar{\mathcal{Y}}^g_k$ for $k = 1, 2, 3$ and $g = 1, 2, \ldots, m_k$. That is, for any k, $\widetilde{\mathcal{Y}}_{k,t}$ takes one of m_k discrete values which are sorted in a decreasing order $\bar{\mathcal{Y}}^1_k < \bar{\mathcal{Y}}^2_k < \ldots < \bar{\mathcal{Y}}^{m_k}_k$. For simplicity, we assume that

[2] In our empirical study of the three-factor GDTSM using the US data, the computing time of the GL methods with 21 grid points for each dimension (that is, the total number of grid points is $21^3 = 9261$) is 138 minutes. However, the method of Terry and Knotek takes $4,220$ minutes (We use Matlab on a 1.7 GHz Intel Core i5 with 4GB DDR3).

each yield factor has the same number of grid points; that is, $m = m_k$ for all k. These m grid points are given by equally spaced points. Specifically,

$$\bar{\mathcal{Y}}_k^g = -\sigma_{y_k}(m-1)^{1/2} + 2\sigma_{y_k}(g-1)/(m-1)^{1/2}$$

for $g = 1, 2, \ldots, m$, where $\sigma_{y_k} = var(\mathcal{Y}_{k,t})$. At time t, the entire system will be in one of $m^3 = m^*$ states; that is, $\widetilde{\mathcal{Y}}_t$ takes one of m^* vectors denoted by $\bar{\mathcal{Y}}^{\,i}$ for $i = 1, 2, \ldots, m^*$. Next, we consider the time-homogeneous *individual* transition probability defined as

$$\tilde{p}_k(i,g) = \Pr(\widetilde{\mathcal{Y}}_{k,t} = \bar{\mathcal{Y}}_k^{\,g} \;\mid\; \widetilde{\mathcal{Y}}_{t-1} = \bar{\mathcal{Y}}^{\,i})$$

such that $\sum_{g=1}^{m} \tilde{p}_k(i,g) = 1$. To generate a Markov chain process which can replicate an underlying process closely, the GL method targets the first and second conditional moments of $\mathcal{Y}_t$ by minimizing the distance of the following moment conditions:

$$(i)\ \sum_{g=1}^{m} \tilde{p}_k(i,g)\bar{\mathcal{Y}}_k^{\,g} - \varphi_k(i) \quad \text{and} \quad (ii)\ \sum_{g=1}^{m} \tilde{p}_k(i,g)\left(\bar{\mathcal{Y}}_k^{\,g} - \varphi_k(i)\right)^2 - \vartheta_k^2,$$

where ϑ_k^2 is the k-th diagonal element of Ω, and $\varphi_k(i)$ denotes the expected value of process $\mathcal{Y}_{k,t+1}$, conditional on $\mathcal{Y}_t = \bar{\mathcal{Y}}^{\,i}$. Letting l_k be an integer-valued function for any k such that $\widetilde{\mathcal{Y}}_{k,t} = \bar{\mathcal{Y}}_k^{l_k(i)}$ when the system is in state i at time t, it holds that $\varphi_k(i) = \sum_{h=1}^{3} a_{k,h}\bar{\mathcal{Y}}_h^{\,l_h(i)}$, where $a_{k,h}$ is each element of A in (4.3.12).

Next, we obtain the m^*-dimensional risk-neutral transition probability matrix ($\widetilde{P}$) whose each element is the probability that $\widetilde{\mathcal{Y}}_t$ will be in a future state j conditional on a current state i. Since η_t are independent, each element of $\widetilde{P}$ is the product of individual transition probabilities: $\widetilde{p}_{ij} = \prod_{k=1}^{3} \tilde{p}_k(i, l_k(j))$ for $i, j = 1, 2, \ldots, m^*$. So far, we construct the discrete values of a transformed process ($\mathcal{Y}_t$) and the transition probability matrix ($\widetilde{P}$). Lastly, we can back up the grid points of $\mathcal{P}_t$ by a reverse transformation.

Remark 1. Quality of Markov-chain approximation As in the literature, we obtain the VAR parameters via simulations based on the transition probabilities and also obtain the parameters from direct simulations of the underlying VAR. Then, we compare the signs and magnitudes of their means. Much of the literature usually focuses on the difference in two decimal points (Tauchen, 1986a; Terry and Knotek, 2011; Gospodinov and Lkhagvasuren, 2014). For details, see Section 4.5.

4.3.2.2 Step 2: Construction of the state-price matrix

Let $(z^1, z^2, \ldots, z^{m^*})$ denote a set of m^* discrete-valued 3×1 vectors for $\mathcal{P}_t$. For one period, we can write a short rate equation and a risk-neutral pricing equation as follows. For $i, j = 1, \ldots, m^*$,

$$r_i = \rho_0 + \rho_1' z^i; \tag{4.3.13}$$

$$q_{ij} = e^{-r_i} \widetilde{p}_{ij}, \tag{4.3.14}$$

where ρ_0 is a constant, ρ_1 is a 3×1 vector, and r_i is the one-period interest rate in state i. Moreover, $\widetilde{p}_{ij}$ is the risk-neutral transition probability obtained by the GL method and q_{ij} is the one-period state price. Then, from (4.3.14), we can compute the state-price matrix.

For example, if $m = 9$ so that $m^* = 729$, we need to construct a 729×729 matrix Q:

$$\begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,729} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,729} \\ \vdots & \vdots & \ddots & \vdots \\ q_{728,1} & q_{728,2} & \cdots & q_{728,729} \\ q_{729,1} & q_{729,2} & \cdots & q_{729,729} \end{bmatrix} = \begin{bmatrix} e^{-r_1} \cdot p_{1,1} & e^{-r_1} \cdot p_{1,2} & \cdots & e^{-r_1} \cdot p_{1,729} \\ e^{-r_2} \cdot p_{2,1} & e^{-r_2} \cdot p_{2,2} & \cdots & e^{-r_2} \cdot p_{2,729} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-r_{728}} \cdot p_{728,1} & e^{-r_{728}} \cdot p_{728,2} & \cdots & e^{-r_{728}} \cdot p_{728,729} \\ e^{-r_{729}} \cdot p_{729,1} & e^{-r_{729}} \cdot p_{729,2} & \cdots & e^{-r_{729}} \cdot p_{729,729} \end{bmatrix}, \tag{4.3.15}$$

where

$$r_i = \rho_0 + \begin{bmatrix} \rho_{1,1} & \rho_{1,2} & \rho_{1,3} \end{bmatrix} \begin{bmatrix} z_1^i \\ z_2^i \\ z_3^i \end{bmatrix}. \tag{4.3.16}$$

Remark 2. Irreducibility of the state-price matrix To apply the PF theorem, Q should be irreducible. To check the irreducibility, we use the result from Berman and Plemmons (1979, Theorem 2.1.3) and Meyer (2000, Lemma 8.3.5): If an $r \times r$ non-negative matrix A is irreducible, then $(I_r + A)^{r-1}$ should be strictly positive.

4.3.2.3 Step 3: Application of the Perron-Frobenius theorem

As described in Section 4.3.1, we identify the recovered probability transition matrix ($\widehat{P}$) by the PF theorem. Suppose that Q is irreducible. Then, from (4.3.6) we have

$$\widehat{p}_{ij} = \varsigma^{-1}(v_j/v_i) \cdot q_{ij}, \tag{4.3.17}$$

where v_i is the i-th element of the perron vector of Q and ς is the corresponding perron root.

Golub and Loan (2013, p. 373) showed that the power method is useful to find the perron vector of a non-negative and irreducible matrix. By iteratively computing the powers of a matrix, the power method approximates a dominant eigenpair (ς, v), where ς is the dominant eigenvalue that is larger in absolute value than all of the other eigenvalues and v is the dominant positive eigenvector associated with ς. For details, see Meyer (2000, p. 533) and Golub and Loan (2013, p. 366).

4.3.2.4 Step 4: State dynamics under the recovered probability measure

To analyze a GDTSM with respect to the $\mathbb{L}$ measure, we need to estimate the factor dynamics under the $\mathbb{L}$ measure. First, we posit the following trivariate VAR(1) process under the $\mathbb{L}$ dynamics:

$$\mathcal{P}_{t+1} = \mu^{\mathbb{L}} + \Phi^{\mathbb{L}}\mathcal{P}_t + \Sigma\epsilon^{\mathbb{L}}_{t+1}, \tag{4.3.18}$$

where $\mu^{\mathbb{L}}$ is a 3×1 vector, $\Phi^{\mathbb{L}}$ is a 3×3 matrix, Σ is a 3×3 lower triangular matrix such that $\Sigma\Sigma' = V$, and $\epsilon^{\mathbb{L}}_t \sim \mathcal{N}(0, I_3)$. Also, we consider the following process which can be

generated by the GL method:

$$\widetilde{\mathcal{P}}_{t+1} = \ddot{\mu}^{\mathbb{L}} + \ddot{\Phi}^{\mathbb{L}}\widetilde{\mathcal{P}}_t + \ddot{\Sigma}\epsilon^{\mathbb{L}}_{t+1} = B^{\mathbb{L}}W_t + \ddot{\Sigma}\epsilon^{\mathbb{L}}_{t+1}, \tag{4.3.19}$$

where $\ddot{\mu}^{\mathbb{L}}$ is a 3×1 vector, $\ddot{\Phi}^{\mathbb{L}}$ is a 3×3 matrix, and $\ddot{\Sigma}$ is a 3×3 lower triangular matrix such that $\ddot{\Sigma}\ddot{\Sigma}' = \ddot{V}$.

(4.3.19) can be estimated as in Tauchen (1986a). Letting a 3×4 matrix $B^{\mathbb{L}} = (\ddot{\mu}^{\mathbb{L}}, \ddot{\Phi}^{\mathbb{L}})$ and a 4×1 vector $W_t = (1, \widetilde{\mathcal{P}}_t')'$, we have $B^{\mathbb{L}} = [E^{\mathbb{L}}(\widetilde{\mathcal{P}}_{t+1}W_t')][E^{\mathbb{L}}(W_tW_t')]^{-1}$, where the expectation can be computed by the recovered transition probability. For details, see Appendix A.4.3.

4.4 Recovery Theorem Revisited

In this section, we summarize the claim of "misspecified recovery" in BHS (2015) and examine this misspecification issue regarding our affine Gaussian dynamic term structure. BHS argued that the $\mathbb{L}$ measure is not necessarily same as the $\mathbb{P}$ measure. Further, they defined the $\mathbb{L}$ measure as the *long-term risk-neutral probability measure* because it absorbs only the martingale component of the SDF (or equivalently, investors' risk aversion to permanent shock).

4.4.1 Misspecified Recovery: Recovery of Long-term Risk-neutral Measure

To figure out what the Recovery theorem really recovers, we review the results from AJ (2005) regarding the SDF decomposition under a discrete-time and finite-state stationary Markov process. The literature has carried out similar analyses in a continuous-time framework (e.g., Hansen and Scheinkman, 2009; Christensen, 2014; BHS, 2015; Qin and Linetsky, 2016; Qin and Linetsky, 2017).

Let S_t denote a pricing kernel. AJ (2005) proposed the following decomposition:

$$S_t = S_t^T S_t^P \quad \text{with} \quad E_t(S_{t+1}^P) = S_t^P, \tag{4.4.1}$$

where S_t^T is the transitory component of a pricing kernel and S_t^P is the permanent component which is a martingale. Correspondingly, the one-period SDF ($s_{t,t+1} = S_{t+1}/S_t$) is factorized as

$$s_{t,t+1} = s_{t,t+1}^T \cdot s_{t,t+1}^P \quad \text{with} \quad E_t(s_{t,t+1}^P) = 1, \tag{4.4.2}$$

where $s_{t,t+1}^T = S_{t+1}^T/S_t^T$ and $s_{t,t+1}^P = S_{t+1}^P/S_t^P$ are the transitory and permanent components of the SDF, respectively. According to AJ (2005, Proposition 3), the transitory component of the SDF is the same as the inverse of the long-bond return ($s_{t,t+1}^T = 1/R_{t,t+1}^\infty$).[3] Thus, the long-term bond can be priced by $s_{t,t+1}^T$ such that $E(s_{t,t+1}^T R_{t,t+1}^\infty) = 1$ (Bakshi and Chabi-Yo, 2012, p. 193).

Recall the PF theorem that yields $Qv = \varsigma v$ (4.3.9) and a pricing equation (4.2.7). Then, we get

$$E_t(s_{t,t+1} v_{t+1}) = \varsigma v_t \quad \text{so that} \quad E_t(s_{t,t+1} v_{t+1}/\varsigma v_t) = 1. \tag{4.4.3}$$

For details, see Hansen and Scheinkman (2009, Proposition 6.2) and BHS (2015, Problem 4.1). Considering that the permanent component is a martingale, each component of the SDF can be defined as follows:

$$s_{t,t+1}^P = \varsigma^{-1} s_{t,t+1} \frac{v_{t+1}}{v_t}; \tag{4.4.4}$$

$$s_{t,t+1}^T = \varsigma \frac{v_t}{v_{t+1}}. \tag{4.4.5}$$

Denote a current state by i and a future state by j. For a single period, the state price

[3] According to AJ (2005), the long-bond return ($R_{t,t+1}^\infty$) is the gross return from holding a bond maturing at an infinite horizon from time t to $t+1$. That is, $R_{t,t+1}^\infty \equiv \lim_{\tau\to\infty} R_{t,t+1}^\tau = \lim_{\tau\to\infty} \frac{V_{t+1}(1_{t+\tau})}{V_t(1_{t+\tau})}$, where $V_t(1_{t+\tau})$ is the current price of a bond maturing at time $t+\tau$.

is priced under the $\mathbb{P}$ measure such that

$$q_{ij} = s_{ij}\, p_{ij} = s_{ij}^{T}\, s_{ij}^{P}\, p_{ij}, \tag{4.4.6}$$

where p_{ij} is the physical probability and s_{ij} is the associated SDF. s_{ij} is referred to as the *original SDF* hereafter. Also, as shown in (4.3.1), the state price is priced under the $\mathbb{Q}$ measure such that

$$q_{ij} = \widetilde{s}_{ij}\, \widetilde{p}_{ij} = e^{-r_i}\, \widetilde{p}_{ij}, \tag{4.4.7}$$

where $\widetilde{p}_{ij}$ is the risk-neutral probability. Moreover, we can write $\widetilde{p}_{ij} = q_{ij}/\bar{q}_i$ such that $\sum_j \widetilde{p}_{ij} = 1$, where $\bar{q}_i = \sum_j q_{ij} = e^{-r_i}$. Here the resulting one-period SDF ($\widetilde{s}_{ij} = e^{-r_i}$) is independent of any tomorrow state j, which implies that all possible tomorrow states j are discounted equally. Consequently, risk adjustment (excepting a time discount factor) is absent from the SDF under the $\mathbb{Q}$ measure; rather, it is absorbed in the corresponding risk-neutral probability.

(4.4.6) and (4.4.7) imply that the SDF should be defined subject to the given probability measure. By the SDF decomposition, we can define another probability measure and the corresponding SDF:

$$\begin{aligned} q_{ij} &= s_{ij}^{T}\, \underbrace{s_{ij}^{P}\, p_{ij}} \\ &= s_{ij}^{T} \cdot \quad \widehat{p}_{ij}, \end{aligned} \tag{4.4.8}$$

where $\widehat{p}_{ij} = s_{ij}^{P}\, p_{ij}$. Also, from (4.4.5) we can see that

$$s_{ij}^{T} = \varsigma\, \frac{v_i}{v_j} = \widehat{s}_{ij}, \tag{4.4.9}$$

where $\widehat{s}_{ij}$ is the *recovered SDF* by the Recovery theorem defined in (4.3.10). Thus, (4.4.9) implies that the recovered SDF is nothing but the transitory component of the original

SDF. For example, $\widehat{s}_{ij}$ can be trend-stationary (BHS, 2015, p. 2). On the other hand, the permanent component (s_{ij}^P) is absorbed in the recovered probability ($\widehat{p}_{ij}$).

Let us examine the above decomposition in detail. The Recovery theorem actually identifies $\widehat{s}_{ij}$ and $\widehat{p}_{ij}$, not s_{ij} and p_{ij}. In fact, since $\widehat{p}_{ij} = s_{ij}^P \, p_{ij}$, the $\mathbb{L}$ measure absorbs risk compensation for exposure to only permanent shocks (or equivalently, the martingale component of s_{ij}). In this sense, BHS (2015) referred to the $\mathbb{L}$ measure as the *long-term risk-neutral probability measure* so that it can be distinguished from the $\mathbb{Q}$ measure which absorbs overall risk aversion except for a time discount factor. Consequently, the difference between the $\mathbb{P}$ and $\mathbb{Q}$ measures reflects all of risk adjustments, while the difference between the $\mathbb{P}$ and $\mathbb{L}$ mirrors risk compensation for exposure to only the long-term components of risk.

Although the $\mathbb{Q}$ and $\mathbb{L}$ measures are distinguishable by definitions, we can find their similarity as well. Both probability measures are adjusted by investors' risk aversion in different degrees: The $\mathbb{Q}$ measure absorbs compensation regarding overall risk, while the $\mathbb{L}$ measure absorbs compensation regarding long-term (martingale) risk. Since it is known that the behavior of the original SDF is dominated by its martingale component (AJ, 2005; Bakshi and Chabi-Yo, 2012), these two measures are not much different from each other. BHS (2015, p. 28) provided empirical examples of the similarity between $\mathbb{Q}$ and $\mathbb{L}$. Both are clearly distinct from the $\mathbb{P}$ measure, however. Qin et al. (2016, Section 5) showed that the $\mathbb{Q}$ and $\mathbb{L}$ measures produce almost identical forecasts, while the forecast under the $\mathbb{P}$ measure is clearly distinguished from them.

Particularly, if interest rates are constant, $\mathbb{Q}$ is identical with $\mathbb{L}$. In this case, the row sums of Q are identical so that riskfree rates are state-independent and $Qe = \exp(-\bar{r})e$. By the PF theorem, e and $\exp(-\bar{r})$ are the perron vector and the perron root of Q, respectively. Also, $Q = \exp(-\bar{r})\widehat{P}$ since $\widehat{P}e = e$, and consequently it follows that $\mathbb{L} = \mathbb{Q}$. Ross (2015, Theorem 2) interpreted this result as $\mathbb{P} = \mathbb{Q}$ since he equated $\mathbb{P}$ with $\mathbb{L}$. In our GDTSM, however, interest rates are not deterministic since they are affine in yield factors.

The question still remains: Under what circumstances can the Recovery theorem recover

the physical probability measure? Obviously, from SDF decompositions (4.4.6) and (4.4.8), we can see that $\widehat{p}_{ij} = p_{ij}$ holds if $s^P_{ij} = 1$. Since s^P_{ij} is a martingale component, it can be considered as the Radon-Nikodym derivative for the change of measure from $\mathbb{P}$ to $\mathbb{L}$; that is, $s^P_{ij} = \widehat{p}_{ij}/p_{ij}$. To sum up, we can say that $\mathbb{P} = \mathbb{L}$ only if this Radon-Nikodym derivative is unity, or equivalently only if the permanent component is degenerate. As shown in (4.4.6), a degenerating martingale component implies that the original SDF is transition-independent since $s_{ij} = s^T_{ij} = \varsigma(v_i/v_j)$. It follows that $s_{ij} = \widehat{s}_{ij}$. Consequently, $\widehat{p}_{ij} = p_{ij}$.

The literature has examined the reliability of a degenerating martingale component. First, AJ (2005) argued that $s^P_{ij} = 1$ is not the case. They theoretically showed that when s^P_{ij} is unity, a return on the long bond maturing at an infinite horizon should be higher than any other assets; however, their empirical studies provided counter-evidence that bond returns with a sufficiently long maturity are much lower than those of equity indexes. For more details, see Qin and Linetsky (2017, p. 303) and BHS (2015, Section 4.3). Second, Bakshi and Chabi-Yo (2012), along with AJ (2005), questioned $s^P_{ij} = 1$ by showing that the lower bound of the permanent component of the original SDF is substantially more volatile than that of the transitory component. In addition, AJ (2005, p. 2004) and BHS (2015, Example 2.2) also presented recursive preferences as an example of the SDF which has a non-trivial martingale component. To sum up, sufficient theoretical and empirical evidence implies that the degeneracy of a martingale component is an implausible restriction and hence $\mathbb{P} \neq \mathbb{L}$ generally.

Suppose that we equate $\mathbb{L}$ with $\mathbb{P}$ even though a martingale component is not negligible. Then, the misspecified $\mathbb{P}$ measure misleads us about risk premia and investors' short-rate expectations. For example, as BHS (2015) and Qin and Linetsky (2016) pointed out, the $\mathbb{L}$ measure makes the long-term risk-return tradeoffs degenerate because assets are priced under the $\mathbb{L}$ measure as if long-term risk premia were zero even in the presence of long-term shocks (e.g., stochastically growing cash flows). Such degeneracy is not likely to hold under the *true* $\mathbb{P}$ measure.

4.4.2 GDTSM under the Long-term Risk-neutral Probability Measure

Now, we analyze our GDTSM under the long-term risk-neutral probability measure ($\mathbb{L}$) by using the SDF decomposition and the change of measure.

4.4.2.1 Long-term risk-neutral dynamics

As described in Section 4.2.1, we start with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration $\mathcal{F}_t$, defined for $0 \leq t \leq T$, where T is a fixed final time. $X = \{X_t : t \in T\}$ is an N-dimensional stationary Markov process. Recall our state dynamics under the $\mathbb{P}$ measure (4.2.1):

$$X_{t+1} = \mu + \Phi X_t + \Sigma \epsilon_{t+1},$$

where $\epsilon_t \sim \mathcal{N}(0, I_N)$. ϵ_t represents a source of risk from unknown shocks at time t. Also, our $\mathbb{Q}$ state dynamics (4.2.8) is

$$X_{t+1} = \mu^{\mathbb{Q}} + \Phi^{\mathbb{Q}} X_t + \Sigma \epsilon_{t+1}^{\mathbb{Q}},$$

where $\mu^{\mathbb{Q}} = \mu - \lambda_0$, $\Phi^{\mathbb{Q}} = \Phi - \lambda_1$, and $\epsilon_t^{\mathbb{Q}} \sim \mathcal{N}(0, I_N)$.

We can consider the change of measure from $\mathbb{P}$ to $\mathbb{L}$ by using the martingale component of the original SDF as the Radon-Nikodym derivative of $\mathbb{L}$ with respect to $\mathbb{P}$; that is, $S^P = d\mathbb{L}/d\mathbb{P}$. As in Section 4.3.1, we also consider the Radon-Nikodym derivative of $\mathbb{L}$ with respect to $\mathbb{Q}$ for the change of measure from $\mathbb{Q}$ to $\mathbb{L}$; that is, $\zeta = d\mathbb{L}/d\mathbb{Q}$. Then, we have that $E_t^{\mathbb{L}}(Z_{t+1}) = E_t(S_{t+1}^P Z_{t+1})/S_t^P = E_t^{\mathbb{Q}}(\zeta_{t+1} Z_{t+1})/\zeta_t$ for any random variable Z_{t+1}, where $E_t^{\mathbb{L}}$ is the conditional expectation under the $\mathbb{L}$ measure.

As seen before, we can represent the price of a τ-period zero-coupon bond at time t as follows:

$$P_t^{(\tau)} = E_t\left(\frac{S_{t+1}}{S_t}\, P_{t+1}^{(\tau-1)}\right) = E_t\left(\frac{S_{t+1}^P}{S_t^P}\frac{S_{t+1}^T}{S_t^T}\, P_{t+1}^{(\tau-1)}\right) = E_t^{\mathbb{L}}\left(\frac{S_{t+1}^T}{S_t^T}\, P_{t+1}^{(\tau-1)}\right). \tag{4.4.10}$$

Also, recall the Radon-Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$ from Section 4.2; that is, $\xi = d\mathbb{Q}/d\mathbb{P}$. Then, in line with (4.2.7) and (4.3.4), we find the relation among three different probability measures:

$$P_t^{(\tau)} = E_t\left(e^{-r_t}\,\frac{\xi_{t+1}}{\xi_t}\,P_{t+1}^{(\tau-1)}\right) = E_t^{\mathbb{Q}}\left(e^{-r_t}\,P_{t+1}^{(\tau-1)}\right) = E_t^{\mathbb{L}}\left(e^{-r_t}\frac{\zeta_t}{\zeta_{t+1}}\,P_{t+1}^{(\tau-1)}\right). \tag{4.4.11}$$

Consequently, the SDF associated with the $\mathbb{L}$ measure is defined as

$$\widehat{s}_{t,t+1} = S_{t+1}^T/S_t^T = \exp(-r_t)\,\zeta_t/\zeta_{t+1}. \tag{4.4.12}$$

Recall the log-normal process for ξ, (4.2.6): $\log\xi_{t+1} = \log\xi_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t'\epsilon_{t+1}$, where λ_t represents the market prices of risk, given by $\lambda_t = \Sigma^{-1}(\lambda_0 + \lambda_1 X_t)$ as in (4.2.4). Similarly, suppose that ζ_t, which is a martingale under the $\mathbb{Q}$ measure, follows the log-normal process:

$$\log\zeta_{t+1} = \log\zeta_t - \frac{1}{2}\lambda_t^{\mathbb{L}\prime}\lambda_t^{\mathbb{L}} - \lambda_t^{\mathbb{L}\prime}\epsilon_{t+1}^{\mathbb{Q}}, \tag{4.4.13}$$

where $\lambda_t^{\mathbb{L}} = \Sigma^{-1}(\lambda_0^{\mathbb{L}} + \lambda_1^{\mathbb{L}} X_t)$. We then recover the $\mathbb{L}$ dynamics of the state variables by using Girsanov's theorem (for details of the proof, see Appendix A.4.4). The $\mathbb{L}$ state dynamics also follows a Gaussian VAR(1) process:

$$X_{t+1} = \mu^{\mathbb{L}} + \Phi^{\mathbb{L}} X_t + \Sigma\epsilon_{t+1}^{\mathbb{L}}, \tag{4.4.14}$$

where $\epsilon_t^{\mathbb{L}} \sim \mathcal{N}(0, I_N)$,

$$\mu^{\mathbb{L}} = \mu^{\mathbb{Q}} - \lambda_0^{\mathbb{L}}, \quad \text{and} \quad \Phi^{\mathbb{L}} = \Phi^{\mathbb{Q}} - \lambda_1^{\mathbb{L}}. \tag{4.4.15}$$

Also, as seen before, $\mu^{\mathbb{Q}} = \mu - \lambda_0$ and $\Phi^{\mathbb{Q}} = \Phi - \lambda_1$. By Girsanov's theorem, we have

$$\epsilon_{t+1}^{\mathbb{Q}} = \epsilon_{t+1} + \lambda_t \quad \text{and} \quad \epsilon_{t+1}^{\mathbb{L}} = \epsilon_{t+1}^{\mathbb{Q}} + \lambda_t^{\mathbb{L}}. \tag{4.4.16}$$

It follows that $\epsilon^{\mathbb{L}}_{t+1} = \epsilon_{t+1} + (\lambda_t + \lambda^{\mathbb{L}}_t)$. Further, suppose that S^P_t, which is a martingale under the $\mathbb{P}$ measure, follows the log-normal process:

$$\log S^P_{t+1} = \log S^P_t - \frac{1}{2}\omega_t^{'}\omega_t - \omega_t^{'}\epsilon_{t+1}, \tag{4.4.17}$$

where ω_t is the prices of risk related to permanent shocks since S^P is only the martingale component of a pricing kernel. In this sense, we refer to it as the *prices of long-term risk* hereafter. In detail, since $S^P = \zeta \cdot \xi$, the following equation holds by the Itô product rule:

$$\lambda_t = \omega_t - \lambda^{\mathbb{L}}_t. \tag{4.4.18}$$

This implies that $-\lambda^{\mathbb{L}}_t$ is defined as the difference between the overall market prices of risk (λ_t) and the prices of long-term risk (ω_t). In this sense, $-\lambda^{\mathbb{L}}_t$ can be considered as the market prices of risk associated with transitory shocks.

Note that, from (4.4.18), we see that when $\lambda_t = -\lambda^{\mathbb{L}}_t$ for all t, ω_t is a vector of zeros. In this case, from (4.4.17) the martingale component of the original SDF is degenerate ($s^P_{t,t+1} = 1$), and consequently the $\mathbb{P}$ and $\mathbb{L}$ measures become identical. This mathematical result is consistent with the previous literature mentioned before.

4.4.2.2 Decomposition of yields and term premia

Under the $\mathbb{L}$ dynamics (4.3.19), we can decompose yields into investors' interest rate expectations and term premia. Basically, this analysis can be conducted in the same way as in Section 4.2; however, each component should be interpreted differently.

Recall the decomposition of yields under the $\mathbb{P}$ measure, (4.2.14):

$$y^{(\tau)}_t = \widetilde{y}^{(\tau)}_t + ytp^{(\tau)}_t. \tag{4.4.19}$$

$\widetilde{y}^{(\tau)}_t$, which is investors' short-rate expectations under the $\mathbb{P}$ measure, is referred to as $\mathbb{P}$

risk-neutral rates. Also, $ytp_t^{(\tau)}$ is risk premia corresponding to overall shocks since $y_t^{(\tau)}$ is priced under the $\mathbb{Q}$ measure which entirely absorbs the original SDF.

Likewise, yields can be decomposed under the $\mathbb{L}$ measure such that

$$y_t = y_t^{\mathbb{L}} + ytp_t^{\mathbb{L}}. \tag{4.4.20}$$

For simplicity, the τ superscripts are suppressed here. $y_t^{\mathbb{L}}$ can be interpreted as the hypothetical yields as if our real world is governed by the $\mathbb{L}$ measure which absorbs risk premia corresponding with only permanent shocks. In this regard, we refer to $y_t^{\mathbb{L}}$ as the *long-term risk-neutral rates* (or equivalently, investors' short-rate expectations under the $\mathbb{L}$ measure).

On the other hand, $ytp_t^{\mathbb{L}}$ is defined as the difference between y_t and $y_t^{\mathbb{L}}$. Also, from (4.4.19) and (4.4.20), we have

$$ytp_t^{\mathbb{L}} = ytp_t - (y_t^{\mathbb{L}} - \widetilde{y}_t). \tag{4.4.21}$$

Obviously, the difference in parentheses is the long-term risk compensation since $y_t^{\mathbb{L}}$ absorbs risk compensation for exposure to permanent shocks, whereas $\widetilde{y}_t$ does not capture any risk compensation. Thus, $ytp_t^{\mathbb{L}}$ can be also defined as the difference between compensation for overall risk exposure and that for long-term risk exposure. In this sense, we refer to $ytp_t^{\mathbb{L}}$ as *short-term risk premia*.

Let ytp_t^{ω} denote the long-term risk compensation; that is, $y_t^{\mathbb{L}} - \widetilde{y}_t = ytp_t - ytp_t^{\mathbb{L}}$. Then, we have the following relation between long-term risk-neutral rates and $\mathbb{P}$ risk-neural rates:

$$y_t^{\mathbb{L}} = \widetilde{y}_t + ytp_t^{\omega}. \tag{4.4.22}$$

We refer to ytp_t^{ω} as *long-term risk premia* hereafter. Recall that we calculated $\widetilde{y}_t$ by using the $\mathbb{P}$ parameter estimates in Section 4.2.1. Similarly, we can compute $y_t^{\mathbb{L}}$ by using the parameter

estimates of the $\mathbb{L}$ dynamics:

$$y_t^{(\tau)\mathbb{L}} = A_\tau^{\mathbb{L}} + B_\tau^{\mathbb{L}\prime} X_t, \quad A_\tau^{\mathbb{L}} = -\frac{1}{\tau} A_\tau(\mu^{\mathbb{L}}, \Phi^{\mathbb{L}}, \delta_0, \delta_1, \Sigma), \quad B_\tau^{\mathbb{L}} = -\frac{1}{\tau} B_\tau(\Phi^{\mathbb{L}}, \delta_1). \tag{4.4.23}$$

Obviously, we also get

$$ytp_t = ytp_t^{\mathbb{L}} + ytp_t^{\omega}. \tag{4.4.24}$$

Thus, overall term premia are decomposed into short-term risk premia and long-term risk premia.

Note that the implication of (4.4.22) is consistent with that of (4.4.18). If long-term risk premia are zero, long-term risk-neutral rates and $\mathbb{P}$ risk-neutral rates are identical so that $\mathbb{P} = \mathbb{L}$. Equivalently, since risk premia are the product of the market prices of risk and the quantity of risk, when the market prices of long-term risk are zero in (4.4.18), we have $\mathbb{P} = \mathbb{L}$ as well.

In a similar fashion to yield decompositions, we can conduct forward rates decompositions as follows. For simplicity, the (τ_j, τ_k) superscripts are suppressed here.

$$f_t = \widetilde{f}_t + ftp_t, \qquad f_t = f_t^{\mathbb{L}} + ftp_t^{\mathbb{L}} \quad \text{and} \quad f_t^{\mathbb{L}} = \widetilde{f}_t + ftp_t^{\omega}. \tag{4.4.25}$$

We refer to $\widetilde{f}_t$ as $\mathbb{P}$ *risk-neutral forward rates*, ftp_t as *forward term premia*, $f_t^{\mathbb{L}}$ as *long-term risk-neutral forward rates*, ftp_t^{ω} as *long-term forward term premia*, and $ftp_t^{\mathbb{L}}$ as *short-term forward term premia* (i.e., $ftp_t^{\mathbb{L}} = ftp_t - ftp_t^{\omega}$).

4.5 Empirical Results

In Section 4.2.3, we analyzed GDTSMs for ten countries by estimating the $\mathbb{P}$ and $\mathbb{Q}$ state dynamics and decomposing five- to ten-year forward rates into overall term premia and $\mathbb{P}$ risk-neutral rates.

Now, we extend this analysis to a new world governed by the long-term risk-neutral

probability measure ($\mathbb{L}$). First, we recover a state dynamics under the $\mathbb{L}$ measure as described in Section 4.3.2. Second, as discussed in Section 4.4.2, forward term premia are extracted from forward rates under the $\mathbb{L}$ measure. Lastly, we identify what the Recovery theorem really recovers and how three different probability measures ($\mathbb{P}$, $\mathbb{Q}$ and $\mathbb{L}$) are linked with one another.

In this section, under a stationary assumption, we exclude six countries from a panel dataset, in which the $\mathbb{Q}$ state dynamics has a nearly unit root; specifically, the largest eigenvalue of $\Phi_{\mathcal{P}}^{\mathbb{Q}}$ is larger than 0.99 (see Table A.4.1). Excluded countries are Australia, Canada, New Zealand, Norway, Sweden and the UK.

4.5.1 Recovered State Dynamics

By using the GL method, the underlying $\mathbb{Q}$ dynamics is approximated by a discrete-state stationary Markov process. We choose $m_k = 9$ for all k so that $m^* = 729$. Gospodinov and Lkhagvasuren (2014) showed that the GL method with a moderate number of grid points (e.g., $m_k = 9$) provides a very precise approximation of the underlying process even for highly persistent data. Another example is Farmer (2014), which also used the GL method with 9 grid points along each dimension to estimate the shadow-rate term structure model. We obtain a set of discrete-valued 3×1 vectors $(z^1, z^2, \ldots, z^{729})$, and a 729×729 risk-neutral transition probability matrix ($\widetilde{P}$).

As described in Remark 1, we check the accuracy of the GL method. First, we generate time series for $10,000$ time periods with a burn-in period of $1,000$ based on grid points and $\widetilde{P}$. Also, we directly simulate a sequence of length $10,000$ with a burn-in period of $1,000$ based on the underlying VAR process. After repeating each simulation $1,000$ times, we calculate the mean of the estimated parameters. Then we compare the mean estimates obtained from two experiments. Table A.4.2 in Appendix reports the result of our quality check. Differences between parameter estimates are very small across all countries. They are nearly identical up to two decimal points.

Table 4.3: Estimation of the $\mathbb{L}$ Dynamics of Yield Factors

	$\mu^{\mathbb{L}}$		$\Phi^{\mathbb{L}}$						Σ					
	$\ddot{\mu}^{\mathbb{L}}$	*(s.e)*	$\ddot{\Phi}^{\mathbb{L}}$			*(s.e)*			$\ddot{\Sigma}$			*(s.e $\times 10^3$)*		
	0.0035	*(0.013)*	1.0071	0.1941	-0.8020	*(0.042)*	*(0.151)*	*(0.789)*	0.0217	0	0	*(0.076)*		
US	-0.0039	*(0.004)*	-0.0278	0.9459	0.8018	*(0.013)*	*(0.046)*	*(0.242)*	0.0007	0.0066	0	*(0.017)*	*(0.007)*	
	0.0060	*(0.002)*	0.0190	-0.0205	0.3652	*(0.006)*	*(0.022)*	*(0.116)*	-0.0007	-0.0015	0.0027	*(0.008)*	*(0.003)*	*(0.002)*
	0.0006	*(0.004)*	0.9767	0.1257	-0.5382	*(0.021)*	*(0.154)*	*(0.612)*	0.0139	0	0	*(0.032)*		
Japan	-0.0008	*(0.001)*	0.0085	0.9900	0.5424	*(0.007)*	*(0.050)*	*(0.198)*	0.0015	0.0042	0	*(0.008)*	*(0.003)*	
	0.0012	*(0.001)*	-0.0053	-0.0436	0.6883	*(0.004)*	*(0.030)*	*(0.118)*	-0.0022	-0.0001	0.0016	*(0.006)*	*(0.002)*	*(0.001)*
	0.0018	*(0.007)*	0.9720	0.1901	-0.7274	*(0.026)*	*(0.128)*	*(0.406)*	0.0161	0	0	*(0.042)*		
Ger.	-0.0016	*(0.003)*	0.0049	0.9588	0.6781	*(0.009)*	*(0.045)*	*(0.145)*	0.0006	0.0057	0	*(0.011)*	*(0.005)*	
	0.0025	*(0.002)*	-0.0037	-0.0178	0.5600	*(0.006)*	*(0.030)*	*(0.095)*	-0.0017	-0.0014	0.0031	*(0.008)*	*(0.003)*	*(0.002)*
	0.0008	*(0.008)*	0.9258	0.2745	0.6059	*(0.026)*	*(0.175)*	*(0.496)*	0.0154	0	0	*(0.038)*		
Switz.	-0.0006	*(0.003)*	0.0226	0.9104	-0.4983	*(0.009)*	*(0.061)*	*(0.174)*	0.0008	0.0053	0	*(0.010)*	*(0.005)*	
	-0.0009	*(0.002)*	0.0019	0.0055	0.8407	*(0.005)*	*(0.033)*	*(0.094)*	0.0007	0.0013	0.0025	*(0.005)*	*(0.002)*	*(0.001)*

Note: This table reports the parameter estimates of the discretized VAR(1) process (4.3.18) under the $\mathbb{L}$ measure, which is induced from the GL method with $m_k = 9$ for all k, and standard errors. All the estimates are reported on an annual basis (by multiplying 4).

Next, we construct a 729×729 state-price matrix (Q) by using the one-period (three-month) interest rates and grid points as in Section 4.3.2.2. Moreover, we confirm that Q is non-negative and irreducible for all four countries as described in Remark 2. In what follows, we obtain the perron root and perron vector of the irreducible Q by employing the power method and compute the recovered state transition probability matrix ($\widehat{P}$). Eventually, we estimate the $\mathbb{L}$ dynamics of the state variables as described in Section 4.3.2.4. Table 4.3 reports parameter estimates for the $\mathbb{L}$ dynamics.

For a robustness check, we increase the number of grid points along each dimension up to $m_k = 21$ ($m^* = 9261$) and repeat the same steps as above. Gospodinov and Lkhagvasuren (2014) noted that in a highly persistent multivariate process, adjusting the number of grid points is not always the best approach to improve approximation quality due to the cross-correlations between factors. In fact, our results with a larger number of grid points ($m_k = 21$) are not much different from our baseline results with $m_k = 9$ (see Table A.4.2 in Appendix). In addition, we employ the method of Terry and Knotek (2011) as an alternative to the GL method; however, we fail to obtain irreducible Q matrices for all countries, except for the US. Even though we adjust the number of grid points between $m_k = 3$ and 27, the method of Terry and Knotek keeps producing reducible matrices.

Table 4.4: Comparisons of Summary Statistics w.r.t. Different Probability Measures

	Max Eig($\Phi_{\mathcal{P}}$)			IRF			Half-life			Volatility			$\triangle$(90-91/08-09)		
	$\mathbb{Q}$	$\mathbb{P}$(BC)	$\mathbb{L}$	$\mathbb{Q}$	$\mathbb{P}$(BC)	$\mathbb{L}$	$\mathbb{Q}$	$\mathbb{P}$(BC)	$\mathbb{L}$	f_t	$\widetilde{f}_t$	$f_t^{\mathbb{L}}$	f_t	$\widetilde{f}_t$	$f_t^{\mathbb{L}}$
US	0.9711	0.9862	0.9708	0.60	0.76	0.59	27.00	49.00	26.00	1.25	0.84	1.17	-387	-239	-367
Japan	0.9873	0.9844	0.9831	0.83	0.72	0.76	63.00	35.00	47.00	1.77	1.47	1.54	-473	-398	-424
Ger.	0.9737	0.9997	0.9718	0.66	0.96	0.64	32.00	-	30.00	1.48	2.09	1.39	-407	-614	-387
Switz.	0.9892	0.9917	0.9851	0.55	0.86	0.52	30.00	52.00	23.00	1.14	2.09	1.03	-259	-477	-236

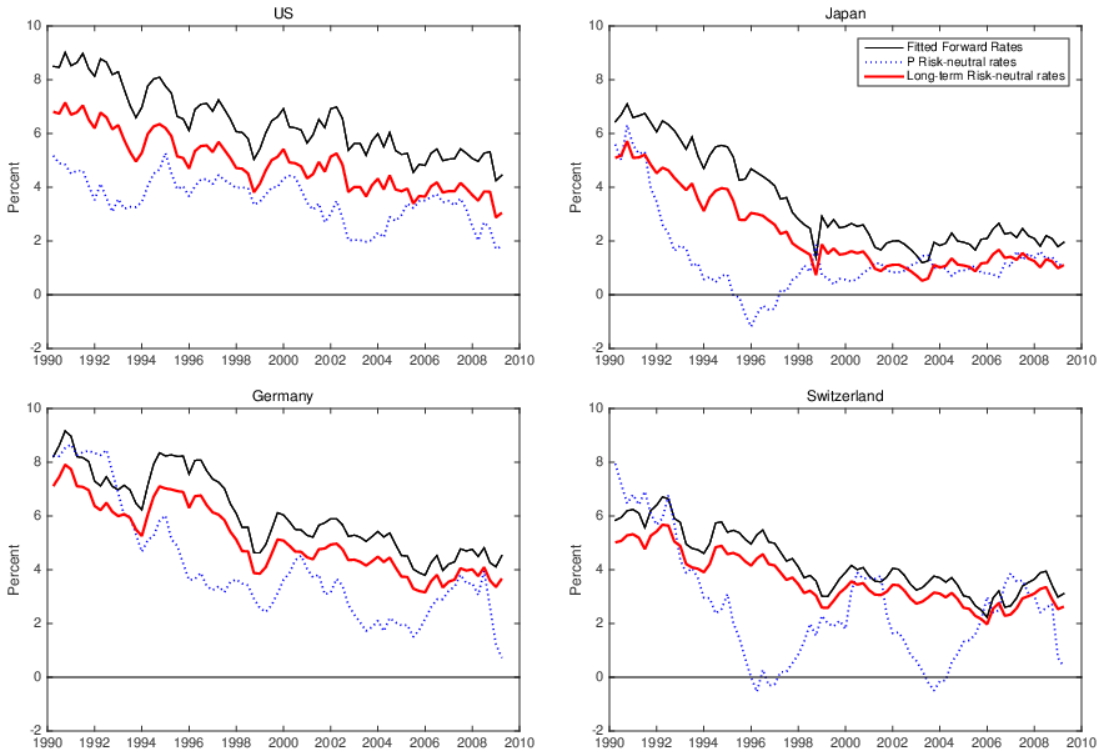
Note: (i) Each statistic under $\mathbb{P}$ and $\mathbb{Q}$ is computed using BC estimates. (ii) IRF is the impulse-response function at horizon of five years of the first yield factors to a level shock. (iii) Half-life is the horizon (quarters) at which the IRF falls first below 0.5. If a computed half-life is larger than 50 years, we do not report it. (iv) The "Volatility" columns report the standard deviations of risk-neutral prices and short-rate expectations under different measures. (v) The last three columns show changes from 1990-1991 to 2008-2009 which are computed by the difference between the mean of observations from 1990:III to 1991:III (the early part of the sample) and from 2008:I to 2009:I (the late part of the sample.) We report in basis points. As defined in (4.4.25), f_t is the fitted five- to ten-year forward rates, $\widetilde{f}_t$ is $\mathbb{P}$ risk-neutral forward rates, and $f_t^{\mathbb{L}}$ is long-term risk-neutral forward rates.

4.5.2 GDTSM Analysis under the Recovered Probability Measure

To figure out the implications of our recovery results, Table 4.4 reports summary statistics of the estimated state dynamics under three different probability measures: the physical measure $\mathbb{P}$, the risk-neutral measure $\mathbb{Q}$, and the recovered measure $\mathbb{L}$. For the $\mathbb{P}$ and $\mathbb{Q}$ measures, we only report statistics obtained from *BC* two-step procedures since the $\mathbb{P}$ persistence obtained from *JSZ* two-step procedures would not reflect reasonable persistence as shown in Section 4.2. Note that statistics under the $\mathbb{Q}$ measure remain the same regardless of bias correction.

Across various measures of persistence (the largest eigenvalue of $\Phi_{\mathcal{P}}$, the impulse response function, and the half-life), we can see that the $\mathbb{L}$ measure produces very similar statistics to the $\mathbb{Q}$ measure, while the $\mathbb{P}$ measure yields very different statistics from the $\mathbb{Q}$ measure (excepting Japan). This is consistent with previous studies which provide theoretical and empirical evidence on the similarity between the $\mathbb{Q}$ and $\mathbb{L}$ measures (see Section 4.4.1). This similarity can be found in the "Volatility" columns as well. The volatilities of long-term risk-neutral forward rates ($f_t^{\mathbb{L}}$) are very similar to those of forward rates (f_t), while they differ greatly from those of $\mathbb{P}$ risk-neutral rates ($\widetilde{f}_t$). It implies that the movement of the fitted forward rates can be explained better by long-term risk-neutral rates than by $\mathbb{P}$ risk-neutral rates. Moreover, we can see the same result across all countries in terms of the changes from

Figure 4.1: Decomposition of Forward Rates – Fitted/Risk-neutral/Long-term Risk-neutral

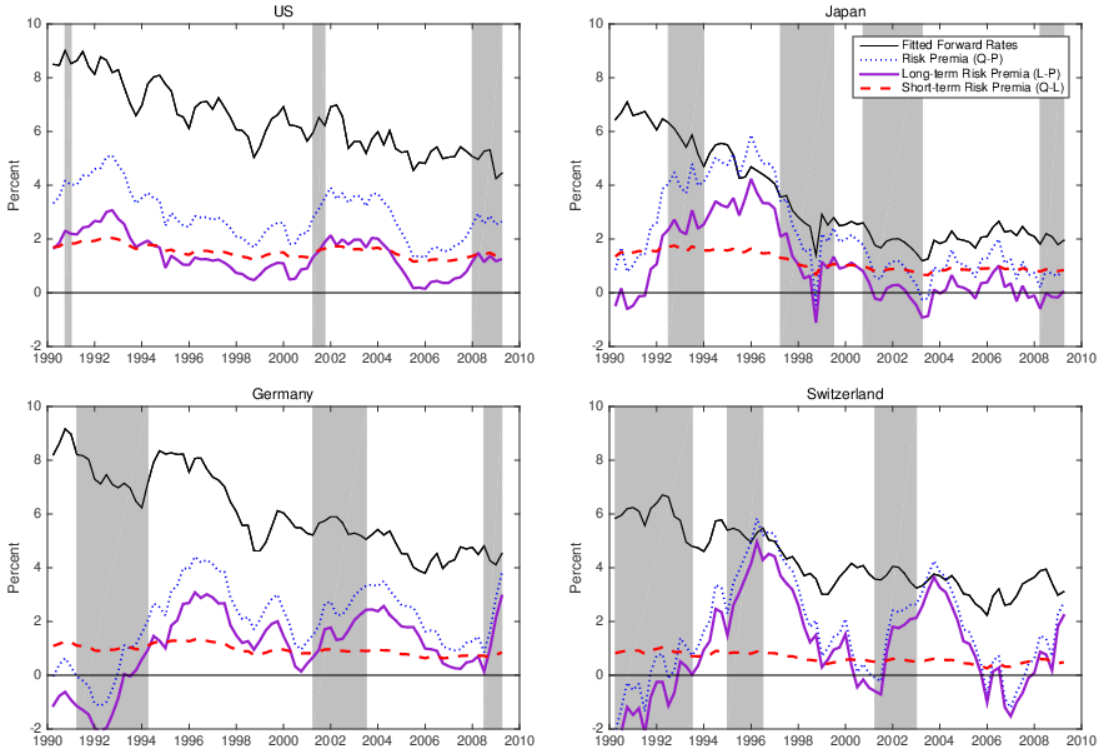


Note: This figure plots the five- to ten-year fitted forward rates, $\mathbb{P}$ risk-neutral forward rates (short-term interest rate expectations) estimated by *BC* two-step procedures, and long-term risk-neutral forward rates estimated by the Recovery theorem.

the early to late sample periods of each component.

Figure 4.1 depicts the historical evolutions of the fitted five- to ten-year forward rates, $\mathbb{P}$ risk-neutral forward rates obtained by *BC* estimates, and long-term risk-neutral forward rates. There are distinct downward trends in long-term risk-neutral forward rates as shown in the fitted forward rates as well; more precisely, both are nearly parallel to each other. Thus, long-term risk-neutral rates contribute extensively to the secular declining trend and volatility of the fitted forward rates. Comparing them with $\mathbb{P}$ risk-neutral forward rates, however, we can see significant differences with respect to level and slope. Across all countries, a gap between the fitted forward rates and long-term risk-neutral forward rates ($f_t - f_t^{\mathbb{L}}$) is pretty small relative to a gap between the fitted forward rates and $\mathbb{P}$ risk-neutral rates ($f_t - \widetilde{f}_t$). Consequently, the former, which implies short-term forward term premia ($ftp_t^{\mathbb{L}}$), should be

Figure 4.2: Decomposition of Forward Term Premia – ftp_t, ftp_t^{ω}, and $ftp_t^{\mathbb{L}}$



Note: This figure plots the five- to ten-year fitted forward rates and the corresponding term premia that are estimated by *JSZ* two-step procedures and *BC* two-step procedures across 10 countries. For each country, the recession periods are indicated by shaded area. Without loss of generality, actual forward rates are omitted, since fitting errors are small.

smaller and flatter than the latter, which captures overall forward term premia (ftp_t).

Our empirical findings coincide very well with theoretical results from the previous literature: (i) the similarity between $\mathbb{Q}$ and $\mathbb{L}$, and (ii) the argument of "misspecified recovery"($\mathbb{P} \neq \mathbb{L}$). Concisely, the recovered investors' expectations from the Recovery theorem are inconsistent with the investors' true (physical) expectations; rather, the recovered expectations represent investors' expectations adjusted by their aversion to long-term risk.

4.5.3 Long-term Risk Premia

Now, we examine term premia in detail. As described in Section 4.4.2, we compute forward term premia by the difference between the fitted forward rates and $\mathbb{P}$ risk-neutral forward rates estimated from *BC* two-step procedures ($ftp_t = f_t - \widetilde{f}_t$). Also, short-term

forward term premia is computed by the difference between the fitted forward rates and long-term risk-neutral forward rates ($ftp_t^{\mathbb{L}} = f_t - f_t^{\mathbb{L}}$). Lastly, we obtain long-term forward term premia (ftp_t^{ω}), which compensate risk-averse investors for exposure to permanent shocks, from the term premia decomposition (4.4.24).

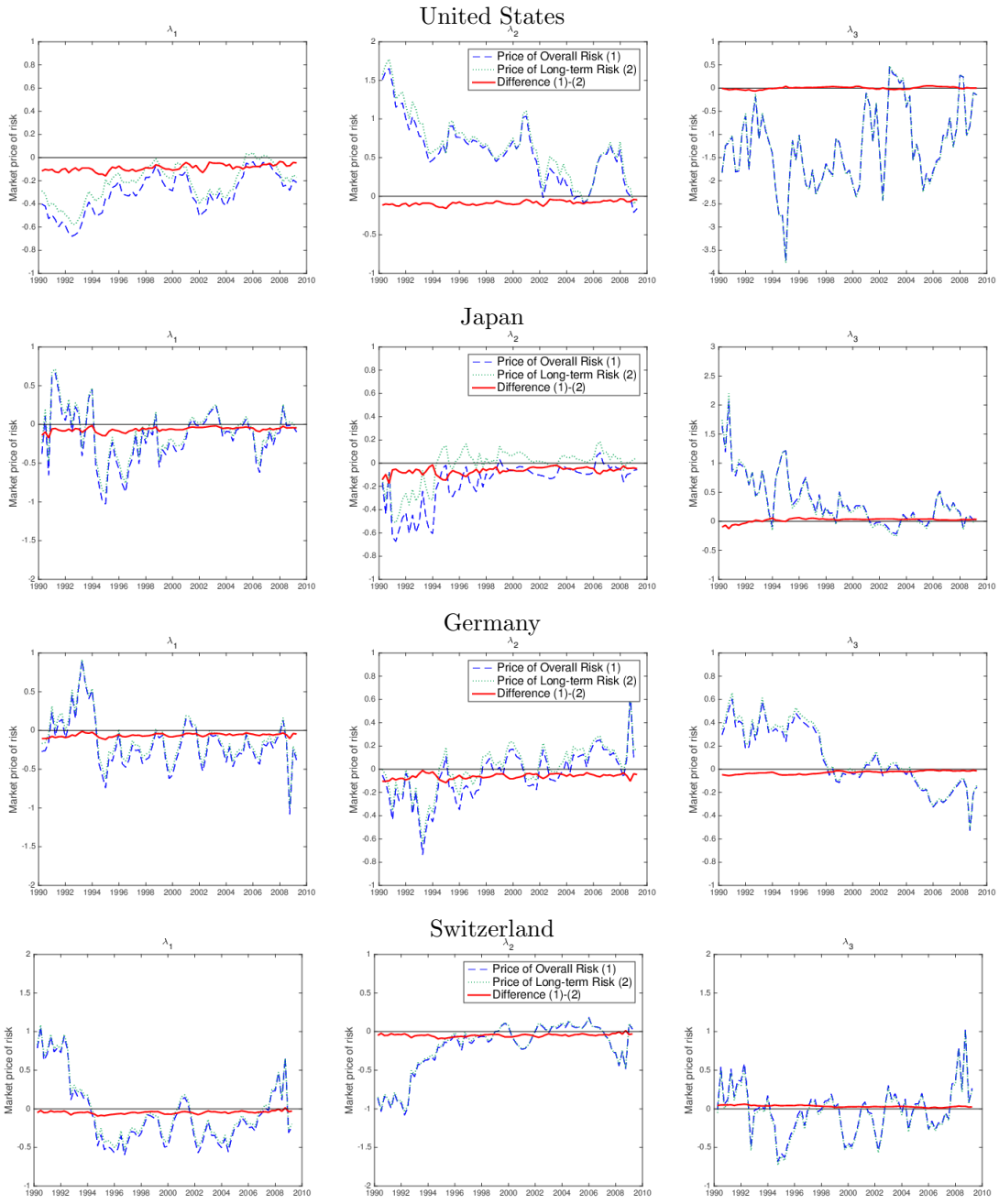
Figure 4.2 depicts the historical evolutions of f_t, ftp_t, $ftp_t^{\mathbb{L}}$, and ftp_t^{ω}. First, overall term premia (ftp_t) and long-term forward term premia (ftp_t^{ω}) are highly volatile over the sample period. Such high volatilities of ftp_t and ftp_t^{ω} are very plausible in the sense that long-term forward term premia contain a martingale component in the original SDF (or equivalently, risk compensation corresponding to permanent shocks), and this martingale component dominates the overall behavior of the original SDF. On the other hand, short-term forward term premia ($ftp_t^{\mathbb{L}}$) are nearly constant and very stable over time. This is because $ftp_t^{\mathbb{L}}$ is risk compensation associated with only transitory shock, or equivalently it contains the transitory component of the original SDF.

Second, overall term premia (ftp_t) and long-term forward term premia (ftp_t^{ω}) almost parallel each other. Contrarily, short-term forward term premia ($ftp_t^{\mathbb{L}}$) are clearly distinguished from them. In addition, $ftp_t^{\mathbb{L}}$ is relatively small in level, and its magnitude is not much different across countries, while ftp_t varies considerably across countries. All of these results imply that overall term premia are mostly attributed to long-term premia, while short-term premia do not significantly affect overall term premia.

Finally, we can easily confirm $\mathbb{L} \neq \mathbb{P}$ over the sample period. In Figure 4.2, we can see clearly that ftp_t^{ω} is extremely volatile and very far from zero for most of the period. Also, $ftp_t^{\mathbb{L}}$ is much different from ftp_t in level and volatility. The misspecified $\mathbb{P}$ measure (actually, the $\mathbb{L}$ measure) mislead us about term premia. In this case, term premia ($ftp_t^{\mathbb{L}}$) are neither as sizable nor time-varying as they should be under the true $\mathbb{P}$ measure.

We can also examine whether or not $\mathbb{L} \neq \mathbb{P}$ in terms of the market prices of risk. In Section 4.4, we theoretically show that if ω_t is a vector of zeros, the $\mathbb{L}$ measure is identical with the $\mathbb{P}$ measure. From (4.4.18), a 3×1 zero vector ω_t implies that $\lambda_t = -\lambda_t^{\mathbb{L}}$, where λ_t

Figure 4.3: Market Prices of Risk Factors



Note: This figure plots the historical evolutions of each component of the market price of overall risk denoted by λ_t, the price of the long-term risk denoted by ω_t, and their difference denoted by $\lambda_t^{\mathbb{L}}$.

is a 3×1 vector of the market prices of overall risk, and $-\lambda_t^{\mathbb{L}}$ is a 3×1 vector of risk prices associated with transitory shocks (difference between λ_t and ω_t). Figure 4.3 depicts the historical evolutions of each element of λ_t, ω_t, and $-\lambda_t^{\mathbb{L}}$ for all countries. The each element

of ω_t is highly volatile and much far away from zero over the sample period. λ_t and $-\lambda_t^{\mathbb{L}}$ are significantly different from each other across all countries in level and slope. Overall, our empirical result supports $\mathbb{L} \neq \mathbb{P}$; that is, the Recovery theorem fails to identify the physical probability measure.

4.6 Concluding Remarks

In this paper, we revisit the Recovery theorem proposed by Ross (2015). In particular, its relevancy and reliability are examined in the framework of the affine Gaussian dynamic term structure model.

First, we apply the Recovery theorem to GDTSM. Using an international panel dataset, a certain probability measure ($\mathbb{L}$) is recovered from state prices constructed by the finite-state Markov-chain approximation method of Gospodinov and Lkhagvasuren (2014), and the state dynamics under the $\mathbb{L}$ measure is estimated. Also, under the $\mathbb{L}$ measure, forward rates are decomposed into investors' short-rate expectations and term premia. For a benchmark against which the Recovery theorem can be tested, we estimate the physical probability measure ($\mathbb{P}$) and a corresponding state dynamics under the $\mathbb{P}$ measure by a conventional maximum likelihood estimation with bias correction. The forward rates decomposition is also conducted under the $\mathbb{P}$ measure.

Second, we provide strong evidence showing that the Recovery theorem misspecifies the physical probability measure. We verify the identity between our $\mathbb{L}$ measure and the long-term risk-neutral measure defined by BHS (2015). Meanwhile, we find the conditions for $\mathbb{P} = \mathbb{L}$ in terms of risk premia as well as the market prices of risk. Our empirical result shows that investors' short-rate expectations and term premia under the $\mathbb{P}$ measure are substantially different from those under the $\mathbb{L}$ measure. Moreover, we characterize term premia under $\mathbb{L}$ as the short-term premia associated with transitory shocks; hence, long-term risk premia corresponding to permanent (martingale) shocks can be extracted from overall risk premia. Our empirical result shows that short-term risk premia are nearly constant

over time, while long-term risk premia are highly volatile and almost parallel overall term premia. Consequently, the secular downward trend and volatility of forward rates are mostly attributed to investors' short-rate expectations under the long-term risk-neutral probability measure, and all important variations in overall term premia can be captured by long-term risk premia. Our result demonstrates that long-term risk matters for asset pricing.

Several extensions are left for future research. As mentioned before, there exists the statistical uncertainty around the point estimates of the $\mathbb{P}$ parameters estimated from BC two-step procedures. Since $\mathbb{P}$ estimates are used as a benchmark, we can seek to validate our result by using alternative estimation procedures or under different model specifications in the GDTSM literature; for example, we may consider the minimum-chi-square estimation of Hamilton and Wu (2012), the use of survey data as additional information as in Kim and Orphanides (2012), risk-parameter restrictions by Bauer (2016), and macro-factor models by Joslin et al. (2014) and Creal and Wu (2015). Next, although a fully specified state-price matrix is necessary for the application of Recovery theorem, it is not practically easy in equity markets. Thus, it is worth checking whether or not a Markov approximation method employed in our fixed-income market study is applicable to equity markets. Moreover, we can examine policy implications on long-term risk premia and investors' long-term risk-neutral expectations.

Appendix

A.4.1. Girsanov's Theorem in a Discrete-time Specification

Let us define an $N \times 1$ vector $\epsilon_t \sim \mathcal{N}(0, I_N)$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathbb{P}$ is the physical probability measure. Let W_t denote the column vector $(W_{1,t}, W_{2,t}, \ldots, W_{N,t})'$. We define an N-dimensional discrete-time Brownian Motion: $W_0 = 0$ and $W_t = \sum_{t=1}^{T} \epsilon_t$, for $t = 1, 2, \ldots, T$. Equivalently, $\epsilon_t = W_t - W_{t-1}$; that is, ϵ are the increments to W. We then find that $\{W_t\}_{t=0}^{T}$ is a martingale under $\mathbb{P}$ and a Markov process as well (Shreve, 2004, Theorem 3.3.4 and 3.5.1). Likewise, let $\widehat{W}_t$ denote a Brownian Motion under another probability measure, $\widehat{\mathbb{P}}$.

Define the random variable Z as

$$Z_t = \exp\left(\sum_{t=1}^{T} -\gamma_t \epsilon_t - \sum_{t=1}^{T} \frac{1}{2}||\gamma_t||^2\right),$$

where $||\cdot||$ denotes the Euclidean norm such that $||\gamma_t|| = \left(\sum_{j=1}^{N} \gamma_{j,t}^2\right)^{1/2}$ for $j = 1, 2, \ldots, N$, and γ_t is called the market prices of risk which are the unit prices of bearing exposure to the increment of W_t. Also, consider

$$\widehat{\epsilon}_{t+1} = \epsilon_{t+1} + \gamma_t,$$

where $\widehat{\epsilon}_t = \widehat{W}_t - \widehat{W}_{t-1}$. More precisely, $\widehat{\epsilon}_t = (\widehat{\epsilon}_{1,t}, \widehat{\epsilon}_{2,t}, \ldots, \widehat{\epsilon}_{N,t})'$ and $\widehat{\epsilon}_{j,t+1} = \epsilon_{j,t+1} + \gamma_{j,t}$.

Then, setting $Z = Z(T)$, $E(Z) = 1$ and the process $\widehat{W}_t$ is an N-dimensional discrete-time Brownian Motion under the $\widehat{\mathbb{P}}$ measure given by

$$\widehat{\mathbb{P}}(A) = \int_A Z(\alpha) d\mathbb{P}(\alpha), \quad \text{for all} \quad A \in \mathcal{F}.$$

We say Z is the Radon-Nikodym derivative of $\widehat{\mathbb{P}}$ with respect to $\mathbb{P}$, and write it as

$$Z = \frac{d\widehat{\mathbb{P}}}{d\mathbb{P}}.$$

Using this Radon-Nikodym derivative, we can find the following relation between two different expectations: the expectation under the original $\mathbb{P}$ measure denoted by $E(X)$ and the expectation under the new probability measure ($\widehat{\mathbb{P}}$) denoted by $\widehat{E}(X)$. For any random variable X, we have

$$\widehat{E}(X) = E(XZ).$$

(Shreve, 2004, Theorem 5.2.3 and Theorem 5.4.1; Duffie, 2010, Ch.6).

A.4.2. Risk-neutral dynamics/ Change of measure

Recall (4.2.1). Consider the dynamics of latent factors under the physical measure ($\mathbb{P}$):

$$X_{t+1} = \mu + \Phi X_t + \Sigma\epsilon_{t+1},$$

where $\epsilon_t \sim \mathcal{N}(0, I_N)$. The Radon-Nikodym derivative process is given by

$$\frac{\xi_{t+1}}{\xi_t} = \exp\left(-\frac{1}{2}\lambda_t'\lambda_t - \lambda_t'\epsilon_{t+1}\right)$$

as in (4.2.6). As defined in (4.2.3), the one-period stochastic discount factor is defined as

$$\frac{S_{t+1}}{S_t} = \exp(-r_t)\frac{\xi_{t+1}}{\xi_t} = \exp\left(-r_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t'\epsilon_{t+1}\right).$$

Also, the market prices of risk are given by $\lambda_t = \Sigma^{-1}(\lambda_0 + \lambda_1 X_t)$ in (4.2.4).

By Shreve (2004, Lemma 5.22) and our risk-price specification, we can derive the conditional moment generating function of a multivariate normal distribution as follows: Since $E^{\mathbb{Q}}(Y|\mathcal{F}_s) = \frac{1}{\xi_s}E(Y\xi_t|\mathcal{F}_s)$ for $0 \le s \le t \le T$, where Y is an $\mathcal{F}_t$-measurable random variable,

we get

$$\begin{aligned}
E^{\mathbb{Q}}\left(\exp(u'X_{t+1})|X_t\right) &= \frac{1}{\xi_t}E(\exp(u'X_{t+1})\cdot\xi_{t+1}|X_t) \\
&= E\left[\exp\left(u'X_{t+1}-\frac{1}{2}\lambda_t^{'}\lambda_t-\lambda_t^{'}\epsilon_{t+1}\right)|X_t\right] \\
&= E\left[\exp\left(u'(\mu+\Phi X_t+\Sigma\epsilon_{t+1})-\frac{1}{2}\lambda_t^{'}\lambda_t-\lambda_t^{'}\epsilon_{t+1}\right)|X_t\right] \\
&= \exp\left(u'(\mu+\Phi X_t-\Sigma\lambda_t)+\frac{1}{2}u'\Sigma\Sigma'u\right) \\
&= \exp\left(u'(\mu-\lambda_0+(\Phi-\lambda_1)X_t)+\frac{1}{2}u'\Sigma\Sigma'u\right).
\end{aligned}$$

Then, the above result implies that the $\mathbb{Q}$ dynamics of X_t is

$$X_{t+1}=\mu^{\mathbb{Q}}+\Phi^{\mathbb{Q}}X_t+\Sigma\epsilon_{t+1}^{\mathbb{Q}},$$

where $\epsilon_t^{\mathbb{Q}}\sim\mathcal{N}(0,I_N)$,

$$\mu^{\mathbb{Q}}=\mu-\lambda_0,\quad \text{and}\quad \Phi^{\mathbb{Q}}=\Phi-\lambda_1.$$

Moreover, by the Girsanov's theorem, we have $\epsilon_{t+1}^{\mathbb{Q}}=\epsilon_{t+1}+\lambda_t$ and the volatility of the state vector (Σ) stays the same under both measures. For the same analysis in a continuous-time specification, see Shreve (2004, p. 213), Piazzesi (2010, p. 702) and Duffie (2010, Ch.6 and Appendix D).

A.4.3. Parameter estimates of the recovered state dynamics

First, obtain the moments of the discretized Markov process $\widetilde{\mathcal{P}}_t=(\widetilde{\mathcal{P}}_{1,t},\widetilde{\mathcal{P}}_{2,t},\widetilde{\mathcal{P}}_{3,t})'$. The conditional mean of $\widetilde{\mathcal{P}}_{k,t+1}$ given a current state i is defined as

$$E^{\mathbb{L}}(\widetilde{\mathcal{P}}_{k,t+1}|\widetilde{\mathcal{P}}_{k,t}=z_k^i)=\textstyle\sum_{j=1}^{m^*}\widehat{p}_{ij}z_k^j=\bar{Z}_k^{(i)},$$

for $k=1,2,3$, $i,j=1,2,\ldots,m^*$, and where $\bar{Z}_k^{(i)}$ denotes the conditional mean of $\widetilde{\mathcal{P}}_{k,t+1}$. Next, the unconditional moments can be defined using the stationary distribution, π, of a finite-state Markov-chain process as follows:

$$E^{\mathbb{L}}(\widetilde{\mathcal{P}}_{k,t}) = \sum_{j=1}^{m^*} \pi_j z_k^j = \bar{Z}_k,$$
$$Var^{\mathbb{L}}(\widetilde{\mathcal{P}}_{k,t}) = \sum_{j=1}^{m^*} \pi_j (z_k^j - \bar{Z}_k)^2,$$
$$Cov^{\mathbb{L}}(\widetilde{\mathcal{P}}_{k,t}, \widetilde{\mathcal{P}}_{h,t}) = \sum_{j=1}^{m^*} \pi_j (z_k^j - \bar{Z}_k)(z_h^j - \bar{Z}_h),$$

for $k, h = 1, 2, 3$, where $\bar{Z}_k$ denotes the unconditional mean of $\widetilde{\mathcal{P}}_{k,t}$, and π_j is the j-th element of an $m^* \times 1$ vector of π that satisfies $\pi_j = \sum_{i=1}^{m^*} \pi_i \cdot \widehat{p}_{ij}$ (Jiang, 2010, p. 324). Moreover, we can obtain additional moments as follows:

$$E^{\mathbb{L}}(\widetilde{\mathcal{P}}_{k,t}^2) = \sum_{j=1}^{m^*} \pi_j (z_k^j)^2 = \bar{Z}_k^2,$$
$$E^{\mathbb{L}}(\widetilde{\mathcal{P}}_{k,t}\widetilde{\mathcal{P}}_{h,t}) = \sum_{j=1}^{m^*} \pi_j z_k^j z_h^j,$$
$$E^{\mathbb{L}}(\widetilde{\mathcal{P}}_{k,t+1}\widetilde{\mathcal{P}}_{k,t}) = \sum_{j=1}^{m^*} \pi_j z_k^j \sum_{l=1}^{m^*} \widehat{p}_{jl} z_k^l = \sum_{j=1}^{m^*} \pi_j z_k^j \bar{Z}_k^{(j)},$$
$$E^{\mathbb{L}}(\widetilde{\mathcal{P}}_{k,t+1}\widetilde{\mathcal{P}}_{h,t}) = \sum_{j=1}^{m^*} \pi_j z_h^j \sum_{l=1}^{m^*} \widehat{p}_{jl} z_k^l = \sum_{j=1}^{m^*} \pi_j z_h^j \bar{Z}_k^{(j)},$$

for $k, j = 1, 2, 3$, and $i, j, l = 1, 2, \ldots, m^*$.

Next, we can estimate parameters induced by $\mathbb{L}$ measure from $B = [E^{\mathbb{L}}(\widetilde{\mathcal{P}}_{t+1}W_t')][E^{\mathbb{L}}(W_t W_t')]^{-1}$, where

$$E^{\mathbb{L}}(W_t W_t') = \begin{bmatrix} \sum_{j=1}^{m^*} \pi_j & \sum_{j=1}^{m^*} \pi_j z_1^j & \sum_{j=1}^{m^*} \pi_j z_2^j & \sum_{j=1}^{m^*} \pi_j z_3^j \\ \sum_{j=1}^{m^*} \pi_j z_1^j & \sum_{j=1}^{m^*} \pi_j (z_1^j)^2 & \sum_{j=1}^{m^*} \pi_j z_1^j z_2^j & \sum_{j=1}^{m^*} \pi_j z_1^j z_3^j \\ \sum_{j=1}^{m^*} \pi_j z_2^j & \sum_{j=1}^{m^*} \pi_j z_2^j z_1^j & \sum_{j=1}^{m^*} \pi_j (z_2^j)^2 & \sum_{j=1}^{m^*} \pi_j z_2^j z_3^j \\ \sum_{j=1}^{m^*} \pi_j z_3^j & \sum_{j=1}^{m^*} \pi_j z_3^j z_1^j & \sum_{j=1}^{m^*} \pi_j z_3^j z_2^j & \sum_{j=1}^{m^*} \pi_j (z_3^j)^2 \end{bmatrix},$$

and

$$E^{\mathbb{L}}(\widetilde{\mathcal{P}}_{t+1}W_t') = \begin{bmatrix} \sum_{j=1}^{m^*} \pi_j z_1^j & \sum_{j=1}^{m^*} \pi_j z_1^j \bar{Z}_1^{(j)} & \sum_{j=1}^{m^*} \pi_j z_2^j \bar{Z}_1^{(j)} & \sum_{j=1}^{m^*} \pi_j z_3^j \bar{Z}_1^{(j)} \\ \sum_{j=1}^{m^*} \pi_j z_2^j & \sum_{j=1}^{m^*} \pi_j z_1^j \bar{Z}_2^{(j)} & \sum_{j=1}^{m^*} \pi_j z_2^j \bar{Z}_2^{(j)} & \sum_{j=1}^{m^*} \pi_j z_3^j \bar{Z}_2^{(j)} \\ \sum_{j=1}^{m^*} \pi_j z_3^j & \sum_{j=1}^{m^*} \pi_j z_1^j \bar{Z}_3^{(j)} & \sum_{j=1}^{m^*} \pi_j z_2^j \bar{Z}_3^{(j)} & \sum_{j=1}^{m^*} \pi_j z_3^j \bar{Z}_3^{(j)} \end{bmatrix}.$$

A.4.4. Long-term Risk-neutral dynamics/ Change of measure

As shown in Appendix A.4.2, we derive the conditional moment generating function of a multivariate normal distribution as follows:

Since $E^{\mathbb{Q}}(Y|\mathcal{F}_s) = \frac{1}{\xi_s}E(Y\xi_t|\mathcal{F}_s)$ and $E^{\mathbb{L}}(Y|\mathcal{F}_s) = \frac{1}{\zeta_s}E^{\mathbb{Q}}(Y\zeta_t|\mathcal{F}_s)$ for $0 \leq s \leq t \leq T$, where Y is an $\mathcal{F}_t$-measurable random variable, we get

$$\begin{aligned}
E^{\mathbb{L}}\left(\exp(u'X_{t+1})|X_t\right) &= \frac{1}{\zeta_t}\, E^{\mathbb{Q}}\left(\exp(u'X_{t+1})\cdot\zeta_{t+1}|X_t\right) \\
&= \frac{1}{\zeta_t}\frac{1}{\xi_t}E\left(\exp(u'X_{t+1})\cdot\xi_{t+1}\zeta_{t+1}|X_t\right) \\
&= E\left[\exp\left(u'X_{t+1} - \frac{1}{2}(\lambda_t'\lambda_t + \lambda_t^{\mathbb{L}'}\lambda_t^{\mathbb{L}}) - \lambda_t'\epsilon_{t+1} - \lambda_t^{\mathbb{L}'}\epsilon_{t+1}^{\mathbb{Q}}\right)|X_t\right] \\
&= E\left[\exp\left(u'(\mu+\Phi X_t) - \frac{1}{2}(\lambda_t'\lambda_t + \lambda_t^{\mathbb{L}'}\lambda_t^{\mathbb{L}}) - \lambda_t^{\mathbb{L}'}\lambda_t + (u_t' - \lambda_t^{\mathbb{L}'})\epsilon_{t+1}\right)|X_t\right] \\
&= \exp\left(u'(\mu+\Phi X_t - \Sigma\lambda_t - \Sigma\lambda_t^{\mathbb{L}}) + \frac{1}{2}u'\Sigma\Sigma'u\right) \\
&= \exp\left(u'\left(\mu - \lambda_0 - \lambda_0^{\mathbb{L}} + (\Phi - \lambda_1 - \lambda_1^{\mathbb{L}})X_t\right) + \frac{1}{2}u'\Sigma\Sigma'u\right).
\end{aligned}$$

Then, the above result implies that the $\mathbb{L}$ dynamics of X_t is

$$X_{t+1} = \mu^{\mathbb{L}} + \Phi^{\mathbb{L}}X_t + \Sigma\epsilon_{t+1}^{\mathbb{L}},$$

where $\mu^{\mathbb{L}} = \mu^{\mathbb{Q}} - \lambda_0^{\mathbb{L}}$ and $\Phi^{\mathbb{L}} = \Phi^{\mathbb{Q}} - \lambda_0^{\mathbb{L}}$ since $\mu^{\mathbb{Q}} = \mu - \lambda_0$ and $\Phi^{\mathbb{Q}} = \Phi - \lambda_1$. Moreover, by the Girsanov's theorem, we have $\epsilon_{t+1}^{\mathbb{L}} = \epsilon_{t+1}^{\mathbb{Q}} + \lambda_t^{\mathbb{L}}$, where $\epsilon_t^{\mathbb{L}} \sim \mathcal{N}(0, I_N)$. The volatility of the state vector (Σ) remains the same under different measures.

Next, recall

$$\frac{d\mathbb{L}}{d\mathbb{P}} = \frac{d\mathbb{L}}{d\mathbb{Q}}\cdot\frac{d\mathbb{Q}}{d\mathbb{P}}, \quad \text{or equivalently} \quad S^P = \zeta\cdot\xi,$$

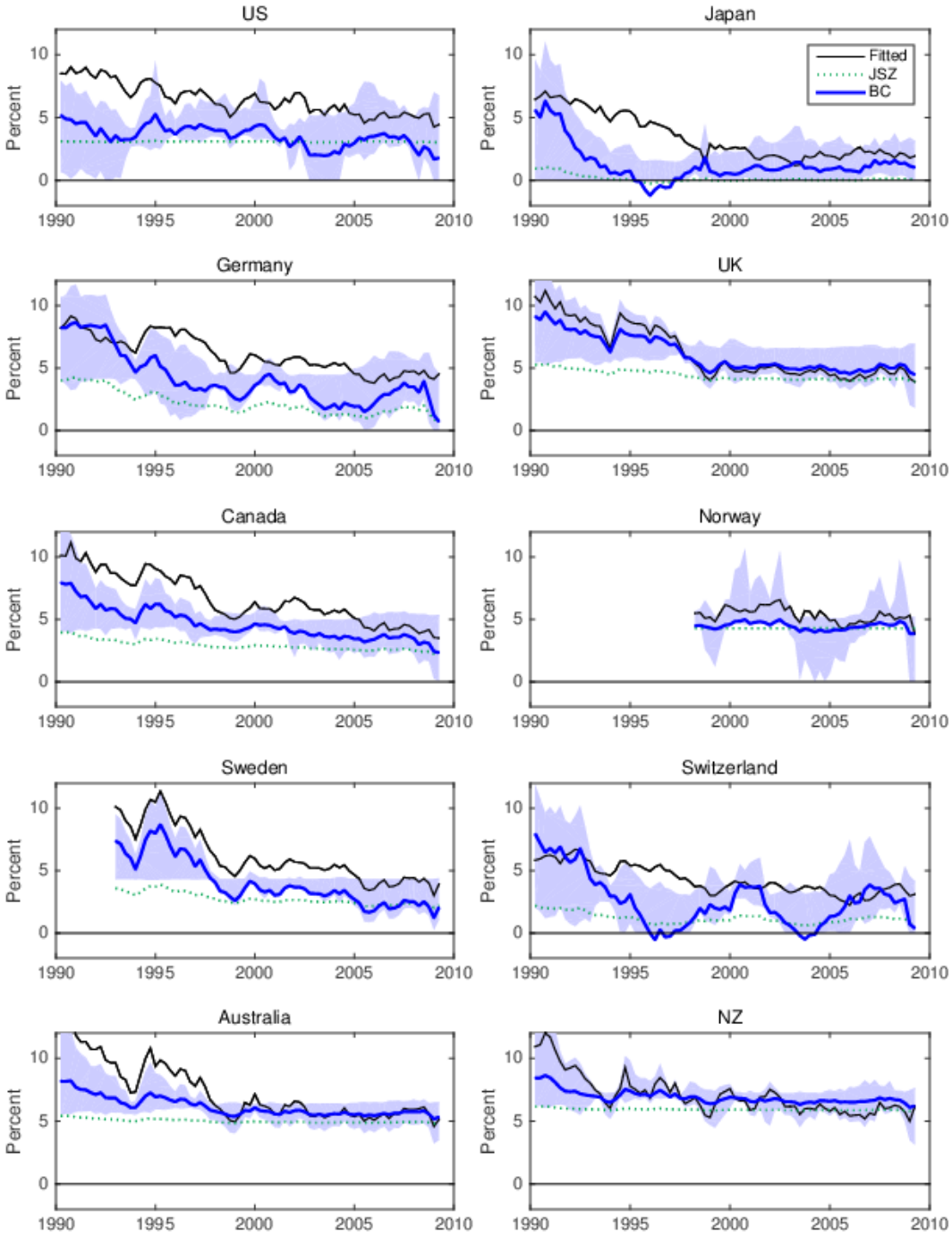
and the instantaneous volatility process of each martingale is defined as:

$$\frac{d\xi_t}{\xi_t} = -\lambda_t dW_t, \quad \frac{d\zeta_t}{\zeta_t} = -\lambda_t^{\mathbb{L}}dW_t^{\mathbb{Q}}, \quad \text{and} \quad \frac{dS_t^P}{S_t^P} = -\omega_t dW_t,$$

where W_t and $W_t^{\mathbb{Q}}$ are Brownian motions under $\mathbb{P}$ and under $\mathbb{Q}$, respectively. Then, by the Itô product rule, we can obtain

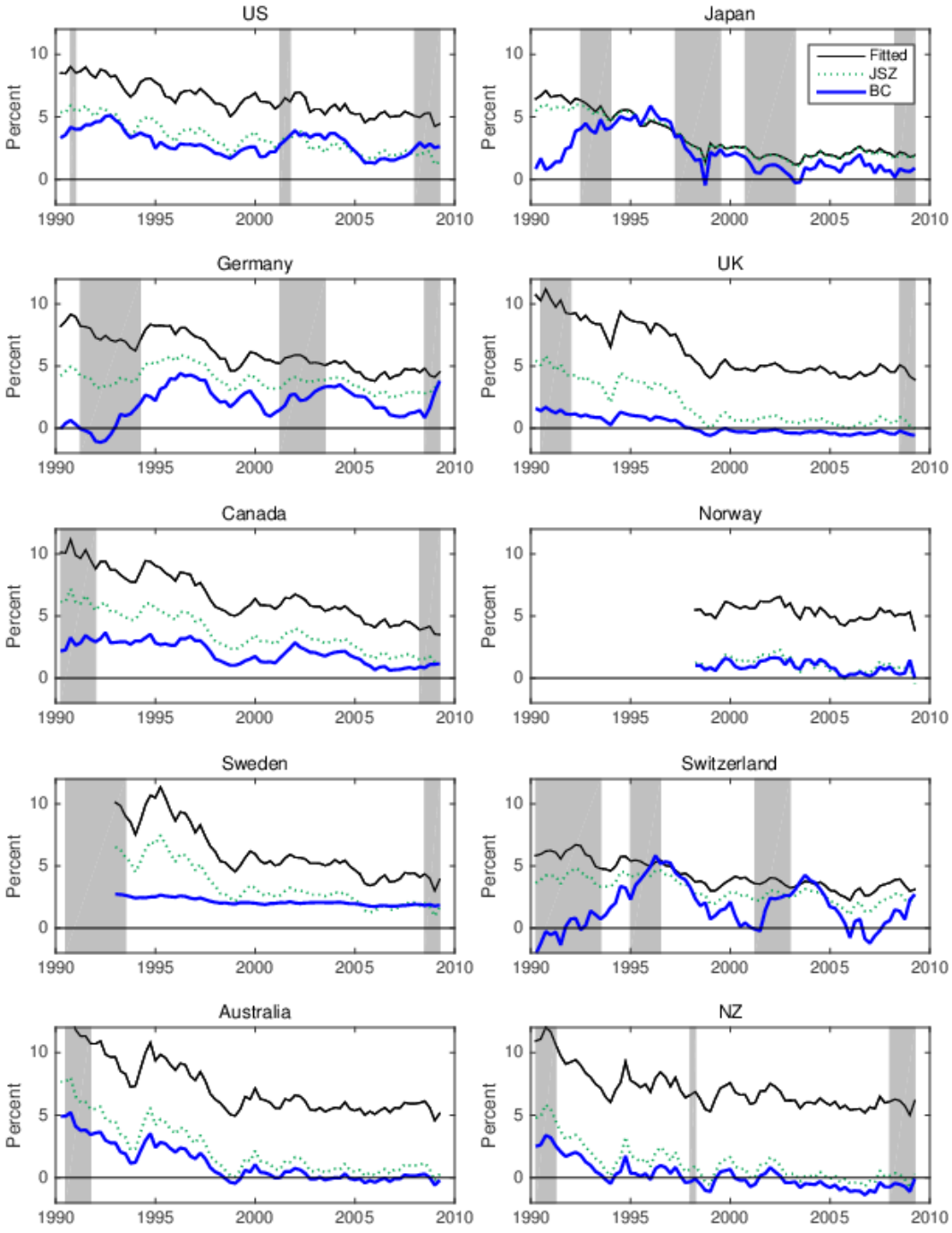
$$\begin{aligned}
-\omega_t dW_t &= -\lambda_t dW_t - \lambda_t^{\mathbb{L}} dW_t^{\mathbb{Q}} + \lambda_t dW_t \cdot \lambda_t^{\mathbb{L}} dW_t^{\mathbb{Q}} \\
&= -\lambda_t dW_t - \lambda_t^{\mathbb{L}}(dW_t + \lambda_t dt) + \lambda_t \lambda_t^{\mathbb{L}} dW_t(dW_t + \lambda_t dt) \\
&= -(\lambda_t + \lambda_t^{\mathbb{L}}) dW_t.
\end{aligned}$$

Figure A.4.1: Decomposition of Forward Rates – Fitted Rates/Risk-neutral Rates



Note: This figure plots the five- to ten-year fitted forward rates and risk-neutral rates (short-term interest rate expectations under the physical probability measure) that are estimated by *JSZ* two-step procedures and *BC* two-step procedures across 10 countries. Without loss of generality, actual forward rates are omitted, since fitting errors are small. Shaded area show bootstrapped 90 percent confidence intervals for *BC risk-neutral rates*.

Figure A.4.2: Decomposition of Forward Rates – Term Premia



Note: This figure plots the five- to ten-year fitted forward rates and the corresponding term premia that are estimated by *JSZ* two-step procedures and *BC* two-step procedures across 10 countries. For each country, the recession periods are indicated by shaded area. Without loss of generality, actual forward rates are omitted, since fitting errors are small.

Table A.4.1: Three-Factor GDTSM Estimation

		$\mathbb{P}$					$\mathbb{Q}$				
		$\mu_{\mathcal{P}}$	$\Phi_{\mathcal{P}}$			eig($\Phi_{\mathcal{P}}$)	r_∞	ϕ	$\Sigma_{\mathcal{P}}$		
		0.0040	0.9438	-0.1753	0.7071	0.9262	0.0643	0.9873	0.0123	0	0
Japan	JSZ	0.0015	0.0060	0.8915	0.4652	0.9262		0.8720	0.0013	0.0041	0
		0.0003	0.0068	0.0064	0.3886	0.3728		0.8131	-0.0017	-0.0002	0.0014
		0.0036	0.9867	-0.1525	0.6045	0.9844	0.0649	0.9873	0.0131	0	0
	BC	-0.0002	0.0023	0.9738	0.4591	0.9844		0.8720	0.0012	0.0043	0
		0.0005	0.0014	0.0071	0.4339	0.4262		0.8129	-0.0017	-0.0002	0.0015
		0.0011	0.9686	0.0245	0.0829	0.9715	0.0698	0.9737	0.0147	0	0
Germany	JSZ	0.0028	-0.0058	0.9101	0.4963	0.9265		0.9382	0.0008	0.0056	0
		-0.0001	0.0067	0.0133	0.5725	0.5532		0.5850	-0.0015	-0.0012	0.003
		-0.0002	0.9977	0.0120	0.0293	0.9997	0.0701	0.9737	0.0152	0	0
	BC	0.0007	-0.0030	0.9493	0.5021	0.9810		0.9382	0.0008	0.0056	0
		-0.0003	0.0039	0.0241	0.6220	0.5883		0.5850	-0.0015	-0.0012	0.0031
		0.0067	0.9304	-0.1922	-0.6330	0.9528	0.0317	0.9929	0.0183	0	0
United Kingdom	JSZ	-0.0035	0.0030	0.8889	-0.6902	0.8166		0.8793	-0.0002	0.0071	0
		-0.0017	0.0115	0.0355	0.7307	0.8166		0.5442	-0.0025	0.0007	0.0027
		0.0033	0.9700	-0.1790	-0.6423	0.9959	0.0334	0.9929	0.0193	0	0
	BC	-0.0001	-0.0051	0.9320	-0.6662	0.8492		0.8795	0.0000	0.0072	0
		-0.0009	0.0065	0.0246	0.7725	0.8492		0.5440	-0.0025	0.0007	0.0028
		0.0048	0.9258	-0.1283	0.2709	0.9546	0.1691	0.9973	0.0210	0	0
Canada	JSZ	-0.0004	-0.0171	0.8401	-0.7038	0.7626		0.8814	0.0018	0.0070	0
		0.0006	0.0087	0.0213	0.4809	0.5296		0.6154	-0.0023	0.0011	0.0028
		0.0016	0.9745	-0.0805	0.2389	0.9883	0.1726	0.9973	0.0220	0	0
	BC	0.0007	-0.0127	0.8895	-0.6580	0.8517		0.8814	0.0019	0.0071	0
		0.0011	0.0041	0.0127	0.5298	0.5538		0.6154	-0.0023	0.0011	0.0029
		0.0157	0.9397	-0.5647	-1.5244	0.7668	2.3751	0.9999	0.0211	0	0
Norway	JSZ	-0.0301	0.0279	0.5350	-0.0808	0.7668		0.8374	-0.0026	0.0075	0
		0.0093	0.0183	0.0326	0.6030	0.5604		0.6992	-0.0008	-0.0007	0.0031
		0.0032	1.0061	-0.5896	-1.4684	0.9157	2.5269	0.9999	0.0214	0	0
	BC	-0.0205	0.0130	0.6305	-0.1410	0.7031		0.8374	-0.0027	0.0075	0
		0.0074	0.0123	0.0166	0.6826	0.7031		0.6993	-0.0009	-0.0007	0.0031
		-0.0015	0.9485	-0.4044	-0.5490	0.9504	2.7754	0.9999	0.0213	0	0
Sweden	JSZ	-0.0049	0.0072	0.7805	-0.6217	0.7858		0.8122	-0.0013	0.0065	0
		0.0007	0.0066	0.0654	0.7320	0.7858		0.8171	-0.0025	-0.0001	0.0023
		-0.0045	0.9874	-0.4074	-0.6007	0.9982	4.1959	0.9999	0.0206	0	0
	BC	-0.0014	-0.0023	0.8333	-0.6098	0.8186		0.8089	-0.0009	0.0065	0
		0.0011	0.0031	0.0533	0.7719	0.8186		0.8201	-0.0022	0.0000	0.0024
		0.0082	0.9508	-0.1240	0.1630	0.9262	0.0775	0.9892	0.0144	0	0
Switzerland	JSZ	0.0037	0.0052	0.8454	-0.3666	0.9262		0.8467	0.0007	0.0051	0
		0.0001	-0.0033	-0.0415	0.6277	0.5721		0.8467	0.0004	0.0013	0.0023
		0.0050	0.9963	-0.0978	0.1541	0.9917	0.0780	0.9892	0.0151	0	0
	BC	0.0005	0.0027	0.9322	-0.3702	0.9917		0.8467	0.0006	0.0052	0
		0.0003	-0.0008	-0.0416	0.7045	0.6500		0.8467	0.0004	0.0012	0.0024
		0.0190	0.8932	-0.1974	-0.1236	0.9252	0.0117	1.0000	0.0176	0	0
Australia	JSZ	0.0018	-0.0142	0.8290	-0.6296	0.6452		0.8526	-0.0002	0.0064	0
		-0.0002	0.0074	0.0387	0.4903	0.6452		0.7221	-0.0028	-0.0007	0.0021
		0.0097	0.9511	-0.1797	-0.0730	0.9804	0.0159	1.0000	0.0182	0	0
	BC	0.0046	-0.0196	0.8880	-0.6258	0.8088		0.8532	-0.0002	0.0064	0
		0.0014	0.0004	0.0253	0.5076	0.5575		0.7237	-0.0028	-0.0004	0.0024
		0.0254	0.8962	-0.0824	-0.8029	0.8911	0.7837	0.9997	0.0180	0	0
New Zealand	JSZ	-0.0019	0.0060	0.8107	-0.8676	0.8058		0.8359	0.0006	0.0087	0
		0.0023	0.0065	-0.0004	0.3098	0.3198		0.5645	-0.0003	0.0004	0.0035
		0.0127	0.9580	-0.0938	-0.7610	0.9587	1.1973	0.9998	0.0211	0	0
	BC	0.0014	0.0005	0.8711	-0.8661	0.8819		0.8360	0.0000	0.0082	0
		0.0023	0.0038	-0.0104	0.3639	0.3524		0.5639	-0.0007	0.0005	0.0035

Note: $\mu_{\mathcal{P}}$, $r_\infty^{\mathbb{Q}}$, and $\Sigma_{\mathcal{P}}$ are reported on an annual basis (by multiplying 4). $\Phi_{\mathcal{P}}$ is $(I_3 + K_{\mathcal{P}})$, where $K_{\mathcal{P}}$ is the mean-reversion coefficient matrix in (4.2.21). $\phi^{\mathbb{Q}}$ here is reported by one plus the ordered eigenvalues of the mean-reversion coefficient matrix; that is $eig(I_3 + K^{\mathbb{Q}})$ in (4.2.18).

Table A.4.2: Accuracy of the GL's Markov Approximation method

		US			Japan			Germany			Switzerland		
$m = 9$													
	a. true	0.005	-0.004	0.006	0.001	-0.001	0.001	0.002	-0.002	0.002	0.001	-0.001	-0.001
$\mu_{\mathcal{P}}^{\mathbb{Q}}$	*b. markov*	0.005	-0.004	0.006	0.002	-0.001	0.001	0.003	-0.001	0.002	0.002	-0.001	-0.001
	c. direct	0.005	-0.004	0.006	0.001	-0.001	0.001	0.002	-0.001	0.002	0.001	-0.001	-0.001
	a. true	1.010	0.202	-0.842	0.979	0.139	-0.594	0.973	0.197	-0.747	0.926	0.290	0.631
		-0.029	0.954	0.829	0.009	1.003	0.568	0.005	0.962	0.682	0.023	0.914	-0.502
		0.019	-0.021	0.365	-0.005	-0.047	0.690	-0.004	-0.019	0.561	0.002	0.006	0.843
	b. markov	1.008	0.200	-0.818	0.979	0.137	-0.569	0.972	0.195	-0.736	0.925	0.289	0.626
$\Phi_{\mathcal{P}}^{\mathbb{Q}}$		-0.028	0.949	0.813	0.009	0.995	0.548	0.005	0.960	0.679	0.023	0.913	-0.502
		0.019	-0.021	0.364	-0.005	-0.047	0.687	-0.004	-0.018	0.560	0.002	0.006	0.842
	c. direct	1.010	0.202	-0.843	0.979	0.139	-0.595	0.973	0.197	-0.747	0.926	0.290	0.631
		-0.029	0.954	0.828	0.009	1.002	0.567	0.005	0.962	0.682	0.023	0.914	-0.503
		0.019	-0.021	0.365	-0.005	-0.047	0.690	-0.004	-0.019	0.561	0.002	0.006	0.842
	a. true	0.407	0.063	-0.033	0.150	0.016	-0.021	0.215	0.011	-0.022	0.209	0.010	0.006
		0.063	0.041	-0.007	0.016	0.019	-0.003	0.011	0.032	-0.008	0.010	0.026	0.007
		-0.033	-0.007	0.005	-0.021	-0.003	0.005	-0.022	-0.008	0.013	0.006	0.007	0.007
	b. markov	0.501	0.077	-0.041	0.251	0.027	-0.035	0.334	0.018	-0.035	0.309	0.015	0.009
$\Sigma_{\mathcal{P}} \times 10^3$		0.077	0.053	-0.009	0.027	0.024	-0.005	0.018	0.034	-0.009	0.015	0.030	0.008
		-0.041	-0.009	0.009	-0.035	-0.005	0.010	-0.035	-0.009	0.015	0.009	0.008	0.008
	c. direct	0.407	0.063	-0.033	0.150	0.016	-0.021	0.215	0.011	-0.022	0.209	0.010	0.006
		0.063	0.041	-0.007	0.016	0.019	-0.003	0.011	0.032	-0.008	0.010	0.026	0.007
		-0.033	-0.007	0.005	-0.021	-0.003	0.005	-0.022	-0.008	0.013	0.006	0.007	0.007
$m = 21$													
	a. true	0.005	-0.004	0.006	0.001	-0.001	0.001	0.002	-0.002	0.002	0.001	-0.001	-0.001
$\mu_{\mathcal{P}}^{\mathbb{Q}}$	*b. markov*	0.005	-0.004	0.006	0.001	-0.001	0.001	0.002	-0.001	0.002	0.001	-0.001	-0.001
	c. direct	0.005	-0.004	0.006	0.001	-0.001	0.001	0.002	-0.001	0.002	0.001	-0.001	-0.001
	a. true	1.010	0.202	-0.842	0.979	0.139	-0.594	0.973	0.197	-0.747	0.926	0.290	0.631
		-0.029	0.954	0.829	0.009	1.003	0.568	0.005	0.962	0.682	0.023	0.914	-0.502
		0.019	-0.021	0.365	-0.005	-0.047	0.690	-0.004	-0.019	0.561	0.002	0.006	0.843
	b. markov	1.010	0.202	-0.843	0.979	0.139	-0.595	0.973	0.197	-0.747	0.926	0.291	0.629
$\Phi_{\mathcal{P}}^{\mathbb{Q}}$		-0.029	0.954	0.828	0.009	1.003	0.567	0.005	0.962	0.681	0.023	0.914	-0.503
		0.019	-0.021	0.365	-0.005	-0.047	0.690	-0.004	-0.019	0.561	0.002	0.006	0.842
	c. direct	1.010	0.202	-0.843	0.979	0.139	-0.595	0.973	0.197	-0.747	0.926	0.290	0.631
		-0.029	0.954	0.828	0.009	1.002	0.567	0.005	0.962	0.682	0.023	0.914	-0.503
		0.019	-0.021	0.365	-0.005	-0.047	0.690	-0.004	-0.019	0.561	0.002	0.006	0.842
	a. true	0.407	0.063	-0.033	0.150	0.016	-0.021	0.215	0.011	-0.022	0.209	0.010	0.006
		0.063	0.041	-0.007	0.016	0.019	-0.003	0.011	0.032	-0.008	0.010	0.026	0.007
		-0.033	-0.007	0.005	-0.021	-0.003	0.005	-0.022	-0.008	0.013	0.006	0.007	0.007
	b. markov	0.418	0.064	-0.034	0.179	0.019	-0.025	0.240	0.013	-0.025	0.230	0.011	0.006
$\Sigma_{\mathcal{P}} \times 10^3$		0.064	0.041	-0.007	0.019	0.019	-0.004	0.013	0.032	-0.008	0.011	0.026	0.007
		-0.034	-0.007	0.007	-0.025	-0.004	0.007	-0.025	-0.008	0.013	0.006	0.007	0.007
	c. direct	0.407	0.063	-0.033	0.150	0.016	-0.021	0.215	0.011	-0.022	0.209	0.010	0.006
		0.063	0.041	-0.007	0.016	0.019	-0.003	0.011	0.032	-0.008	0.010	0.026	0.007
		-0.033	-0.007	0.005	-0.021	-0.003	0.005	-0.022	-0.008	0.013	0.006	0.007	0.007

Note: This table presents the results of the accuracy check for the GL method with respect to two different number of grid points along each dimension, $m = 9$ and 21. (a) "*true*" represents the coefficients in the underlying data generating process of yield factors under the $\mathbb{Q}$ measure, and (b) "*markov*" represents the induced mean estimates from the GL method. (c) "*direct*" represents the mean estimates obtained from the direct simulation of the underlying VAR(1). Consistently, $\mu_{\mathcal{P}}^{\mathbb{Q}}$, $\Phi_{\mathcal{P}}^{\mathbb{Q}}$, and $\Sigma_{\mathcal{P}}$ are reported on an annual basis (by multiplying 4).

References

Alvarez, F., and U. J. Jermann (2005), "Using asset prices to measure the persistence of the marginal utility of wealth," Econometrica, *73*, 1977–2016.

Ang, A., and M. Piazzesi (2003), "A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables," Journal of Monetary Economics, *50*, 745–787.

Audrino, F., R. Huitema, and M. Ludwig (2015), "An empirical analysis of the Ross recovery theorem," Unpublished manuscript. Available at SSRN, http://ssrn.com/abstract=2433170

Aydin, H. I., and Y. Yildirim (2015), "Extracting expectations in affine term structure models," Unpublished manuscript.

Bakshi, G., and F. Chabi-Yo (2012), "Variance bounds on the permanent and transitory components of stochastic discount factors," Journal of Financial Economics, *105*, 191–208.

Bakshi, G., F. Chabi-Yo, and X. Gao (2015), "A Recovery That We Can Trust? Deducing and Testing the Restrictions of the Recovery Theorem," Unpublished manuscript.

Bauer, M. D. (2016), "Restrictions on risk prices in dynamic term structure models," Journal of Business and Economic Statistics, forthcoming.

Bauer, M. D., G. D. Rudebusch, and J. C. Wu (2012), "Correcting estimation bias in dynamic term structure models," Journal of Business and Economic Statistics, *30*, 454–467.

Bauer, M. D., G. D. Rudebusch, and J. C. Wu (2014), "Term premia and inflation uncertainty: Empirical evidence from an international panel dataset: Comment," The American Economic Review, *104*, 323–337.

Berman, A., and R. J. Plemmons (1979), Nonnegative matrices in The Mathematical Sciences, Computer Science and Applied Mathematics. New York, NY: Academic press, INC.

Borovička, J., L. P. Hansen, and J. A. Scheinkman (2015), "Misspecified recovery," Journal of Finance, *71*, 2493–2544.

Carr, P., and Yu, J. (2012), "Risk, return, and Ross recovery," Journal of Derivatives, *20*, 38–59.

Christensen, T. (2014), "Nonparametric stochastic discount factor decomposition," Unpublished manuscript. Available at arXiv:1412.4428

Cochrane, J. H. (2009), Asset Pricing (Revised Edition). Princeton, NJ: Princeton University Press.

Cochrane, J. H., and M. Piazzesi (2005), "Bond risk premia," American Economic Review, *95*, 138–160.

Cochrane, J. H., and M. Piazzesi (2009), "Decomposing the yield curve." In AFA 2010 Atlanta Meetings Paper.

Creal, D. D., and J. C. Wu (2015), "Estimation of affine term structure models with spanned or unspanned stochastic volatility," Journal of Econometrics, *185*, 60–81.

Dai, Q., and K. J. Singleton (2000), "Specification analysis of affine term structure models," The Journal of Finance, *55*, 1943–1978.

Duffee, G. R. (2002), "Term premia and interest rate forecasts in affine models," The Journal of Finance, *57*, 405–443.

Duffee, G. R. (2011), "Forecasting with the term structure: The role of no-arbitrage restrictions," Working papers (No. 576), Johns Hopkins University, Department of Economics.

Duffee, G. R., and R. H. Stanton (2012), "Estimation of dynamic term structure models," The Quarterly Journal of Finance, *2*, 1250008.

Duffie, D. (2010), Dynamic asset pricing theory (3rd ed.). Princeton, NJ: Princeton University Press.

Duffie, D., and R. Kan (1996), "A yield-factor model of interest rates," Mathematical Finance, *6*, 379–406.

Farmer, L. E. (2014), "Markov-chain approximation and estimation of nonlinear, non-Gaussian state space models," Unpublished manuscript.

Farmer, L. E., and A. A. Toda (2015), "Discretizing stochastic processes with exact conditional moments," Unpublished manuscript. Available at SSRN, http://ssrn.com/abstract=2585859

Floden, M. (2008), "A note on the accuracy of Markov-chain approximations to highly persistent AR (1) processes," Economics Letters, *99*, 516–520.

Galindev, R., and D. Lkhagvasuren (2010), "Discretization of highly persistent correlated AR (1) shocks," Journal of Economic Dynamics and Control, *34*, 1260–1276.

Golub, G. H., and C. F. Van Loan (2013), Matrix computations (4th ed.). Baltimore, MD: JHU Press.

Gospodinov, N., and D. Lkhagvasuren (2014), "A moment-matching method for approximating vector autoregressive processes by finite-state Markov chains," Journal of Applied Econometrics, *29*, 843–859.

Gürkaynak, R. S., and J. H. Wright (2012), "Macroeconomics and the term structure," Journal of Economic Literature, *50*, 331–367.

Hamilton, J. D., and J. C. Wu (2012), "Identification and estimation of Gaussian affine term structure models," Journal of Econometrics, *168*, 315–331.

Hansen, L. P., and J. A. Scheinkman (2009), "Long-term risk: An operator approach," Econometrica, *77*, 177–234.

Hansen, L. P., (2012), "Dynamic valuation decomposition within stochastic economies," Econometrica, *80*, 911–967.

Jiang, J. (2010), Large sample techniques for statistics, Springer Texts in Statistics. New York: Springer.

Joslin, S., M. Priebsch, and K. J. Singleton (2014), "Risk premiums in dynamic term structure models with unspanned macro risks," The Journal of Finance, *69*, 1197–1233.

Joslin, S., K. J. Singleton, and H. Zhu (2011), "A new perspective on Gaussian dynamic term structure models," Review of Financial Studies, *24*, 926–970.

Kim, D. H., and A. Orphanides (2012), "Term structure estimation with survey data on interest rate forecasts," Journal of Financial and Quantitative Analysis, *47*, 241–272.

Kopecky, K. A., and R. M. Suen (2010), "Finite state Markov-chain approximations to highly persistent processes," Review of Economic Dynamics, *13*, 701–714.

Litterman, R., and J. Scheinkman (1991), "Common factors affecting bond returns," The Journal of Fixed Income, *1*, 54–61.

Martin, I., and S. Ross (2013), "The long bond," Unpublished manuscript, London School of Economics and MIT.

Meyer, C. D. (2000), Matrix analysis and applied linear algebra (Vol. 2). Philadelphia, PA: Siam.

Qin, L., and V. Linetsky (2016), "Positive eigenfunctions of Markovian pricing operators: Hansen-Scheinkman factorization, Ross recovery, and Long-term pricing," Operations Research, *64*, 99–117.

Qin, L., and V. Linetsky (2017), "Long-term risk: A martingale approach," Econometrica, *85*, 299–312.

Qin, L., V. Linetsky, and Y. Nie (2016), "Long forward probabilities, Recovery and the term structure of bond risk premiums," Unpublished manuscript. Available at SSRN, http://ssrn.com/abstract=2721366

Piazzesi, M. (2010), "Affine term structure models." In Aït-Sahalia, Y., and L. P. Hansen (Eds.), Handbook of Financial Econometrics. New York: Elsevier, pp. 691–766.

Ross, S. (2015), "The recovery theorem," The Journal of Finance, *70*, 615–648.

Rouwenhorst, G. (1995), "Asset pricing implications of equilibrium business cycle models," In Cooley, T (Eds.), Frontiers of Business Cycle Research. Princeton, NJ: Princeton University Press, pp. 294–330.

Shreve, S. E. (2004), Stochastic calculus for finance II: Continuous-time models (Vol. 11), Springer Science and Business Media. New York, NY: Springer.

Spears, T. (2013), "On estimating the risk-neutral and real-world probability measures," Doctoral dissertation, Oxford University.

Tauchen, G. (1986a), "Finite state markov-chain approximations to univariate and vector autoregressions," Economics Letters, *20*, 177–181.

Tauchen, G. (1986b), "Statistical properties of generalized method-of-moments estimators of structural parameters obtained from financial market data," Journal of Business and Economic Statistics, *4*, 397–416.

Terry, S. J., and E. S. Knotek (2011), "Markov-chain approximations of vector autoregressions: Application of general multivariate-normal integration techniques," Economics Letters, *110*, 4–6.

Tran, N. K., and S. Xia (2014), "Specified recovery," Unpublished manuscript.

Tsui, H. M. (2013), "Ross Recovery theorem and its extension," Doctoral dissertation, Oxford University.

Walden, J. (2017), "Recovery with unbounded diffusion processes," Review of Finance, rfw068.

Wright, J. H. (2011), "Term premia and inflation uncertainty: Empirical evidence from an international panel dataset," The American Economic Review, *101*, 1514–1534.

Vita

Name of Author: Jaewoo Oh

Place of Birth: Seoul, Republic of Korea

Date of Birth: February 29, 1976

Education

08/2012–12/2013	Master of Public Administration, Syracuse University, The Maxwell School, Syracuse, NY
03/1995–08/2000	Bachelor of Arts in Business, Yonsei University, Seoul, Republic of Korea

Experience and Employment

08/2016–05/2017	Research Specialist, Department of Economics, University of Connecticut
01/2014–05/2016	Graduate Assistant, Department of Economics, Syracuse University

Awards and Fellowships

2014-2016	University Graduate Assistantship, Syracuse University
2014-2016	Maxwell Dean Summer Fellowship, Syracuse University
2015	Travel Grant, Syracuse University
2012-2014	Korean Government Fellowship for Overseas Studies, Korean Government
2006	Minister's Prize for Outstanding Achievement, Ministry of Strategy and Finance, Korea
2001	Commissioner's Prize for Outstanding Achievement, Korea Customs Service, Korea
1995-2000	Scholarship for Honors, Yonsei University, Korea