UNIVERSITY OF CALIFORNIA, SAN DIEGO

Essays in Econometrics

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Economics

by

Qihui Chen

Committee in charge:

Professor Andres Santos, Co-Chair
Professor Yixiao Sun, Co-Chair
Professor Ian Abramson
Professor Brendan K. Beare
Professor Dimitris N. Politis

2017

The dissertation of Qihui Chen is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Co-Chair

_____

Co-Chair

University of California, San Diego

2017

DEDICATION

To my parents, older brother, and younger sister

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Table of Contents vi

List of Figures vii

List of Tables viii

List of Symbols ix

Acknowledgements x

Vita xi

Abstract of the Dissertation xii

Chapter 1 Inference on Functions under First Order Degeneracy 1
- 1.1 Introduction 3
- 1.2 Setup and Background 8
  - 1.2.1 General Setup 8
  - 1.2.2 Related Examples 9
  - 1.2.3 Concepts of Differentiability 12
  - 1.2.4 Second Order Delta Method 17
- 1.3 The Bootstrap 19
  - 1.3.1 Bootstrap Setup 20
  - 1.3.2 Failures of the Standard Bootstrap 22
  - 1.3.3 The Babu Correction 24
  - 1.3.4 A Modified Bootstrap 26
  - 1.3.5 Estimation of the Derivative 29
- 1.4 Hypothesis Testing 31
  - 1.4.1 Local Perturbations 32
  - 1.4.2 Local Size and Power 34
- 1.5 Application: Testing for Common CH Features 37
  - 1.5.1 The Setup 38
  - 1.5.2 A Modified J Test 40
  - 1.5.3 Simulation Studies 44
- 1.6 Conclusion 47
- 1.7 Acknowledgement 47
- 1.8 Appendix 50
  - 1.8.1 Proofs of Main Results 50
  - 1.8.2 Results for Examples 1.2.1 - 1.2.6 67
  - 1.8.3 Proofs for Section 1.5 78

Chapter 2 Improved Inference on the Rank of a Matrix with Applications to IV and Cointegration Models . . . 87
2.1 Introduction . . . 89
2.2 Examples and Motivation . . . 93
2.2.1 Examples . . . 93
2.2.2 Motivation . . . 99
2.3 Asymptotic Analysis . . . 102
2.3.1 Differential Properties . . . 103
2.3.2 The Asymptotic Distributions . . . 106
2.3.3 The Bootstrap . . . 109
2.4 Simulations and Applications . . . 114
2.4.1 Simulation Studies . . . 114
2.4.2 Testing for Identification in SDF Models . . . 118
2.4.3 Rank Determination . . . 123
2.5 Conclusion . . . 127
2.6 Acknowledgement . . . 127
2.7 Appendix . . . 127
2.7.1 Proofs of Main Results . . . 127
2.7.2 Results for Examples 2.2.1-2.2.7 . . . 144
2.7.3 Kleibergen and Paap (2006)'s Test . . . 151
2.7.4 Parameters in Section 2.4.1 . . . 154

Chapter 3 Robust and Optimal Estimation for Partially Linear Instrumental Variables Models with Partial Identification . . . 157
3.1 Introduction . . . 159
3.2 The Model . . . 163
3.3 Robust Estimation . . . 165
3.3.1 Strong-Norm Convergence . . . 168
3.3.2 Weak-Norm Convergence Rate . . . 171
3.3.3 Asymptotic Normality . . . 171
3.4 Optimal Estimation . . . 175
3.4.1 Optimal Penalty . . . 176
3.4.2 Two-Step Procedure . . . 177
3.5 Variance Estimation . . . 179
3.6 Simulation Studies . . . 180
3.7 Conclusion . . . 185
3.8 Acknowledgement . . . 185
3.9 Appendix . . . 185
3.9.1 Proofs of Main Results . . . 185
3.9.2 Useful Lemmas . . . 194

References . . . 206

LIST OF FIGURES

Figure 2.1: The rejection rate of the multiple testing method based on the Kleibergen and Paap (2006) test with 5% nominal level 102

Figure 2.2: Comparison between the Kleibergen and Paap (2006) test and the multiple testing method based on it with 5% nominal level 103

Figure 2.3: The rejection rate of our tests with 5% nominal level 115

Figure 2.4: Comparison between our tests and the multiple testing method based on the Kleibergen and Paap (2006) test with 5% nominal level 116

Figure 2.5: Comparison between the sequential testing procedures based on our tests and the Kleibergen and Paap (2006) test with $\delta = 0.1$ 126

Figure 2.6: Comparison between the sequential testing procedures based on our tests and the Kleibergen and Paap (2006) test with $\delta = 0.12$ 126

Figure 2.7: The distribution function of the weak limit of rk(1) when $\Pi_0 = \mathbf{0}_{2\times 2}$ 153

Figure 3.1: The distribution of $\sqrt{n}(\hat{\beta}_P - \beta_0)$ for various $P(\cdot)$ when $\lambda_n = n^{-4/5}$ and $n = 100$ 182

Figure 3.2: The distribution of $\sqrt{n}(\hat{\beta}_P - \beta_0)$ for various $P(\cdot)$ when $\lambda_n = n^{-4/5}$ and $n = 500$ 182

Figure 3.3: The distribution of $\sqrt{n}(\hat{\beta}_P - \beta_0)$ for various $P(\cdot)$ when $\lambda_n = n^{-4/5}$ and $n = 1000$ 183

Figure 3.4: The distribution of $\sqrt{n}(\hat{\beta}_P - \beta_0)$ for various $P(\cdot)$ when $\lambda_n = n^{-4/5}$ and $n = 2000$ 183

LIST OF TABLES

Table 1.1: Simulation Designs 46

Table 1.2: Rejection rates under the null: Design D1 48

Table 1.3: Rejection rates under the null: Design D3 48

Table 1.4: Rejection rates under the null: Design D4 49

Table 1.5: Rejection rates under the alternative 49

Table 2.1: Rejection rates for $r = 2$ under DGP1 119

Table 2.2: Rejection rates for $r = 3$ under DGP1 120

Table 2.3: Rejection rates for $r = 3$ under DGP2 when $T = 330$ 121

Table 2.4: p values for different tests 123

Table 3.1: The bias of $\hat{\beta}_P$ for various $P(\cdot)$, λ_n and n 184

Table 3.2: The standard deviation of $\sqrt{n}\hat{\beta}_P$ for various $P(\cdot)$, λ_n and n 184

Table 3.3: List of simplified notation 186

Table 3.4: List of sequences 186

LIST OF SYMBOLS

NORMS AND METRICS

$\|\cdot\|_\infty$:	For a function $f : T \to \mathbf{R}$, $\|f\|_\infty \equiv \sup_{t\in T} \|f(t)\|$.
$d(a, B)$:	For B a subset of a metric space (T, d), $d(a, B) \equiv \inf_{b\in B} d(a, b)$.
$d_H(A, B)$:	For sets A, B, $d_H(A, B) \equiv \max\{\sup_{a\in A} d(a, B), \sup_{b\in B} d(b, A)\}$.
$\|A\|$:	The Frobenius norm of a matrix $A \in \mathbf{M}^{m\times k}$.
$\|\cdot\|_{L^q(\mathrm{P})}$:	For a measure P and function f, $\|f\|^q_{L^q(\mathrm{P})} \equiv \int \|f\|^q d\,\mathrm{P}$.
$N(\epsilon, \mathcal{F}, \|\cdot\|^*)$:	The ϵ covering number for a class under $\|\cdot\|^*$.
$N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|^*)$:	The ϵ bracketing number for a class under $\|\cdot\|^*$.
$J_{[]}(\delta, \mathcal{F}, \|\cdot\|^*)$:	The entropy integral $\int_0^\delta \{1 + \log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|^*)\}^{1/2} d\epsilon$.

SETS AND SPACES

$\mathbb{B}, \mathbb{D}, \mathbb{E}$:	Normed or Banach spaces.
$\mathbf{M}^{m\times k}$:	The space of $m \times k$ real matrices.
$C(T)$:	The space of continuous functions on T.
$C^1(T)$:	The space of continuously differentiable functions on T.
$\ell^\infty(T)$:	The space of bounded functions on T with norm $\|\cdot\|_\infty$.
$A^\epsilon, \epsilon > 0$:	For A a subset of a metric space (T, d), $A^\epsilon \equiv \{a \in T : d(a, A) \le \epsilon\}$.
$\mathbb{S}^{m\times k}$:	A subset of $\mathbf{M}^{m\times k}$: $\mathbb{S}^{m\times k} \equiv \{U \in \mathbf{M}^{m\times k} : U^\intercal U = I_k\}$.

OTHERS

$1\{A\}$:	The indicator function of a subset A.
$a \lesssim b$:	$a \le Mb$ for some constant M that is universal in the dissertation.
$A^\intercal$:	The transpose of $A \in \mathbf{M}^{m\times k}$.
$\mathrm{tr}(A)$:	The trace of a square matrix $A \in \mathbf{M}^{k\times k}$.
$\mathrm{vec}(A)$:	The column vectorization of $A \in \mathbf{M}^{m\times k}$.
$\sigma_j(A)$:	The jth largest singular value of a matrix $A \in \mathbf{M}^{m\times k}$.
$\varphi : A \twoheadrightarrow B$:	A correspondence from a set A to another set B.

ACKNOWLEDGEMENTS

Many thanks to my dissertation committee members: Andres Santos, Yixiao Sun, Brendan Beare, Dimitris Politis, and Ian Abramson. I am also extremely grateful to my co-author, Zheng Fang, my prior classmates at Singapore Management University, Bo Chen and Ye Chen, and my master thesis advisors, Liangjun Su and Sainan Jin. Lastly, I am thankful to my parents, older brother and younger sister for their understanding and support.

Chapters 1 and 2, in part, are currently being prepared for submission for publication of the material. Chen, Qihui; Fang, Zheng. Chapter 3, in part, is currently being prepared for submission for publication of the material. Chen, Qihui. The dissertation author was the primary investigator and author of these materials.

VITA

2008 B.S. in Mathematics, Xiamen University

2008 B.A. in Economics, Xiamen University

2011 M.A. in Economics, Xiamen University

2011 M.S. in Economics, Singapore Management University

2017 Ph.D. in Economics, University of California, San Diego

PUBLICATIONS

"Testing Homogeneity in Panel Data Models with Interactive Fixed Effects," *Econometric Theory*, 29(6), 1079-1135 (with Liangjun Su)

"Improvement in Finite-Sample Properties of GMM-Based Wald Tests," *Computational Statistics*, 28(2), 735-749 (with Yu Ren)

ABSTRACT OF THE DISSERTATION

Essays in Econometrics

by

Qihui Chen

Doctor of Philosophy in Economics

University of California, San Diego, 2017

Professor Andres Santos, Co-Chair

Professor Yixiao Sun, Co-Chair

This dissertation studied two main topics: inference methods for directionally differentiable functions with first order degeneracy and estimation methods for semiparametric instrumental variables (IV) models with partial identification. Chapter 1 presents a general asymptotic framework for conducting inference on directionally differentiable functions with a zero first order derivative. Chapter 2 applies the theory to develop improved inference methods for the rank of a matrix, which is important in estimation and inference IV models. Chapter 3 studies robust and optimal estimation in a canonical semiparametric IV model – partially linear IV model – with nonparametric partial identification.

Chapter 1

Inference on Functions under First Order Degeneracy

Abstract

This chapter presents a unified second order asymptotic framework for conducting inference on parameters of the form $\phi(\theta_0)$, where θ_0 is unknown but can be estimated by $\hat{\theta}_n$, and ϕ is a known map that admits null first order derivative at θ_0. For a large number of examples in the literature, the second order Delta method reveals a nondegenerate weak limit for the plug-in estimator $\phi(\hat{\theta}_n)$. We show, however, that the "standard" bootstrap is consistent if and only if the second order derivative $\phi''_{\theta_0} = 0$ under regularity conditions, i.e., the standard bootstrap is inconsistent if $\phi''_{\theta_0} \neq 0$, and provides degenerate limits unhelpful for inference otherwise. We thus identify a source of bootstrap failures distinct from that in Fang and Santos (2015) because the problem persists even if ϕ is differentiable. We show that the correction procedure in Babu (1984) can be extended to our general setup. Alternatively, a modified bootstrap is proposed to accommodate nondifferentiable maps. Both approaches are shown to provide local size control under restrictions on $\hat{\theta}_n$ and ϕ''_{θ_0}. As an illustration, we develop a test of common conditional heteroskedastic (CH) features

that allows the existence of multiple common CH features. In fact, this chapter contains new results on the J-test in GMM settings that allow partial identification and/or degeneracy of Jacobian matrices.

1.1 Introduction

There is a large number of inference problems in economics and statistics in which the parameter of interest is of the form $\phi(\theta_0)$, where θ_0 is an unknown parameter depending on the underlying distribution of the data and ϕ is a known map. In these settings, it is common practice to employ the plug-in estimator $\phi(\hat{\theta}_n)$, where $\hat{\theta}_n$ is an estimator for θ_0, as a building block for conducting inference on $\phi(\theta_0)$. The Delta method asserts that if $r_n\{\hat{\theta}_n - \theta_0\} \xrightarrow{L} \mathbb{G}$ for some sequence $r_n \uparrow \infty$, then

$$r_n\{\phi(\hat{\theta}_n) - \phi(\theta_0)\} \xrightarrow{L} \phi'_{\theta_0}(\mathbb{G}) \ , \tag{1.1}$$

provided ϕ is at least Hadamard directionally differentiable at θ_0, where ϕ'_{θ_0} is the derivative of ϕ at θ_0 (Shapiro, 1991; Dümbgen, 1993). As powerful as the Delta method has proven to be (van der Vaart, 1998; Fang and Santos, 2015), an implicit and yet crucial assumption for the convergence (1.1) to be useful for inferential purposes is that $\phi'_{\theta_0}(\mathbb{G})$ or ϕ'_{θ_0} is nondegenerate, i.e., $\phi'_{\theta_0} \neq 0$. Unfortunately, such *first order degeneracy* arises frequently in asymptotic analysis, with applications including Wald tests or Wald type functionals (Wald, 1943; Engle, 1984), unconditional and conditional moment inequality models (Andrews and Soares, 2010; Andrews and Shi, 2013), Cramér-von Mises functionals (Darling, 1957), the study of stochastic dominance (Linton et al., 2010), and the J-test for overidentification in GMM settings (Hall and Horowitz, 1996).

In the presence of first order degeneracy, one may resort to a higher order analysis for the sake of a nondegenerate limiting distribution. Shapiro (2000) established that if ϕ is second order Hadamard directionally differentiable (see Definition 1.2.2), then

$$r_n^2\{\phi(\hat{\theta}_n) - \phi(\theta_0) - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)\} \xrightarrow{L} \phi''_{\theta_0}(\mathbb{G}) \ , \tag{1.2}$$

where ϕ''_{θ_0} denotes the second order derivative of ϕ at θ_0. Thus, when first order degeneracy

occurs, (1.2) suggests that we may base our asymptotic analysis on

$$r_n^2\{\phi(\hat{\theta}_n) - \phi(\theta_0)\} \xrightarrow{L} \phi''_{\theta_0}(\mathbb{G}) \ . \tag{1.3}$$

On the other hand, the common feature that the aforementioned examples share is that ϕ is second order Hadamard (directionally) differentiable and that the resulting derivative is nondegenerate. Usefulness of the limiting distribution in (1.3), however, relies on our ability to consistently estimate it. In this regard, Efron (1979)'s bootstrap seems to be a potential option. Specifically, if $\hat{\theta}_n^*$ is a bootstrap analog of $\hat{\theta}_n$ that works for estimating the law of $\mathbb{G}$, then in view of (1.3) one may hope that

$$r_n^2\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\} \tag{1.4}$$

can be employed as an estimator for the law of $\phi''_{\theta_0}(\mathbb{G})$, at least when ϕ is smooth. Unfortunately, there are simple examples where the law of (1.4) conditional on the data, referred to as the standard bootstrap, fails to provide consistent estimates (Babu, 1984).

As the first contribution of this paper, we show that the standard bootstrap (1.4) is consistent if and only if $\phi''_{\theta_0} = 0$, whenever $\mathbb{G}$ is centered Gaussian. Thus, the standard bootstrap is necessarily inconsistent when ϕ''_{θ_0} is nondegenerate, while when ϕ''_{θ_0} is degenerate, the resulting asymptotic distribution is degenerate and hence not useful for inference. We thus conclude that the failure of the standard bootstrap is an inherent implication of first order degeneracy. It is worth noting that the failure of the standard bootstrap persists even when ϕ is differentiable. Hence, we identify a source of bootstrap inconsistency completely different from that in Fang and Santos (2015) — i.e., nondifferentiability of the map ϕ.

Heuristically, the reason why the standard bootstrap fails is that even though $r_n^2\phi'_{\theta_0}(\hat{\theta}_n - \theta_0) = 0$ in the "real world", its bootstrap counterpart is nondegenerate, i.e., $r_n^2\phi'_{\hat{\theta}_n}(\hat{\theta}_n^* - \hat{\theta}_n) = O_p(1)$, echoing Efron (1979)'s point that the bootstrap provides approximate frequency statements rather than approximate likelihood statements. This observation was picked up by Babu (1984) who provided a consistent resampling procedure by including

the first order correction term:

$$r_n^2\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n) - \phi'_{\hat{\theta}_n}(\hat{\theta}_n^* - \hat{\theta}_n)\} \ . \tag{1.5}$$

As the second contribution, we generalize the above modified bootstrap (1.5), referred to as the Babu correction, to settings that accommodate infinite dimensional models and a wide range of bootstrap schemes for $\hat{\theta}_n^*$. However, we stress that the Babu correction is inappropriate when ϕ is only Hadamard directionally differentiable.

As the third contribution of the paper, we follow Fang and Santos (2015) and provide a modified bootstrap which is consistent regardless of the presence of first order degeneracy and nondifferentiability of ϕ. The insight we exploit is that the weak limit $\phi''_{\theta_0}(\mathbb{G})$ in (1.3) is a composition of the limit $\mathbb{G}$ and the second order derivative ϕ''_{θ_0}. Therefore, we may estimate the law of $\phi''_{\theta_0}(\mathbb{G})$ by composing a suitable estimator $\hat{\phi}''_n$ for ϕ''_{θ_0} with a bootstrap approximation $r_n\{\hat{\theta}_n^* - \hat{\theta}_n\}$ for $\mathbb{G}$. Since the conditions on $\hat{\phi}''_n$ proposed by Fang and Santos (2015) in order for this approach to work are either demanding or hard to check in our setup, we provide a high level condition that is easy to verify. We further demonstrate that numerical differentiation provides a desirable estimator $\hat{\phi}''_n$ in general; alternatively, we show how to estimate ϕ''_{θ_0} by exploiting its structure in particular examples. Interestingly, we note that the above procedure is a combination of bootstrap and analytic asymptotic approximations, while initially the former was intended as a substitute for and improvement upon the latter (Horowitz, 2001).

It is often the case that a hypothesis on θ_0 can be formulated as: for some ϕ,

$$\mathrm{H}_0 : \phi(\theta_0) = 0 \qquad \mathrm{H}_1 : \phi(\theta_0) > 0 \ . \tag{1.6}$$

In turn, the above asymptotic framework suggests that we employ $r_n^2\phi(\hat{\theta}_n)$ as the test statistic in the presence of first order degeneracy. Pointwise size control then follows immediately by employing critical values based on our resampling procedures. As argued by Imbens and Manski (2004) and Andrews and Guggenberger (2009a), however, pointwise asymp-

totic approximations may be unreliable when $\phi(\hat{\theta}_n)$ is irregular, i.e., when the asymptotic distribution of $\phi(\hat{\theta}_n)$ is sensitive to local perturbations of the distribution of the data. In our setup, such irregularity is inherent to first order degeneracy. We show that our test ensures local size control provided $\hat{\theta}_n$ is regular and $\tau \circ \phi''_{\theta_0}$ is subadditive for some strictly increasing map τ. We note that unlike Fang and Santos (2015), however, ϕ''_{θ_0} itself often fails to be subadditive.

Our framework includes many existing results as special cases. To further demonstrate the applicability of our framework, we develop a test of common conditional heteroskedastic (CH) features studied by Dovonon and Renault (2013) but under weaker assumptions that allow partial identification, i.e., allow the existence of more than one common CH feature. This is important because it is unknown *a priori* how many common features there are and in the context of asset pricing the number is presumably large (Engle et al., 1990). Monte Carlo simulations indicate that our tests substantially alleviate size distortion and have good power performance. Our approach may also be used to develop tests for other common features (Engle and Kozicki, 1993). In fact, our paper contains new results on the J-test in GMM settings that allow partial identification and/or degeneracy of Jacobian matrices.

There have been extensive studies on the validity of bootstrap schemes (Hall, 1992; van der Vaart and Wellner, 1996a; Horowitz, 2001). It was realized soon after Efron (1979) that the bootstrap is not always successful (Bickel and Freedman, 1981); see also Andrews (2000) for a summary. Babu (1984) provided a simple example of bootstrap failure due to first order degeneracy, and established the validity of the Babu correction for the special case studied there. Shao (1994) showed that m out of n resampling can well serve as an alternative remedy, while Bertail et al. (1999) provided a two step modified subsampling procedure which involves estimation of the convergence rate in the first stage. Both methods entail the choice of tuning parameters while our proposal often works without such nuisances when ϕ is differentiable. Datta (1995) revisited Babu's example and offered a bias correction procedure that depends on a first stage shrinkage type estimator. Somewhat sim-

ilar methods were later proposed in Andrews (2000) and Giurcanu (2012). Interestingly, bootstrap inconsistency for some U and V statistics can also be attributed to first order degeneracy (Bickel et al., 1997).

Bootstrap inconsistency arising from nondifferentiability was studied in Dümbgen (1993), Andrews (2000), and recently in Fang and Santos (2015) who formally established that differentiability of ϕ is a necessary as well as sufficient condition for the standard bootstrap to work under mild regularities. Our work complements theirs by identifying a different source of bootstrap failure. In the literature, the two sources are often mixed together, for example, in Romano and Shaikh (2010), Andrews and Soares (2010), Linton et al. (2010), and Andrews and Shi (2013). The second order analysis of resampling schemes such as jackknife and bootstrap has been employed in the statistics literature, though the focus has been on bias and variance estimation where typically the stronger concept of second order Fréchet differentiability is imposed (Efron, 1979; Beran, 1984; Rao and Wu, 1985; Shao and Wu, 1989; Shao, 1991). The numerical differentiation approach of estimating derivatives was implicit in Dümbgen (1993)'s rescaled bootstrap, recently employed by Song (2014), and comprehensively studied by Hong and Li (2015) including discussions on second order asymptotics. Our work complements Hong and Li (2015) by providing a more general condition that may be used to verify "consistency" of derivative estimators (not necessarily constructed via numerical differentiation).

The remainder of this paper is structured as follows. Section 1.2 formalizes the general setup, shows the wide applicability of our framework by introducing related examples, and establishes the asymptotic framework by presenting a mild extension of the second order Delta method. Section 1.3 characterizes the inherent difficulties caused by first order degeneracy, extends the Babu correction to our general setup, and offers a flexible modified bootstrap procedure. Section 1.4 demonstrates that our procedure is robust to local perturbations of the distribution of the data under regularity conditions. Section 1.5 develops a test of common CH features that accommodates partial identification, while Section 1.6 concludes. Proofs are collected in the appendices.

1.2 Setup and Background

In this section, we formalize the general setup, introduce related examples, and review notions of differentiability based on which we present the second order Delta method.

1.2.1 General Setup

The treatment in the paper is general in the sense that we allow both the parameter θ_0 and the map ϕ to take values in infinite dimensional spaces, though attention is confined to real-valued ϕ when studying tests. In particular, we assume $\theta_0 \in \mathbb{D}_\phi \subset \mathbb{D}$ and $\phi : \mathbb{D}_\phi \to \mathbb{E}$, where $\mathbb{D}$ and $\mathbb{E}$ are normed spaces with norms $\|\cdot\|_\mathbb{D}$ and $\|\cdot\|_\mathbb{E}$ respectively. Moreover, the data generating process is general as well in that the model can be parametric, semiparametric and nonparametric and that the data $\{X_i\}_{i=1}^n$ need not be i.i.d.. However, we do impose i.i.d. assumption in our local analysis, but only for simplicity. The results there can presumably be extended to general asymptotically normal experiments (van der Vaart and Wellner, 1990).

The common probability space on which all (random) maps are defined is the canonical one. For example, in the simplest i.i.d. setup, we think of the data $\{X_i\}_{i=1}^n$ as the coordinate projections on the first n coordinates in the product probability space $(\prod_{i=1}^\infty \mathscr{X}, \bigotimes_{i=1}^\infty \mathcal{A}, \prod_{i=1}^\infty P)$ where $(\mathscr{X}, \mathcal{A})$ is the sample space each X_i lives in and P is the common Borel probability measure that governs each X_i. In the presence of bootstrap weights, we further think of the product space as the "first ∞" coordinates of the even "larger" product space $\big((\prod_{i=1}^\infty \mathscr{X}) \times \mathscr{W}, (\bigotimes_{i=1}^\infty \mathcal{A}) \otimes \mathcal{W}, (\prod_{i=1}^\infty P) \times Q\big)$, where $(\mathscr{W}, \mathcal{W}, Q)$ governs the infinite sequence of bootstrap weights.

Given the generality of our setup, weak convergence throughout the paper is meant in the Hoffmann-Jørgensen sense (van der Vaart and Wellner, 1996a). Expectations and probabilities should therefore be interpreted as outer expectations and outer probabilities respectively defined relative to the canonical probability space, though we obviate the

distinction in the notation. The notation is made explicit in the appendices whenever differentiating between inner and outer expectations is necessary.

1.2.2 Related Examples

To fix ideas, we now turn to related examples that serve to illustrate the wide applicability of our framework. The first example is taken from Babu (1984), which provides an easy illustration of bootstrap inconsistency in the presence of first order degeneracy even if the transformation ϕ is smooth.

Example 1.2.1 (Wald Functional: Squared Mean)**.** Let $X \in \mathbf{R}$ be a random variable, and suppose that we are interested in conducting inference on

$$\phi(\theta_0) = (E[X])^2 \ . \tag{1.7}$$

Here, $\theta_0 = E[X]$, $\mathbb{D} = \mathbb{E} = \mathbf{R}$, and $\phi : \mathbf{R} \to \mathbf{R}$ is defined by $\phi(\theta) = \theta^2$. In fact, ϕ is a special case of the more general quadratic functionals of the form $\|W\theta\|^2$ for $\theta \in \mathbf{R}^k$ and W a $k \times k$ weighting matrix. This seemingly toy example also arises in VAR models for inference on impulse responses (Benkwitz et al., 2000) and in some nonseparable models with structural measurement errors (Hoderlein and Winter, 2010). ■

The second example is a special case of the unconditional moment inequality models studied in Chernozhukov et al. (2007), Romano and Shaikh (2008, 2010), Andrews and Guggenberger (2009b), and Andrews and Soares (2010).

Example 1.2.2 (Unconditional Moment Inequalities)**.** Let $X \in \mathbf{R}$ be a scalar random variable and suppose we want to test the moment inequality $E[X] \leq 0$. The modified method of moments approach is based on estimating the functional

$$\phi(\theta_0) = (\max\{\theta_0, 0\})^2 \ , \tag{1.8}$$

where $\theta_0 = E[X]$, $\mathbb{D} = \mathbb{E} = \mathbf{R}$, and $\phi : \mathbf{R} \to \mathbf{R}$ is defined by $\phi(\theta) = (\max\{\theta, 0\})^2$. The functional ϕ can be easily adapted to handle general moment inequality models. ■

The third example concerns the classical Cramér-von Mises functional employed to test goodness of fit (Darling, 1957; van der Vaart, 1998), which has been widely adopted in economics and statistics.

Example 1.2.3 (Cramér-von Mises Functional)**.** Suppose that we are interested in testing if the distribution function of a random vector $X \in \mathbf{R}^{d_x}$ is a given function F_0. The Cramér-von Mises approach considers the functional

$$\phi(\theta_0) = \int (F - F_0)^2 \, dF_0 \ .$$

Here, $\theta_0 = F$, $\mathbb{D} = \ell^\infty(\mathbf{R}^{d_x})$, $\mathbb{E} = \mathbf{R}$, and $\phi : \ell^\infty(\mathbf{R}^{d_x}) \to \mathbf{R}$ is defined to be $\phi(\theta) = \int (\theta - F_0)^2 \, dF_0$. More generally, it is possible to test if F belongs to a parametric family $\{F_\gamma : \gamma \in \Gamma\}$ by studying $\phi(\theta_0) = \inf_{\gamma \in \Gamma} \int (\theta_0 - F_\gamma)^2 \, dF_\gamma$. ■

The fourth example, closely related to but significantly different from Example 1.2.3, is based on Linton et al. (2010) for testing stochastic dominance.

Example 1.2.4 (Stochastic Dominance)**.** Let $X = (X^{(1)}, X^{(2)})^\intercal \in \mathbf{R}^2$ be continuously distributed, and define the marginal cdfs $F^{(j)}(u) \equiv P(X^{(j)} \leq u)$ for $j \in \{1, 2\}$. For a positive integrable weighting function $w : \mathbf{R} \to \mathbf{R}^+ \equiv \{x \in \mathbf{R} : x \geq 0\}$, Linton et al. (2010) estimate

$$\phi(\theta_0) = \int_{\mathbf{R}} \max\{F^{(1)}(u) - F^{(2)}(u), 0\}^2 w(u) du \ , \tag{1.9}$$

to construct a test of whether $X^{(1)}$ first order stochastically dominates $X^{(2)}$. In this example, we set $\theta_0 = (F^{(1)}, F^{(2)})$, $\mathbb{D} = \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$, $\mathbb{E} = \mathbf{R}$ and $\phi(\theta) = \int \max\{\theta^{(1)}(u) - \theta^{(2)}(u), 0\}^2 w(u) du$ for any $\theta \equiv (\theta^{(1)}, \theta^{(2)}) \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$. We note that the Cramér-von Mises type functionals in Andrews and Shi (2013, 2014) shares the common structure of the functional ϕ in (1.9) and hence can be taken care of by our framework as well. ■

The fifth example is a special case of the Kolmogorov-Smirnov type functionals for inference on conditional moment inequalities studied by Andrews and Shi (2013).

Example 1.2.5 (Conditional Moment Inequalities)**.** Let $Z \in \mathbf{R}^2$ and $W \in \mathbf{R}^{d_w}$ be random vectors satisfying $E[Z^{(1)}|W] \leq 0$ and $E[Z^{(2)}|W] = 0$. For a suitably chosen class of nonnegative functions $\mathcal{F}$ on $\mathbf{R}^{d_w}$, the above conditional moment inequality is equivalent to $E[Z^{(1)}f(W)] \leq 0$ and $E[Z^{(2)}f(W)] = 0$ for all $f \in \mathcal{F}$. Andrews and Shi (2013) propose testing the above restriction by estimating the functional

$$\phi(\theta_0) = \sup_{f \in \mathcal{F}} \{[\max(E[Z^{(1)}f(W)], 0)]^2 + (E[Z^{(2)}f(W)])^2\} \ . \tag{1.10}$$

Here, $\theta_0 \in \ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$ satisfies $\theta_0(f) = E[Zf(W)]$ for all $f \in \mathcal{F}$, $\mathbb{D} = \ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$, $\mathbb{E} = \mathbf{R}$, and $\phi : \mathbb{D} \to \mathbb{E}$ is given by $\phi(\theta) = \sup_{f \in \mathcal{F}} \{[\max(\theta^{(1)}(f), 0)]^2 + [\theta^{(2)}(f)]^2\}$. ■

Our final example provides new results on the J-test of overidentification in GMM settings proposed by Sargan (1958, 1959) and further developed in Hansen (1982). The novelty here lies in the accommodation of partial identification and Jacobian matrices not of full column rank.

Example 1.2.6 (Overidentification Test)**.** Let $X \in \mathbf{R}^{d_x}$ be a random vector and consider the model defined by the moment restriction $E[g(X, \gamma_0)] = 0$ for some $\gamma_0 \in \Gamma \subset \mathbf{R}^k$ where $g : \mathbf{R}^{d_x} \times \Gamma \to \mathbf{R}^m$ is a known function with $m > k$. The conventional J-test can be recast by estimating the functional ϕ defined as: for some known $m \times m$ symmetric positive definite matrix W,

$$\phi(\theta_0) = \inf_{\gamma \in \Gamma} E[g(X, \gamma)]^\intercal W E[g(X, \gamma)] \ . \tag{1.11}$$

Here, $\theta_0 \in \prod_{j=1}^m \ell^\infty(\Gamma)$ is defined by $\theta_0(\gamma) = E[g(X, \gamma)]$, $\mathbb{D} = \prod_{j=1}^m \ell^\infty(\Gamma)$, $\mathbb{E} = \mathbf{R}$, and $\phi : \prod_{j=1}^m \ell^\infty(\Gamma) \to \mathbf{R}$ is defined by $\phi(\theta) = \inf_{\gamma \in \Gamma} \theta(\gamma)^\intercal W \theta(\gamma)$. The bootstrap for the J statistic has been studied by Hall and Horowitz (1996) and Andrews (2002). Note that θ_0 is always identified even though γ_0 is potentially partially identified. ■

1.2.3 Concepts of Differentiability

All examples in the previous subsection exhibit first order degeneracy, i.e., there exist points θ in $\mathbb{D}$ such that the first order derivative ϕ'_θ is 0 and in some cases ϕ is not even differentiable at θ, which can be seen from Examples 1.2.1 and 1.2.2 respectively. As such, we resort to a second order expansion that handles first order degeneracy and meanwhile accommodates potential nondifferentiability of ϕ. Let us proceed by recalling notions of first order differentiability (Shapiro, 1990; Fang and Santos, 2015)

Definition 1.2.1. Let $\mathbb{D}$ and $\mathbb{E}$ be normed spaces equipped with norms $\|\cdot\|_{\mathbb{D}}$ and $\|\cdot\|_{\mathbb{E}}$ respectively, and $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$.

(i) The map ϕ is said to be *Hadamard differentiable* at $\theta \in \mathbb{D}_\phi$ *tangentially* to a set $\mathbb{D}_0 \subseteq \mathbb{D}$, if there is a continuous linear map $\phi'_\theta : \mathbb{D}_0 \to \mathbb{E}$ such that:

$$\lim_{n\to\infty} \left\| \frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} - \phi'_\theta(h) \right\|_{\mathbb{E}} = 0 \ , \tag{1.12}$$

for all sequences $\{h_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbf{R}$ such that $t_n \to 0$, $h_n \to h \in \mathbb{D}_0$ as $n \to \infty$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for all n.

(ii) The map ϕ is said to be *Hadamard directionally differentiable* at $\theta \in \mathbb{D}_\phi$ *tangentially* to a set $\mathbb{D}_0 \subseteq \mathbb{D}$, if there is a continuous map $\phi'_\theta : \mathbb{D}_0 \to \mathbb{E}$ such that:$\phi'_\theta : \mathbb{D} \to \mathbb{E}$ such that:[1]

$$\lim_{n\to\infty} \left\| \frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} - \phi'_\theta(h) \right\|_{\mathbb{E}} = 0 \ , \tag{1.13}$$

for all sequences $\{h_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbf{R}_+$ such that $t_n \downarrow 0$, $h_n \to h \in \mathbb{D}_0$ as $n \to \infty$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for all n.

Inspecting Definition 1.2.1, we see that the main difference between Hadamard differentiability and directional differentiability lies in the linearity of the derivative. This turns out to be the exact gap between these two notions of differentiability. In particular,

[1] We note that the "tangential set" in Shapiro (1991) refers to the domain of ϕ (i.e., $\mathbb{D}_\phi$ in our context), whereas here it refers to the domain $\mathbb{D}_0$ of the derivative ϕ'_θ.

(1.13) ensures that the Hadamard directional derivative ϕ'_θ is necessarily continuous and positively homogeneous of degree one, though potentially nonlinear (Shapiro, 1990).

Given the introduced notions of differentiability and in view of the remarkable fact that Delta method is valid under even Hadamard directional differentiability in terms of deriving asymptotic distributions (Shapiro, 1991; Dümbgen, 1993), it seems a natural next step to invoke the Delta method. However, in the presence of first order degeneracy, the resulting limiting distribution is degenerate at zero, rendering substantial challenges for inferential purposes.

In essence, the Delta method is a stochastic version of Taylor expansion. Therefore, one could go one step further to explore the quadratic term when the linear term is degenerate. We thus follow Shapiro (2000) and define

Definition 1.2.2. Let $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ be a map as in Definition 1.2.1.

(i) Suppose that $\phi : \mathbb{D}_\phi \to \mathbb{E}$ is Hadamard differentiable tangentially to $\mathbb{D}_0 \subset \mathbb{D}$ such that the derivative $\phi'_\theta : \mathbb{D}_0 \to \mathbb{E}$ is well defined on $\mathbb{D}$. We say that ϕ is *second order Hadamard differentiable* at $\theta \in \mathbb{D}_\phi$ *tangentially* to $\mathbb{D}_0$ if there is a bilinear map $\Phi''_\theta : \mathbb{D}_0 \times \mathbb{D}_0 \to \mathbb{E}$ such that: for $\phi''_\theta(h) \equiv \Phi''_\theta(h, h)$,

$$\lim_{n\to\infty} \| \frac{\phi(\theta + t_n h_n) - \phi(\theta) - t_n \phi'_\theta(h_n)}{t_n^2} - \phi''_\theta(h) \|_{\mathbb{E}} = 0 \ , \tag{1.14}$$

for all sequences $\{h_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbf{R}^+$ such that $t_n \to 0$, $h_n \to h \in \mathbb{D}_0$ as $n \to \infty$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for all n.

(ii) Suppose that $\phi : \mathbb{D}_\phi \to \mathbb{E}$ is Hadamard directionally differentiable tangentially to $\mathbb{D}_0 \subset \mathbb{D}$ such that the derivative $\phi'_\theta : \mathbb{D}_0 \to \mathbb{E}$ is well defined on $\mathbb{D}$. We say that ϕ is *second order Hadamard directionally differentiable* at $\theta \in \mathbb{D}_\phi$ *tangentially* to $\mathbb{D}_0$ if

there is a map $\phi''_\theta : \mathbb{D}_0 \to \mathbb{E}$ such that:[2]

$$\lim_{n\to\infty} \| \frac{\phi(\theta + t_n h_n) - \phi(\theta) - t_n \phi'_\theta(h_n)}{t_n^2} - \phi''_\theta(h) \|_{\mathbb{E}} = 0 \ , \tag{1.15}$$

for all sequences $\{h_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbf{R}^+$ such that $t_n \downarrow 0$, $h_n \to h \in \mathbb{D}_0$ as $n \to \infty$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for all n.

The second order derivative ϕ''_θ in both cases is necessarily continuous on $\mathbb{D}_0$, which can be shown in a straightforward manner as in the proof of Proposition 3.1 in Shapiro (1990). Similar in spirit to Definition 1.2.1, the key difference between the above two notions of second order differentiability is that the former is a quadratic form corresponding to a bilinear map while the latter is in general only positively homogeneous of degree two, i.e., $\phi''_\theta(th) = t^2 \phi''_\theta(h)$ for all $t \geq 0$ and all $h \in \mathbb{D}_0$. The definition of second order Hadamard (resp. directional) differentiability is defined given first order Hadamard (resp. directional) differentiability. However, it is possible that ϕ is first order Hadamard differentiable but only second order Hadamard directionally differentiable (see Example 1.2.2). It is also possible that ϕ is first order Hadamard directional differentiable and yet its second order derivative is a bilinear map.[3] In all our examples, ϕ is first order Hadamard differentiable though ϕ'_θ may be degenerate; see Subsection 1.2.3.1. We stress that requiring ϕ'_θ to be well defined on the entirety of $\mathbb{D}$ does not demand differentiability on $\mathbb{D}$. Instead, it just means that ϕ'_θ can take elements potentially not in $\mathbb{D}_0$ as arguments. Finally, we note that first and second order (directional) derivatives share the same domain $\mathbb{D}_0$.

Clearly, the second order is not the end of the story. If ϕ''_θ in turn is degenerate, one can go beyond the second order; see Remark 1.2.1. We do not pursue this possibility at length in the current paper.

Remark 1.2.1. Suppose that $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ is $(p-1)$-th order Hadamard directionally differentiable tangentially to $\mathbb{D}_0 \subset \mathbb{D}$ such that the derivative $\phi^{(j)}_\theta : \mathbb{D}_0 \to \mathbb{E}$ is well defined on $\mathbb{D}$ for all $j = 1, \ldots, p-1$, where $p \geq 2$. Then we say that ϕ is pth order Hadamard

[2]Compared with Shapiro (2000), we omitted $\frac{1}{2}$ in the denominator for notational compactness.
[3]For example, consider the map $\phi : \mathbf{R} \to \mathbf{R}$ defined by $\phi(\theta) = \max\{\theta, 0\}$.

directionally differentiable at $\theta \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$ if there is a map $\phi_\theta^{(p)} : \mathbb{D}_0 \to \mathbb{E}$ such that:

$$\phi(\theta + t_n h_n) = \phi(\theta) + \sum_{j=1}^{p-1} t_n^j \phi_\theta^{(j)}(h_n) + t_n^p \phi_\theta^{(p)}(h) + o(t_n^p) \ , \tag{1.16}$$

for all sequences $\{h_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbf{R}^+$ such that $t_n \downarrow 0$, $h_n \to h \in \mathbb{D}_0$ as $n \to \infty$ and $\theta + t_n h_n \in \mathbb{D}_\phi$ for all n. ■

1.2.3.1 Examples Revisited

We now turn to the examples introduced in Section 1.2.2 and demonstrate how they fit into the scope of our analysis by calculating the derivatives. From now on, we shall focus on Examples 1.2.1, 1.2.4 and 1.2.6 exclusively for conciseness; Examples 1.2.2, 1.2.3 and 1.2.5 will be treated in Appendix 1.8.2.

Example 1.2.1 (Continued). In this example, the functional involved is second order Hadamard differentiable. Trivially we have

$$\phi'_\theta(h) = 2\theta h \ , \ \phi''_\theta(h) = h^2 \ . \tag{1.17}$$

Note that the first order derivative ϕ'_θ is degenerate when $\theta = 0$, whereas ϕ''_θ is everywhere nondegenerate. The bilinear map $\Phi''_\theta : \mathbf{R}^2 \to \mathbf{R}$ here is given by $\Phi''_\theta(h, g) = hg$. ■

Examples 1.2.4 involves a functional whose domain is infinite dimensional.

Example 1.2.4 (Continued). By Lemma 1.8.4, ϕ is first order Hadamard differentiable at any $\theta \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ with $\phi'_\theta : \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R}) \to \mathbf{R}$ satisfying for any $h = (h^{(1)}, h^{(2)}) \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$

$$\phi'_\theta(h) = 2 \int_{B_+(\theta)} [\theta^{(1)}(u) - \theta^{(2)}(u)][h^{(1)}(u) - h^{(2)}(u)] w(u) du \ ,$$

where $B_+(\theta) \equiv \{u \in \mathbf{R} : \theta^{(1)}(u) > \theta^{(2)}(u)\}$. Note that $\phi'_\theta(h) = 0$ if $B_+(\theta)$ has Lebesgue

measure zero, i.e., $\theta^{(1)} \leq \theta^{(2)}$ almost everywhere. Moreover, ϕ is second order Hadamard directionally differentiable at any $\theta \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ with the derivative $\phi''_\theta : \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R}) \to \mathbf{R}$ satisfying

$$\phi''_\theta(h) = \int_{B_0(\theta)} \max\{h^{(1)}(u) - h^{(2)}(u), 0\}^2 w(u)du + \int_{B_+(\theta)} [h^{(1)}(u) - h^{(2)}(u)]^2 w(u)du \ ,$$

for any $h = (h^{(1)}, h^{(2)}) \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$, where $B_0(\theta) \equiv \{u \in \mathbf{R} : \theta^{(1)}(u) = \theta^{(2)}(u)\}$ is referred to as the contact set of $\theta^{(1)}$ and $\theta^{(2)}$. If $\theta^{(1)} \leq \theta^{(2)}$, then $\phi''_\theta(h)$ simplifies to $\phi''_\theta(h) = \int_{B_0(\theta)} \max\{h^{(1)}(u) - h^{(2)}(u)\}^2 w(u)du$. If in addition the contact set $B_0(\theta)$ has Lebesgue measure zero, then ϕ''_θ in turn is degenerate, corresponding to the degenerate limits obtained in Theorem 1 of Linton et al. (2010). ■

In Example 1.2.6, the domain $\mathbb{D}_0$ of the derivative ϕ''_{θ_0} is a strict subset of $\mathbb{D}$.

Example 1.2.6 (Continued). Consider $\theta \in \prod_{j=1}^m \ell^\infty(\Gamma)$ such that $\theta(\gamma) = 0$ for some $\gamma \in \Gamma$. Then ϕ is Hadamard differentiable at θ and $\phi'_\theta(h) = 0$ for all $h \in \prod_{j=1}^m \ell^\infty(\Gamma)$. Suppose further that Γ is compact and that $\Gamma_0(\theta) \equiv \{\gamma \in \Gamma : \theta(\gamma) = 0\}$ is in the interior of Γ. For $C^1(\Gamma)$ the space of continuously differentiable functions on Γ, if $\theta \in \prod_{j=1}^m C^1(\Gamma)$, then by Lemma 1.8.6, under additional regularity conditions, ϕ is second order Hadamard directionally differentiable at θ tangentially to $\prod_{j=1}^m C(\Gamma)$ with the derivative given by: for any $h \in \prod_{j=1}^m C(\Gamma)$,

$$\phi''_\theta(h) = \min_{\gamma_0 \in \Gamma_0(\theta)} \min_{v \in \mathbf{R}^k} \{h(\gamma_0) - J(\gamma_0)v\}^\intercal W \{h(\gamma_0) - J(\gamma_0)v\} \ ,$$

where $J(\gamma_0) \equiv \frac{d\theta(\gamma)}{d\gamma^\intercal}\big|_{\gamma=\gamma_0}$ is the Jacobian matrix. When γ_0 is point identified and $J(\gamma_0)$ is of full column rank, ϕ becomes second order Hadamard differentiable with

$$\phi''_\theta(h) = h(\gamma_0)^\intercal M(\gamma_0) W M(\gamma_0) h(\gamma_0) \ ,$$

where $M(\gamma_0) = I_m - J(\gamma_0)[J(\gamma_0)^\intercal J(\gamma_0)]^{-1} J(\gamma_0)^\intercal$ with I_m the identity matrix of size m. We note that $\Gamma_0(\theta)$ being in the interior of Γ is not essential and can be relaxed by introducing

relevant notions of cones (Bonnans and Shapiro, 2000). ■

1.2.4 Second Order Delta Method

The Delta method for potentially directionally differentiable maps as well as differentiable ones has proven powerful in asymptotic analysis (van der Vaart, 1998; Shapiro, 1991; Fang and Santos, 2015; Hansen, 2015). Unfortunately, it is insufficient to handle substantial challenges for inference arising from first order degeneracy. Heuristically, if $r_n\{\hat{\theta}_n - \theta_0\} \xrightarrow{L} \mathbb{G}$ and $\phi'_{\theta_0} = 0$, then the Delta method implies that

$$r_n\{\phi(\hat{\theta}_n) - \phi(\theta_0)\} \xrightarrow{L} \phi'_{\theta_0}(\mathbb{G}) \equiv 0 \ .$$

For real-valued ϕ, the usual confidence interval for $\phi(\theta_0)$ at asymptotic level $1-\alpha$ is

$$[\phi(\hat{\theta}_n) - \frac{c_{1-\alpha/2}}{r_n}, \phi(\hat{\theta}_n) - \frac{c_{\alpha/2}}{r_n}] = \{\phi(\hat{\theta}_n)\} \ , \tag{1.18}$$

where the c_α is the α-th quantile of $\phi'_{\theta_0}(\mathbb{G}) \equiv 0$ and is zero for all $\alpha \in (0,1)$. Clearly, $P(\phi(\theta_0) \in \{\phi(\hat{\theta}_n)\}) = 0$ if, for example, $\phi(\hat{\theta}_n)$ is a continuous random variable.

To circumvent the above difficulty, we resort to higher order expansions and aim to establish a mild extension of Theorem 2.3 in Shapiro (2000) by accommodating weak convergence in the Hoffmann-Jørgensen sense and dispensing with the convexity of $\mathbb{D}_\phi$ and separability of $\mathbb{D}$. We proceed by imposing the following assumptions.

Assumption 1.2.1. *(i) $\mathbb{D}$ and $\mathbb{E}$ are normed spaces with norms $\|\cdot\|_\mathbb{D}$ and $\|\cdot\|_\mathbb{E}$ respectively; (ii) $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{E}$ is second order Hadamard directionally differentiable at $\theta_0 \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0 \subset \mathbb{D}$.*

Assumption 1.2.2. *(i) There is $\hat{\theta}_n : \{X_i\}_{i=1}^n \to \mathbb{D}_\phi$ such that $r_n\{\hat{\theta}_n - \theta_0\} \xrightarrow{L} \mathbb{G}$ in $\mathbb{D}$ for some $r_n \uparrow \infty$; (ii) $\mathbb{G}$ is tight and $P(\mathbb{G} \in \mathbb{D}_0) = 1$.*

Assumption 1.2.3. *(i) ϕ''_θ can be continuously extended to $\mathbb{D}$; (ii) $\mathbb{D}_0$ is closed under vector addition, i.e., $h_1 + h_2 \in \mathbb{D}_0$ whenever $h_1, h_2 \in \mathbb{D}_0$; (iii) $\phi'_{\theta_0}(h) = 0$ for all $h \in \mathbb{D}_0$.*

Assumption 1.2.1 formalizes the requirement that $\phi : \mathbb{D}_\phi \to \mathbb{E}$ is second order Hadamard directionally differentiable at θ_0, which allows us to conduct higher order expansions. By definition, this necessitates first order Hadamard directional differentiability of ϕ at θ_0 tangentially to $\mathbb{D}_0$ and that ϕ'_{θ_0} is well defined on the entirety of $\mathbb{D}$. Assumption 1.2.2(i) characterizes another key ingredient: there is an estimator $\hat{\theta}_n$ for θ_0 that admits a weak limit $\mathbb{G}$ at a potentially non-$\sqrt{n}$ rate r_n; see Remark 1.3.1. Assumption 1.2.2(ii) ensures that the support of $\mathbb{G}$ is included in the domain of the derivative ϕ''_{θ_0} so that $\phi''_{\theta_0}(\mathbb{G})$ is well defined, while tightness of $\mathbb{G}$ is only a minimal requirement.[4] Assumption 1.2.3(i) allows us to view the map ϕ''_{θ_0} as well defined and continuous on the entire space $\mathbb{D}$, and is automatically satisfied whenever $\mathbb{D}_0$ is closed (Dugundji, 1951, Theorem 4.1). We emphasize, however, that Assumption 1.2.3(i) does not require differentiability of $\phi : \mathbb{D}_\phi \to \mathbb{E}$ tangentially to $\mathbb{D}$, i.e., the extension of ϕ''_{θ_0} need not satisfy (1.15) for $h \in \mathbb{D} \setminus \mathbb{D}_0$. Assumption 1.2.3(ii) imposes that $\mathbb{D}_0$ be closed under addition which, since $\mathbb{D}_0$ is necessarily a cone, is equivalent to demanding that $\mathbb{D}_0$ be convex.[5] This mild requirement is only employed in some of our results and helps ensure that, when multiple extensions of ϕ''_{θ_0} exist, the choice of extension has no impact in our arguments. Finally, Assumption 1.2.3(iii) formalizes the defining feature of the paper, i.e., first order degeneracy of ϕ.

Given Assumptions 1.2.1 and 1.2.2, we are able to establish a second order Delta method. Assumption 1.2.3(i) is needed to obtain a strengthening in which ϕ''_{θ_0} takes elements potentially in $\mathbb{D} \setminus \mathbb{D}_0$ as arguments.

Theorem 1.2.1. *If Assumptions 1.2.1 and 1.2.2 hold, then*

$$r_n^2\{\phi(\hat{\theta}_n) - \phi(\theta_0) - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)\} \xrightarrow{L} \phi''_{\theta_0}(\mathbb{G}) \ . \tag{1.19}$$

If in addition Assumption 1.2.3(i) holds, then

$$r_n^2\{\phi(\hat{\theta}_n) - \phi(\theta_0) - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)\} = \phi''_{\theta_0}(r_n\{\hat{\theta}_n - \theta_0\}) + o_p(1) \ . \tag{1.20}$$

[4]The support of $\mathbb{G}_0$ is the set of all $x \in \mathbb{D}$ having the property that $P(\mathbb{G}_0 \in U) > 0$ for each open set $U \subset \mathbb{D}$ containing x.

[5]We note that convexity of $\mathbb{D}_0$ is only needed for the stronger version of the Delta method in (1.20).

The essence of Theorem 1.2.1 is in complete accord with that underlying the first order Delta method. In particular, the definition of second order Hadamard directional differentiability is engineered so that the second order Delta method is nothing more than a stochastic version of the Taylor expansion of order two, i.e.,

$$\phi(\theta_0 + t_n h_n) = \phi(\theta_0) + t_n \phi'_{\theta_0}(h_n) + t_n^2 \phi''_{\theta_0}(h) + o(t_n^2) \ ,$$

where t_n corresponds to r_n^{-1}, and h_n to $r_n\{\hat{\theta}_n - \theta_0\}$. Note that Theorem 1.2.1 is valid regardless of the nature of the differentiability (i.e., fully differentiable or directionally differentiable) and the presence of first order degeneracy. When ϕ'_{θ_0} is degenerate, the convergence (1.19) simplifies to

$$r_n^2\{\phi(\hat{\theta}_n) - \phi(\theta_0)\} \xrightarrow{L} \phi''_{\theta_0}(\mathbb{G}) \ . \tag{1.21}$$

Finally, we note that higher order versions of the Delta method can be developed along the lines of Remark 1.2.1; see Remark 1.2.2.

Remark 1.2.2. Suppose that Assumptions 1.2.1(i) and 1.2.2 hold and that ϕ is p-th order Hadamard directionally differentiable at $\theta_0 \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$. It follows that

$$r_n^p\big[\phi(\hat{\theta}_n) - \{\phi(\theta_0) + \sum_{j=1}^{p-1} \phi_{\theta_0}^{(j)}(\hat{\theta}_n - \theta_0)\}\big] \xrightarrow{L} \phi_{\theta_0}^{(p)}(\mathbb{G}) \ .$$

If in addition $\phi_{\theta_0}^{(p)}$ can be continuously extended to $\mathbb{D}$, then

$$r_n^p\big[\phi(\hat{\theta}_n) - \{\phi(\theta_0) + \sum_{j=1}^{p-1} \phi_{\theta_0}^{(j)}(\hat{\theta}_n - \theta_0)\}\big] = \phi_{\theta_0}^{(p)}(r_n\{\hat{\theta}_n - \theta_0\}) + o_p(1) \ . \qquad \blacksquare$$

1.3 The Bootstrap

Establishing asymptotic distributions as in Theorem 1.2.1 is the first step towards conducting statistical inference on $\phi(\theta_0)$, the usefulness of which relies on our ability to

accurately estimate the limiting law. In this section, we discuss how first order degeneracy of ϕ can complicate inference using the standard bootstrap based on first and especially second order asymptotics, and provide alternative consistent resampling schemes.

1.3.1 Bootstrap Setup

Throughout, we let $\hat{\theta}_n^*$ denote a "bootstrapped version" of $\hat{\theta}_n$, which is defined as a function mapping the data $\{X_i\}_{i=1}^n$ and random weights $\{W_i\}_{i=1}^n$ that are independent of $\{X_i\}_{i=1}^n$ into the domain $\mathbb{D}_\phi$ of ϕ. This general definition allows us to include diverse resampling schemes such as nonparametric, Bayesian, block, score, more generally multiplier and exchangeable bootstrap as special cases. We shall assume the limiting distribution $\mathbb{G}$ of $r_n\{\hat{\theta}_n - \theta_0\}$ can be consistently estimated by the law of $r_n\{\hat{\theta}_n^* - \hat{\theta}_n\}$ conditional on the data.

Next, making sense of bootstrap consistency necessitates a metric that quantifies distances between probability measures. As is standard in the literature, we employ the bounded Lipschitz metric d_{BL} formalized by Dudley (1966, 1968): for two Borel probability measures L_1 and L_2 on $\mathbb{D}$, define

$$d_{\mathrm{BL}}(L_1, L_2) \equiv \sup_{f \in \mathrm{BL}_1(\mathbb{D})} \left| \int f \, dL_1 - \int f \, dL_2 \right| ,$$

where we recall that $\mathrm{BL}_1(\mathbb{D})$ denotes the set of Lipschitz functionals whose absolute level and Lipschitz constant are bounded by one, i.e.,

$$\mathrm{BL}_1(\mathbb{D}) \equiv \{ f : \mathbb{D} \to \mathbf{R} : \sup_{t \in \mathbb{D}} |f(t)| + \sup_{t_1, t_2 \in \mathbb{D}, t_1 \neq t_2} \frac{|f(t_1) - f(t_2)|}{\|t_1 - t_2\|_{\mathbb{D}}} \leq 1 \} .$$

Since weak convergence in the Hoffmann-Jørgensen sense to separable limits can be metrized by d_{BL} (Dudley, 1990; van der Vaart and Wellner, 1990), we may now measure the distance between the "law" of $\hat{\mathbb{G}}_n^* \equiv r_n\{\hat{\theta}_n^* - \hat{\theta}_n\}$ conditional on $\{X_i\}$ and the limiting

law of $r_n\{\hat{\theta}_n - \theta_0\}$ by

$$d_{\mathrm{BL}}(\hat{\mathbb{G}}_n^*, \mathbb{G}) = \sup_{f \in \mathrm{BL}_1(\mathbb{D})} |E_W[f(r_n\{\hat{\theta}_n^* - \hat{\theta}_n\})] - E[f(\mathbb{G})]| \ , \tag{1.22}$$

where E_W denotes expectation with respect to the bootstrap weights $\{W_i\}_{i=1}^n$ holding the data $\{X_i\}_{i=1}^n$ fixed. Employing the distribution of $r_n\{\hat{\theta}_n^* - \hat{\theta}_n\}$ conditional on the data as an approximation to the distribution of $\mathbb{G}$ is then asymptotically justified if their distance, equivalently (1.22), converges in probability to zero.

We formalize the above discussion by imposing the following assumptions on $\hat{\theta}_n^*$.

Assumption 1.3.1. *(i)* $\hat{\theta}_n^* : \{X_i, W_i\}_{i=1}^n \to \mathbb{D}_\phi$ *with* $\{W_i\}_{i=1}^n$ *independent of* $\{X_i\}_{i=1}^n$*; (ii)* $\hat{\theta}_n^*$ *satisfies* $\sup_{f \in BL_1(\mathbb{D})} |E_W[f(r_n\{\hat{\theta}_n^* - \hat{\theta}_n\})] - E[f(\mathbb{G})]| = o_p(1)$.

Assumption 1.3.2. *(i)* $E[f(r_n\{\hat{\theta}_n^* - \hat{\theta}_n\})^*] - E[f(r_n\{\hat{\theta}_n^* - \hat{\theta}_n\})_*] \to 0$ *for all* $f \in BL_1(\mathbb{D})$ *where* $f(r_n\{\hat{\theta}_n^* - \hat{\theta}_n\})^*$ *and* $f(r_n\{\hat{\theta}_n^* - \hat{\theta}_n\})_*$ *denote minimal measurable majorant and maximal measurable minorant (with respect to* $\{X_i, W_i\}_{i=1}^n$ *jointly) respectively; (ii)* $f(r_n\{\hat{\theta}_n^* - \hat{\theta}_n\})$ *is a measurable function of* $\{W_i\}_{i=1}^n$ *outer almost surely in* $\{X_i\}_{i=1}^n$ *for any continuous and bounded* $f : \mathbb{D} \to \mathbf{R}$.

Assumption 1.3.1(i) formally defines the bootstrap analog $\hat{\theta}_n^*$ of $\hat{\theta}_n$, while Assumption 1.3.1(ii) simply imposes the consistency of the "law" of $r_n\{\hat{\theta}_n^* - \hat{\theta}_n\}$ conditional on the data for the law of $\mathbb{G}$, i.e., the bootstrap "works" for the estimator $\hat{\theta}_n$. Assumption 1.3.2 is of technical concern. In particular, Assumption 1.3.2(i) can often be established as a result of bootstrap consistency (van der Vaart and Wellner, 1996a), while Assumption 1.3.2(ii) is easy to verify for particular resampling schemes. For example, if $\{W_i\}_{i=1}^n \mapsto f(r_n\{\hat{\theta}_n^* - \hat{\theta}_n\})$ is continuous, then Assumption 1.3.2(ii) is fulfilled. When θ_0 is Euclidean-valued, i.e., $\mathbb{D} = \mathbf{R}^k$ with $k \in \mathbf{N}$, one can dispense with Assumption 1.3.2.

1.3.2 Failures of the Standard Bootstrap

We now turn to the challenges for inferences using the standard bootstrap caused by first order degeneracy. As is well known in the literature, the law of

$$r_n\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\} \tag{1.23}$$

conditional on the data provides a consistent estimator of the law of $\phi'_{\theta_0}(\mathbb{G})$ provided ϕ is Hadamard differentiable (van der Vaart and Wellner, 1996a), which in particular includes the case when $\phi'_{\theta_0} = 0$. In other words, the standard bootstrap, meaning the law of (1.23) conditional on the data, is consistent for the law of $\phi'_{\theta_0}(\mathbb{G})$ regardless of the presence of first order degeneracy.

Substantial difficulties, however, arise from using (1.23) for inferential purposes when first order degeneracy does occur. Ignoring the first order degeneracy or perhaps as a way to avoid ridiculous confidence intervals such as (1.18), one might consider the following confidence interval for real-valued $\phi(\theta_0)$:

$$[\phi(\hat{\theta}_n) - \frac{\tilde{c}_{1-\alpha/2}}{r_n}, \phi(\hat{\theta}_n) - \frac{\tilde{c}_{\alpha/2}}{r_n}] \ , \tag{1.24}$$

where $\tilde{c}_{1-\alpha}$ is the $(1-\alpha)$-th bootstrapped quantile for $\alpha \in (0,1)$ defined as

$$\tilde{c}_{1-\alpha} \equiv \inf\{c \in \mathbf{R} : P_W(r_n\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\} \leq c) \geq 1-\alpha\} \ .$$

However, establishing the validity of (1.24) as a level $1-\alpha$ confidence interval for $\phi(\theta_0)$ is problematic because $\tilde{c}_{1-\alpha} \xrightarrow{p} 0$ for all $\alpha \in (0,1)$ and 0 is a discontinuity point of the cdf of the limit (see Lemma 1.8.1).

In fact, simple algebra reveals that (1.24) is numerically identical to

$$[\phi(\hat{\theta}_n) - \frac{\bar{c}_{1-\alpha/2}}{r_n^2}, \phi(\hat{\theta}_n) - \frac{\bar{c}_{\alpha/2}}{r_n^2}] \ , \tag{1.25}$$

where $\bar{c}_\alpha$ is defined as

$$\bar{c}_{1-\alpha} \equiv \inf\{c \in \mathbf{R} : P_W(r_n^2\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\} \leq c) \geq 1-\alpha\} \ .$$

In other words, $\bar{c}_\alpha$ is the α-th bootstrapped quantile of the standard bootstrap based on second order asymptotics:

$$r_n^2\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\} \ . \tag{1.26}$$

As illustrated by Babu (1984) for the squared mean example, the law of (1.26) conditional on the data is inconsistent for the law of $\phi''_{\theta_0}(\mathbb{G})$ when $\theta_0 = 0$, the point at which first order degeneracy arises. We next demonstrate that the bootstrap failure is not peculiar to this example by generalizing it to our general setup. As the following theorem shows, for centered Gaussian $\mathbb{G}$, the second order standard bootstrap is consistent if and only if ϕ''_{θ_0} is degenerate.

Theorem 1.3.1. *Suppose that Assumptions 1.2.1, 1.2.2, 1.2.3, 1.3.1 and 1.3.2 hold, and that $\mathbb{G}$ is centered Gaussian. Then $\phi''_{\theta_0} = 0$ on the support of $\mathbb{G}$ if and only if*

$$\sup_{f \in \mathrm{BL}_1(\mathbb{E})} |E_W[f(r_n^2\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\})] - E[f(\phi''_{\theta_0}(\mathbb{G}))]| = o_p(1) \ . \tag{1.27}$$

If, in addition, ϕ is second order Hadamard differentiable, then the conclusion holds without requiring $\mathbb{G}$ be centered Gaussian.

The sufficiency part of the theorem is somewhat expected and not a deep result, while the necessity is perhaps surprising and has far-reaching implications for statistical inference as we shall detail shortly. The proof of the latter consists of two steps: in the first step, we show that bootstrap consistency as in (1.27) implies existence of a bilinear map Φ''_{θ_0} corresponding to ϕ''_{θ_0}, in similar fashion as the proof of Theorem 3.1 in Fang and Santos (2015); in the second step, we establish that Φ''_{θ_0} and hence ϕ''_{θ_0} is necessarily degenerate. Both steps involve the insights of equating distributions through their characteristic

functionals as in van der Vaart (1991) and Hirano and Porter (2012).

Theorem 1.3.1 implies that, in the presence of first order degeneracy, if the second order derivative ϕ''_{θ_0} is nondegenerate, then the standard bootstrap based on second order asymptotics is necessarily inconsistent whenever $\mathbb{G}$ is centered Gaussian. If ϕ''_{θ_0} is degenerate, we have a degenerate limiting distribution that can not be directly used for inference. We thus conclude that bootstrap failure is an inherent implication of models with first order degeneracy.

Heuristically, the reason why the standard bootstrap fails is that even though $r_n^2\phi'_{\theta_0}(\hat{\theta}_n - \theta_0) = 0$ in the "real world", its bootstrap counterpart is non-negligible. To see this, consider the squared mean example. If $\theta_0 = 0$, then

$$n\phi'_{\hat{\theta}_n}(\hat{\theta}_n^* - \hat{\theta}_n) = n2\hat{\theta}_n \cdot \{\hat{\theta}_n^* - \hat{\theta}_n\} = 2\sqrt{n}\{\hat{\theta}_n - \theta_0\} \cdot \sqrt{n}\{\hat{\theta}_n^* - \hat{\theta}_n\} = O_p(1) \ .$$

This is an emphatic reflection of Efron (1979)'s caveat that the bootstrap, as well as other resampling schemes, provides frequency approximations rather than likelihood approximations. These heuristics suggest that the standard bootstrap might work if the first order term $r_n^2\phi'_{\hat{\theta}_n}(\hat{\theta}_n^* - \hat{\theta}_n)$ is included, which turns out to be true for sufficiently smooth maps; see Theorem 1.3.2.

It is worth noting that Theorem 1.3.1 holds even if ϕ is smooth. Consequently, first order degeneracy is a source of bootstrap inconsistency completely different from that discussed in Fang and Santos (2015), i.e., nondifferentiability of ϕ. In addition, we note that, without the qualifier that $\mathbb{G}$ is centered Gaussian, bootstrap consistency (1.27) holds if and only if $\phi''_{\theta_0}(\mathbb{G}+h) - \phi''_{\theta_0}(h) \stackrel{d}{=} \phi''_{\theta_0}(\mathbb{G})$ for all $h \in \text{Supp}(\mathbb{G})$ under mild support conditions; see Theorem A.1 in Fang and Santos (2015).

1.3.3 The Babu Correction

We now extend the Babu correction under our more general setup. We proceed by imposing the following assumption.

Assumption 1.3.3. *(i) The map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{E}$ is second order Hadamard differentiable at $\theta_0 \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$; (ii) ϕ is first order Hadamard differentiable at every point in some neighborhood of θ_0 tangentially to $\mathbb{D}_0$ such that* [6]

$$\lim_{n\to\infty} \| \frac{\phi'_{\theta_0+t_n g_n}(h_n) - \phi'_{\theta_0}(h_n)}{t_n} - 2\Phi''_{\theta_0}(g,h) \|_{\mathbb{E}} = 0 \ , \tag{1.28}$$

for all sequences $\{g_n, h_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbf{R}^+$ such that $t_n \downarrow 0$, $(g_n, h_n) \to (g, h) \in \mathbb{D}_0 \times \mathbb{D}_0$ as $n \to \infty$ and $\theta + t_n g_n, \theta + t_n h_n \in \mathbb{D}_\phi$ for all sufficiently large n, where $\Phi''_{\theta_0} : \mathbb{D}_0 \times \mathbb{D}_0 \to \mathbb{E}$ is the bilinear map underlying ϕ''_{θ_0}.

Assumption 1.3.3(i) defines the scope of our current discussion: the Babu correction shall be applied to smooth maps. Assumption 1.3.3(ii) is stronger than ϕ being simply second order Hadamard differentiable, in that it requires the existence of first order derivative at all points in a neighborhood of θ_0 such that (1.3.3) holds. Assumption 1.3.3 is fulfilled for the setup considered in Babu (1984) and for Examples 1.2.1 and 1.2.3, but violated for the remaining examples.

Under Assumption 1.3.3, the corrected bootstrap

$$r_n^2\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n) - \phi'_{\hat{\theta}_n}(\hat{\theta}_n^* - \hat{\theta}_n)\} \tag{1.29}$$

is consistent for the law of $\phi''_{\theta_0}(\mathbb{G})$ regardless of the degeneracy of ϕ'_{θ_0}.

Theorem 1.3.2. *Suppose that Assumptions 1.2.1(i), 1.2.2, 1.2.3(ii), 1.3.1, 1.3.2 and 1.3.3 holds. If the bilinear form Φ''_{θ_0} can be continuously extended to $\mathbb{D} \times \mathbb{D}$, then*

$$\sup_{f\in \mathrm{BL}_1(\mathbb{E})} |E_W[f(r_n^2\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n) - \phi'_{\hat{\theta}_n}(\hat{\theta}_n^* - \hat{\theta}_n)\})] - E[f(\phi''_{\theta_0}(\mathbb{G}))]| = o_p(1) \ . \tag{1.30}$$

Theorem 1.3.2 generalizes Babu (1984) considerably in that it accommodates semiparametric and nonparametric models, and allows wider resampling schemes beyond the nonparametric bootstrap of Efron (1979). The Babu correction works nicely with smooth

[6] The appearance of the factor 2 is due to omission of the factor 1/2 in Definition 1.2.2.

maps in the sense of Assumption 1.3.3, but unfortunately is inadequate to handle nonsmooth ones. This is because when ϕ is only second order directionally differentiable, often times the derivative ϕ''_{θ_0} is not "continuous" in θ_0, implying that the Babu correction (1.29) is unable to estimate ϕ''_{θ_0} properly and in this way results in inconsistent estimates. For this reason, we next provide yet another resampling method which accommodates nondifferentiable maps.

1.3.4 A Modified Bootstrap

In this subsection, we shall present a modified bootstrap following Fang and Santos (2015) that is consistent for the law of $\phi''_{\theta_0}(\mathbb{G})$, and adaptive to both the presence of first order degeneracy and nondifferentiability of ϕ.

The heuristics underlying our proposal, however, are connected to those in Fang and Santos (2015) in a subtle way. In the context of first order asymptotics where ϕ is only directionally differentiable, inconsistency of the standard bootstrap arises from its inability to properly estimate the directional derivative ϕ'_{θ_0}. In our setup, however, there are examples in which the derivative ϕ''_{θ_0} is a known map; see Examples 1.2.1 and 1.2.3 which are all differentiable maps. The standard bootstrap in these settings fails because there is a non-negligible term being neglected. However, in all other examples where ϕ is not smooth enough, Fang and Santos (2015)'s arguments will come into play as well.

In any case, the second order weak limit $\phi''_{\theta_0}(\mathbb{G})$ is a composition of the derivative ϕ''_{θ_0} and the limit $\mathbb{G}$ of $\hat{\theta}_n$, as is the first order limit $\phi'_{\theta_0}(\mathbb{G})$. Thus, the law of $\phi''_{\theta_0}(\mathbb{G})$ can be estimated by composing a suitable estimator $\hat{\phi}''_n$ for ϕ''_{θ_0} with a consistent bootstrap approximation for the law of $\mathbb{G}$, in exactly the same fashion as the resampling scheme proposed by Fang and Santos (2015). That is, we propose employing the law of

$$\hat{\phi}''_n(r_n\{\hat{\theta}^*_n - \hat{\theta}_n\}) \tag{1.31}$$

conditional on the data as an approximation for the law of $\phi''_{\theta_0}(\mathbb{G})$, where $\hat{\phi}''_n : \mathbb{D} \to \mathbb{E}$ is a

suitable estimator of ϕ''_{θ_0}. In cases when ϕ''_{θ_0} is a known map, we may simply set $\hat{\phi}''_n = \phi''_{\theta_0}$ for all $n \in \mathbf{N}$.

Certainly, we would like $\hat{\phi}''_n$ to converge to ϕ''_{θ_0} in some sense as $n \to \infty$. This can be made precise as follows.

Assumption 1.3.4. *$\hat{\phi}''_n : \mathbb{D} \to \mathbb{E}$ is a function of $\{X_i\}_{i=1}^n$ satisfying that for every sequence $\{h_n\} \subset \mathbb{D}$ and every $h \in \mathbb{D}_0$ such that $h_n \to h$ as $n \to \infty$,*

$$\hat{\phi}''_n(h_n) \xrightarrow{p} \phi''_{\theta_0}(h) \ . \tag{1.32}$$

Assumption 1.3.4 says that $\hat{\phi}''_n$ converges in probability to ϕ''_{θ_0} along any convergent sequence $h_n \to h$ as $n \to \infty$. It is worth noting that Assumption 1.3.4 is equivalent to requiring: for every compact set $K \subset \mathbb{D}_0$ and every $\epsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P\Big(\sup_{h \in K^\delta} \|\hat{\phi}''_n(h) - \phi''_{\theta_0}(h)\|_{\mathbb{E}} > \epsilon \Big) = 0 \ , \tag{1.33}$$

where $K^\delta \equiv \{a \in \mathbb{D} : \inf_{b \in K} \|a - b\|_{\mathbb{D}} < \delta\}$; see Lemma 1.8.2. Condition (1.33) was employed in Fang and Santos (2015) who also provided several sufficient conditions for it to hold. For example, if $\hat{\phi}''_n : \mathbb{D} \to \mathbb{E}$ is Lipschitz continuous, then pointwise consistency of $\hat{\phi}''_n$ suffices for (1.33). Unfortunately, second order derivatives often lack uniform continuity and hence those sufficient conditions are inapplicable. Nonetheless, condition (1.32) is straightforward to verify in all our examples.

Given the equivalence of conditions (1.32) and (1.33), consistency of our modified bootstrap (1.31) follows from Theorem 3.2 in Fang and Santos (2015).

Theorem 1.3.3. *Under Assumptions 1.2.1, 1.2.2, 1.2.3(i), 1.3.1, 1.3.2 and 1.3.4, it follows that*

$$\sup_{f \in \mathrm{BL}_1(\mathbb{E})} |E_W[f(\hat{\phi}''_n(r_n\{\hat{\theta}^*_n - \hat{\theta}_n\}))] - E[f(\phi''_{\theta_0}(\mathbb{G}))]| = o_p(1) \ . \tag{1.34}$$

Theorem 1.3.3 shows that the law of $\hat{\phi}''_n(r_n\{\hat{\theta}^*_n - \hat{\theta}_n\})$ conditional on the data is indeed

consistent for the law of $\phi''_{\theta_0}(\mathbb{G})$, regardless of the degree of smoothness of ϕ and degeneracy of ϕ'_{θ_0}. Interestingly, the resampling scheme in Theorem 1.3.3 is a mixture of the classical bootstrap and analytical asymptotic approximations. Finally, we note that Assumption 1.3.4 allows us to think of Theorem 1.3.3 as a variant of the extended continuous mapping theorem.

We now briefly compare the Babu correction, the above composition procedure and the recentered bootstrap (Hall and Horowitz, 1996; Horowitz, 2001). In some cases (for instance, Example 1.2.1 and the regular J-test), they coincide with each other. However, the Babu correction applies to general smooth functionals, rather than just quadratic forms, and hence can be thought of as a generalization of the recentered bootstrap. The composition procedure, which works for an even larger class of functionals, is a direct approach by exploiting the structure of the limits, and hence is more tractable.

Remark 1.3.1. Examples where the convergence rate is not $\sqrt{n}$ include inference on the means of kernel density estimators (Hall, 1992),[7] smoothed maximum score estimators (Horowitz, 2002), and cointegration regressions (Chang et al., 2006). For nonstandard convergence rates, however, the bootstrap process $r_n\{\hat{\theta}^*_n - \hat{\theta}_n\}$ can fail to consistently estimate the law of $\mathbb{G}$, violating Assumption 1.3.1(ii). Fortunately, as far as Theorem 1.3.3 is concerned, any consistent estimator, which need not satisfy Assumption 1.3.1(ii), will do. For example, in cube-root estimation problems, one could instead employ some smoothed bootstrap $r_n\{\tilde{\theta}^*_n - \tilde{\theta}_n\}$ where $\tilde{\theta}^*_n$ and $\tilde{\theta}_n$ are some smoothed estimators, or m out of n resampling (or subsampling) $m_n\{\hat{\theta}^*_{m_n} - \hat{\theta}_n\}$ where $\hat{\theta}^*_{m_n}$ is a bootstrap estimator based on subsamples of size m_n. In the context of estimating nonincreasing density functions, see Kosorok (2008b) and Sen et al. (2010); for bootstrapping the maximum score estimators, see Delgado et al. (2001) and Patra et al. (2015). ■

[7] We stress that our assumptions on the primitive parameter θ_0 exclude cases where θ_0 is equal to a density function evaluated at a particular point due to the bias term.

1.3.5 Estimation of the Derivative

Given the posited bootstrap consistency for the law of $\mathbb{G}$, the remaining crucial piece towards consistent bootstrap for the law of $\phi''_{\theta_0}(\mathbb{G})$ based on Theorem 1.3.3 is then an estimator $\hat{\phi}''_n$ of the derivative ϕ''_{θ_0} that satisfies Assumption 1.3.4. There are two general approaches for estimation of ϕ''_{θ_0}: one by exploiting the structure of ϕ''_{θ_0}, and the other one based on numerical differentiation as we describe now.

When first order degeneracy occurs, we have

$$\phi''_{\theta_0}(h) = \lim_{n\to\infty} \frac{\phi(\theta_0 + t_n h) - \phi(\theta_0)}{t_n^2} \ . \tag{1.35}$$

Following Song (2014) and Hong and Li (2015), we may then estimate ϕ''_{θ_0} via numerical differentiation as follows: for any $h \in \mathbb{D}$,

$$\hat{\phi}''_n(h) = \frac{\phi(\hat{\theta}_n + t_n h) - \phi(\hat{\theta}_n)}{t_n^2} \ . \tag{1.36}$$

If t_n tends to zero at a suitable rate, the sense of which is made precise by the following assumption, then $\hat{\phi}''_n$ is a good estimator for ϕ''_{θ_0} in the sense of Assumption 1.3.4.

Assumption 1.3.5. *$\{t_n\}_{n=1}^{\infty}$ is a sequences of scalars such that $t_n \downarrow 0$ and $r_n t_n \to \infty$.*

The next proposition confirms the validity of the numerical estimator (1.36).

Proposition 1.3.1. *If Assumptions 1.2.1, 1.2.2, 1.2.3(iii) and 1.3.5 hold, then the numerical estimator $\hat{\phi}''_n$ in* (1.36) *satisfies Assumption 1.3.4.*

We note that numerical differentiation can also be employed to estimate the derivative when ϕ'_{θ_0} is degenerate only at points in a proper subset of the parameter space; see Remark 1.3.2. Proposition 1.3.1 provides a tractable way of estimating the derivative ϕ''_{θ_0}. On the other hand, the expression of ϕ''_{θ_0} itself often suggests an obvious estimator as we elaborate in next subsection.

Remark 1.3.2. If ϕ'_{θ_0} is possibly nondegenerate, then we may estimate ϕ''_{θ_0} by: for $h \in \mathbb{D}$,

$$\hat{\phi}''_n(h) = \frac{\phi(\hat{\theta}_n + t_n h) - \phi(\hat{\theta}_n) - t_n \hat{\phi}'_n(h)}{t_n^2} \ , \tag{1.37}$$

where $\hat{\phi}'_n(h)$ is given by:

$$\hat{\phi}'_n(h) = \frac{\phi(\hat{\theta}_n + s_n h) - \phi(\hat{\theta}_n)}{s_n} \ , \tag{1.38}$$

and $\{t_n, s_n\}$ are tuning parameters that tend to zero. We emphasize that s_n should not be taken to be equal to t_n because otherwise we have $\hat{\phi}''_n(h) = 0$ numerically for all $h \in \mathbb{D}$. In fact, in order for $\hat{\phi}'_n$ and $\hat{\phi}''_n$ to possess desired estimation properties, we need put restrictions on the rate at which s_n, t_n approach zero. ■

1.3.5.1 Examples Revisited

We now demonstrate how to exploit the structure of the derivative for the purpose of derivative estimation. Examples 1.2.1 is trivial since ϕ''_{θ_0} is a known map and hence one can simply set $\hat{\phi}''_n = \phi''_{\theta_0}$.

Example 1.2.4 (Continued). Let $\hat{B}_0(\theta_0)$ be an estimator of $B_0(\theta_0)$. Then we may estimate ϕ''_{θ_0} by

$$\hat{\phi}''_n(h) = \int_{\hat{B}_0(\theta)} \max\{h^{(1)}(u) - h^{(2)}(u), 0\}^2 w(u) du \ . \tag{1.39}$$

It is a simple exercise to verify that Assumption 1.3.4 is satisfied provided

$$\int_{\mathbf{R}} 1\{u \in \hat{B}_0(\theta_0) \triangle B_0(\theta_0)\} w(u) du \xrightarrow{p} 0 \ , \tag{1.40}$$

where $A \triangle B$ denotes the set difference between sets A and B. Such a construction corresponds to the bootstrap procedure studied in Linton et al. (2010). ■

Example 1.2.6 (Continued). In the classical case when $\Gamma_0(\theta)$ is singleton, we may esti-

mate ϕ''_{θ_0} based on the GMM estimator $\hat{\gamma}_n$ and the estimated Jacobian matrix $\hat{J}_n$. Generally, there are two unknown objects involved in the second order derivative: the identified set $\Gamma_0(\theta)$ and $J(\gamma_0)$. Let $\mathbf{M}^{m\times k}$ be the space of $m \times k$ matrices. Suppose that $\hat{\Gamma}_n \subset \Gamma$ is a d_H-consistent estimator for $\Gamma_0(\theta)$, and $\hat{J}_n : \Gamma \to \mathbf{M}^{m\times k}$ an estimator for $J : \Gamma \to \mathbf{M}^{m\times k}$ such that $\sup_{\gamma\in\Gamma} \|\hat{J}_n(\gamma) - J(\gamma)\| \xrightarrow{p} 0$. Then we may estimate ϕ''_{θ_0} by

$$\hat{\phi}''_n(h) = \min_{\gamma_0\in\hat{\Gamma}_n} \min_{v\in B_n} \{h(\gamma_0) - \hat{J}_n(\gamma_0)v\}^\intercal W \{h(\gamma_0) - \hat{J}_n(\gamma_0)v\} \ , \tag{1.41}$$

where $B_n \equiv \{v \in \mathbf{R}^k : \|v\| \leq t_n^{-1}\}$ for $t_n \downarrow 0$ satisfying $t_n\sqrt{n} \to \infty$. Consistency of $\hat{\Gamma}_n$ can be established by appealing to Chernozhukov et al. (2007), while uniform consistency of $\hat{J}_n$ can be derived using Glivenko-Cantelli type arguments. Following the proof of Lemma 1.8.9, it is straightforward to show that $\hat{\phi}''_n$ satisfies Assumption 1.3.4. ■

1.4 Hypothesis Testing

Resampling methods such as bootstrap have many powerful applications in statistical analysis. For instance, jackknife and bootstrap initially were intended primarily as tools for bias reduction and variance estimation (Efron, 1979). If, however, ϕ is nondifferentiable, biases can not be fully eliminated and bias reduction can cause large variances (Hirano and Porter, 2012; Fang, 2016). In this section, we instead study the hypothesis

$$\mathrm{H}_0 : \phi(\theta_0) = 0 \qquad \mathrm{H}_1 : \phi(\theta_0) > 0 \ . \tag{1.42}$$

Under first order degeneracy, as is the case in all our examples, we propose using $r_n^2\phi(\hat{\theta}_n)$ as the test statistic, which, according to Theorems 1.3.2 and 1.3.3, provides pointwise size control by rejecting H_0 if $r_n^2\phi(\hat{\theta}_n) > \hat{c}_{1-\alpha}$ where $\hat{c}_{1-\alpha}$ is the critical value constructed from the Babu correction or our proposed bootstrap, i.e.,

$$\hat{c}_{1-\alpha} = \inf\{c \in \mathbf{R} : P_W(r_n^2\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n) - \phi'_{\hat{\theta}_n}(\hat{\theta}_n^* - \hat{\theta}_n)\} \leq c) \geq 1 - \alpha\} \ , \tag{1.43}$$

or

$$\hat{c}_{1-\alpha} = \inf\{c \in \mathbf{R} : P_W(\hat{\phi}''_n(r_n\{\hat{\theta}^*_n - \hat{\theta}_n\}) \leq c) \geq 1-\alpha\} \ . \tag{1.44}$$

Note that $\hat{c}_{1-\alpha}$ is generally infeasible in that it is constructed based on the "exact" distribution of $\hat{\phi}''_n(r_n\{\hat{\theta}^*_n - \hat{\theta}_n\})$ conditional on the data. Nonetheless, it can be estimated by Monte Carlo simulation which in turn invites additional random error but can be made presumably arbitrarily small by choosing the number of bootstrap samples (Efron, 1979; Hall, 1992; Horowitz, 2001). Thus, as standard in the bootstrap literature, we think of $\hat{c}_{1-\alpha}$ as known in what follows.

In fact, under additional restrictions, our test can provide local size control. This property is particularly attractive because of the irregularity arising from nondifferentiability of ϕ. In this case, pointwise asymptotic approximations can be misleading (Imbens and Manski, 2004; Andrews and Guggenberger, 2009a). Interestingly, it turns out that there is another source of irregularity due to second order asymptotics (see Lemma 1.4.1). We next proceed to investigate the behaviors of our procedure under local perturbations to the underlying distribution of the data, as characterized in next subsection.

1.4.1 Local Perturbations

We first introduce relevant concepts following Bickel et al. (1998). In what follows we specialize our setup to the the i.i.d. setting for simplicity.[8] In particular, the data $\{X_i\}_{i=1}^n$ is presumed to have a common probability measure $P \in \mathcal{P}$, where $\mathcal{P}$ is a collection of Borel probability measures that possibly generate the data. Further, we think of the parameter θ_0 as a map $\theta : \mathcal{P} \to \mathbb{D}_\phi$, i.e., $\theta_0 = \theta(P)$. Formally, we impose the following:

Assumption 1.4.1. *(i) $\{X_i\}_{i=1}^n$ is an i.i.d. sequence with each $X_i \in \mathbf{R}^{d_x}$ distributed according to $P \in \mathcal{P}$; (ii) $\theta_0 \equiv \theta(P)$ for some known map $\theta : \mathcal{P} \to \mathbb{D}_\phi$ and $\phi(\theta_0) = 0$.*

[8]Generally, we may consider models that are locally asymptotically quadratic (van der Vaart, 1998; Ploberger and Phillips, 2012).

Given the model $\mathcal{P}$ defined in Assumption 1.4.1, we now formalize the notion of local perturbations to the true probability measure P. Intuitively, a local perturbation can be thought as a sequence of probability measures contained in $\mathcal{P}$ that approaches P. Since the set of probability measures is not a vector space, an appropriate embedding is needed to make precise sense of this idea. This is simplified by considering one dimensional parametric models containing P and contained in $\mathcal{P}$ (Stein, 1956).

Definition 1.4.1. A function $t \mapsto P_t$ mapping a neighborhood $(-\epsilon, \epsilon)$ of zero into $\mathcal{P}$ is called a differentiable path passing through P if $P_0 = P$ and for some $h : \mathbf{R}^{d_x} \to \mathbf{R}$,

$$\lim_{t \to 0} \int \left[\frac{dP_t^{1/2} - dP^{1/2}}{t} - \frac{1}{2} h \, dP^{1/2} \right]^2 = 0 \ . \tag{1.45}$$

Intuitively, a differentiable path is just a parametric model in $\mathcal{P}$ and indexed by $t \in (-\epsilon, \epsilon)$ such that it is getting close to P sufficiently fast as $t \to 0$. The function h is referred to as the score function of P and satisfies $\int h \, dP = 0$ and $h \in L^2(P)$.

The perturbations on P are fundamental in that they affect everything that is built on the model, which in particular includes the parameter $\theta : \mathcal{P} \to \mathbb{D}_\phi$ and the estimator $\hat{\theta}_n : \{X_i\}_{i=1}^n \to \mathbb{D}_\phi$. In this paper, we shall only consider θ and $\hat{\theta}_n$ that are well behaved with respect to these local perturbations. This is formalized by the following assumption.

Assumption 1.4.2. *(i) For every differentiable path $\{P_t\}$ in $\mathcal{P}$ with score function h, $\theta : \mathcal{P} \to \mathbb{D}_\phi$ is regular in the sense that there exists $\theta_0'(h) \in \mathbb{D}_0$ such that $\|\theta(P_t) - \theta(P) - t\theta_0'(h)\|_{\mathbb{D}} = o(t)$ (as $t \to 0$); (ii) $\hat{\theta}_n$ is a regular estimator for $\theta(P)$.*[9]

Assumption 1.4.2(i) is a smoothness condition on the parameter $\theta : \mathcal{P} \to \mathbb{D}_\phi$, while Assumption 1.4.2(ii) means that $\hat{\theta}_n$ is asymptotically invariant to local perturbations. Assumption 1.4.2(i) and (ii) in fact are closely related, though themselves alone do not imply one another. In particular, regularity of $\hat{\theta}_n$ plus a mild condition implies regularity of $\theta : \mathcal{P} \to \mathbb{D}_\phi$, and vice versa (van der Vaart, 1991; Hirano and Porter, 2012).

[9]Formally, $\hat{\theta}_n$ is a regular estimator if for every differentiable path $\{P_t\}$ in $\mathcal{P}$ with score function h, we have $r_n\{\hat{\theta}_n - \theta(P_n)\} \xrightarrow{L_n} \mathbb{G}$, where $P_n \equiv P_{1/r_n}$ and L_n denotes the law under $\prod_{i=1}^n P_n$.

Given the above regularity conditions, we now proceed to characterize local behaviors of our test statistic.

Lemma 1.4.1. *Let $\{P_t\}$ be a differentiable path with score function h. Suppose that Assumptions 1.2.1, 1.2.2, 1.2.3(ii)(iii), 1.4.1 and 1.4.2 hold. Then,*

$$r_n^2 \phi(\hat{\theta}_n) \xrightarrow{L_n} \phi''_{\theta_0}(\mathbb{G} + \theta'_0(h)) \ , \tag{1.46}$$

where L_n denotes the law under $\prod_{i=1}^n P_n$ with $P_n \equiv P_{1/r_n}$ by abuse of notation.

Lemma 1.4.1 indicates that the asymptotic distribution of $r_n^2 \phi(\hat{\theta}_n)$ varies as a function of the score h, and in this sense exhibits second order irregularity, even if the map ϕ is both first and second order differentiable and $\hat{\theta}_n$ is regular. This is perhaps surprising *ex ante* and yet somewhat expected *ex post*. One important implication of Lemma 1.4.1 is that one should carefully evaluate how sensitive the statistical procedures under consideration is, in the presence of first order degeneracy.

1.4.2 Local Size and Power

Having derived the asymptotic distributions of $r_n^2 \phi(\hat{\theta}_n)$ under local perturbations, we are now in a position to establish local power performance and local size control of our test. We consider differentiable paths $\{P_t\}$ in $\mathcal{P}$ that also belong to the set

$$\mathcal{H} \equiv \{\{P_t\} : \text{ (i) } \phi(\theta(P_t)) = 0 \text{ if } t \leq 0, \text{ and (ii) } \phi(\theta(P_t)) > 0 \text{ if } t > 0\} \ .$$

Thus, a path $\{P_t\} \in \mathcal{H}$ is such that $\{P_t\}$ satisfies the null hypothesis whenever $t \leq 0$, but switches to satisfying the alternative hypothesis at all $t > 0$. One can think of $\mathcal{H}$ as a simple device to study local size and power in a compact way. Further, we denote the power function at sample size n for the test that rejects whenever $r_n^2 \phi(\hat{\theta}_n) > \hat{c}_{1-\alpha}$ by

$$\pi_n(P_{\eta/r_n}) \equiv P_n^n(r_n^2 \phi(\hat{\theta}_n) > \hat{c}_{1-\alpha}) \ ,$$

where we write $P_n \equiv P_{\eta/r_n}$ and $P_n^n \equiv \prod_{i=1}^n P_n$. The following additional assumption ensures local size control of our test.

Assumption 1.4.3. *(i)* $\mathbb{E} = \mathbf{R}$*; (ii) The cdf of* $\phi''_{\theta_0}(\mathbb{G})$ *is strictly increasing and continuous at its* $(1-\alpha)$*-th quantile* $c_{1-\alpha}$*; (iii) There exists a strictly increasing function* $\tau : \phi''_{\theta_0}(\mathbb{D}_0) \to \mathbf{R}$ *such that* $\tau(0) = 0$ *and* $\tau \circ \phi''_{\theta_0} : \mathbb{D}_0 \to \mathbf{R}$ *is subadditive.*

Assumption 1.4.3(i) formalizes the requirement that ϕ be scalar valued. Assumption 1.4.3(ii) requires strict monotonicity of the cdf of $\phi''_{\theta_0}(\mathbb{G})$ at $c_{1-\alpha}$ which ensures consistency of the critical value $\hat{c}_{1-\alpha}$, and continuity which ensures the test controls size at least pointwise in P. Subadditivity of $\tau \circ \phi''_{\theta_0}$ as required in Assumption 1.4.3(iii) is crucial for establishing local size control of our test. This condition was imposed directly on the first order derivative in Fang and Santos (2015). In our setup, ϕ''_{θ_0} itself often violates subadditivity because it is closely related to quadratic forms. Nonetheless, in all but Example 1.2.6, $\tau \circ \phi''_{\theta_0}$ is subadditive for $\tau : \mathbf{R}^+ \to \mathbf{R}^+$ given by $\tau(\nu) = \sqrt{\nu}$.[10]

The following theorem derives the asymptotic limits of the power function $\pi_n(P_{\eta/r_n})$.

Theorem 1.4.1. *Let Assumptions 1.2.1, 1.2.2, 1.2.3, 1.3.1, 1.3.2, 1.3.4, 1.4.1, 1.4.2 and 1.4.3(i)-(ii) hold. It then follows that for any differentiable path* $\{P_t\}$ *in* $\mathcal{H}$ *with score function* h*, and every* $\eta \in \mathbf{R}$ *we have*

$$\liminf_{n\to\infty} \pi_n(P_{\eta/r_n}) \geq P(\phi''_{\theta_0}(\mathbb{G} + \theta'_0(\eta h)) > c_{1-\alpha}) \ . \tag{1.47}$$

If in addition Assumption 1.4.3(iii) also holds, then we can conclude that for any $\eta \leq 0$

$$\limsup_{n\to\infty} \pi_n(P_{\eta/r_n}) \leq \alpha \ . \tag{1.48}$$

The first claim of the theorem establishes a lower bound for the power function under local perturbations to the null which includes in particular local alternatives. In fact,

[10]For Example 1.2.6, it turns out that $\sqrt{\phi''_{\theta_0}(\cdot)}$ is subadditive when γ_0 is point identified, though the main motivation for us being general there is to accommodate partial identification as well as the Jacobian matrix being degenerate.

the lower bound is sharp whenever $c_{1-\alpha}$ is a continuity point of the cdf of $\phi''_{\theta_0}(\mathbb{G} + \eta\theta'_0(h))$, in which case (1.47) holds with equality. The role of Assumption 1.4.3(iii) can be seen from (1.47) and the inequalities

$$\begin{aligned} P(\phi''_{\theta_0}(\mathbb{G} + \eta\theta'_0(h)) > c_{1-\alpha}) &= P(\tau \circ \phi''_{\theta_0}(\mathbb{G} + \theta'_0(\eta h)) > \tau(c_{1-\alpha})) \\ &\leq P(\tau \circ \phi''_{\theta_0}(\mathbb{G}_0) + \tau \circ \phi''_{\theta_0}(\theta'_0(\eta h)) > \tau(c_{1-\alpha})) \\ &= P(\tau \circ \phi''_{\theta_0}(\mathbb{G}_0) > \tau(c_{1-\alpha})) \\ &= P(\phi''_{\theta_0}(\mathbb{G}_0) > c_{1-\alpha}) \leq \alpha \ , \end{aligned}$$

where the second equality is due to $\phi''_{\theta_0}(\theta'_0(\eta h)) = 0$ and $\tau(0) = 0$.[11]

To conclude this section, we note that it is possible to develop a testing procedure adaptive to potential first order degeneracy, that is, in settings where ϕ is not always first order degenerate under the null. We emphasize that $r_n^2\phi(\hat{\theta}_n)$ fails to be a valid statistic since it diverges to infinity at those nondegenerate points, and so does

$$r_n^2\{\phi(\hat{\theta}_n) - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)\} \ ,$$

because θ_0 might not be identified given $\phi(\theta_0) = 0$. By introducing an appropriate selection rule, we can combine first and second order asymptotics to provide a more general testing procedure; see Remark 1.4.1. Development of adaptiveness not only serves to maintain generality of our theory, but also is necessary when constructing confidence sets for $\phi(\theta_0)$; see Remark 1.4.2.

Remark 1.4.1. If ϕ'_{θ_0} is only degenerate at some but not all points under the null, then one may employ the statistic

$$T_n \equiv r_n\phi(\hat{\theta}_n) \cdot 1\{\frac{r_n\phi(\hat{\theta}_n)}{\kappa_n} > 1\} + r_n^2\phi(\hat{\theta}_n) \cdot 1\{\frac{r_n\phi(\hat{\theta}_n)}{\kappa_n} \leq 1\} \ ,$$

where $\kappa_n \downarrow 0$ satisfying $\kappa_n r_n \to \infty$ as $n \to \infty$. Heuristically, if ϕ'_{θ_0} is nondegenerate,

[11] This is because $\phi''_{\theta_0}(\eta\theta'_0(h)) = \lim_{n\to\infty} n\{\phi(\theta(P_n)) - \phi(\theta(P))\} = 0$ by Assumption 1.2.1 and $\{P_n\}$ being a local perturbation under the null.

then $r_n\phi(\hat{\theta}_n)/\kappa_n = O_p(1)/o_p(1) \xrightarrow{p} \infty$ and thus with probability approaching one $T_n = r_n\phi(\hat{\theta}_n)$ which has nondegenerate weak limit $\phi'_{\theta_0}(\mathbb{G})$. If ϕ'_{θ_0} is degenerate, then $r_n\phi(\hat{\theta}_n)/\kappa_n = r_n^2\phi(\hat{\theta}_n)/\kappa_n r_n = O_p(1)/\kappa_n r_n \xrightarrow{p} 0$ and therefore with probability approaching one $T_n = r_n^2\phi(\hat{\theta}_n)$ which has nondegenerate weak limit $\phi''_{\theta_0}(\mathbb{G})$. Accordingly we may construct the corresponding critical value as

$$\hat{c}^*_{1-\alpha} \equiv \tilde{c}_{1-\alpha} \cdot 1\{\frac{r_n\phi(\hat{\theta}_n)}{\kappa_n} > 1\} + \hat{c}_{1-\alpha} \cdot 1\{\frac{r_n\phi(\hat{\theta}_n)}{\kappa_n} \leq 1\} \ , \tag{1.49}$$

where for $\alpha \in (0,1)$ and some estimator $\hat{\phi}'_n$ of ϕ'_{θ_0},

$$\tilde{c}_{1-\alpha} \equiv \inf\{c \in \mathbf{R} : P_W(\hat{\phi}'_n(r_n\{\hat{\theta}^*_n - \hat{\theta}_n\}) \leq c) \geq 1-\alpha\} \ .$$

The indicator functions above serve as a rule for selecting proper statistics based on degeneracy of (a finite sample analogue of) ϕ'_{θ_0}. ■

Remark 1.4.2. Confidence regions for $\nu_0 \equiv \phi(\theta_0) \in \mathbb{E}$ can be constructed by test inversion based on the statistic

$$T_n(\nu_0) \equiv r_n\psi(\hat{\theta}_n) \cdot 1\{\frac{r_n\psi(\hat{\theta}_n)}{\kappa_n} > 1\} + r_n^2\psi(\hat{\theta}_n) \cdot 1\{\frac{r_n\psi(\hat{\theta}_n)}{\kappa_n} \leq 1\} \ , \tag{1.50}$$

where $\psi : \mathbb{D}_\phi \to \mathbf{R}$ is given by $\psi(\theta) \equiv \|\phi(\theta) - \nu_0\|_{\mathbb{E}}$. Critical values can be constructed in a similar fashion as in Remark 1.4.1. By the chain rule (Shapiro, 1990, Proposition 3.6), it is straightforward to see that $\psi'_{\theta_0} = \|\phi'_{\theta_0}\|_{\mathbb{E}}$ and so $\phi'_{\theta_0} = 0$ if and only if $\psi'_{\theta_0} = 0$. Moreover, $\psi''_{\theta_0} = \|\phi''_{\theta_0}\|_{\mathbb{E}}$ when $\psi'_{\theta_0} = 0$. In general, confidence regions thus constructed are less conservative than the plug-in type confidence regions $\phi(\mathcal{C}_{n,\theta})$ with $\mathcal{C}_{n,\theta}$ some level $1-\alpha$ confidence region for θ_0. ■

1.5 Application: Testing for Common CH Features

In this section, we apply our framework to develop a robust test of common conditionally heteroskedastic (CH) factor structure by allowing partial identification. Let $\{Y_t\}_{t=1}^T$

be a k-dimensional time series. According to Engle and Kozicki (1993), a feature that is present in each component of Y_t is said to be common to Y_t if there exists a linear combination of Y_t that fails to have the feature.[12] A canonical example is the notion of cointegration developed by Engle and Granger (1987) in order to characterize the common feature of stochastic trend.

1.5.1 The Setup

Following Engle et al. (1990) and Dovonon and Renault (2013), suppose that the k-dimensional process $\{Y_t\}$ satisfies

$$\mathrm{Var}(Y_{t+1}|\mathcal{F}_t) = \Lambda D_t \Lambda^\intercal + \Omega \ , \tag{1.51}$$

where Λ is a $k \times p$ matrix of full column rank with $p < k$, D_t a $p \times p$ diagonal matrix with diagonal (random) elements σ_{jt}^2 for $j = 1, \ldots, p$, Ω a $k \times k$ positive semidefinite matrix, and $\{\mathcal{F}_t\}_{t=1}^{\infty}$ a filtration to which $\{Y_t\}_{t=1}^{\infty}$ and $\{\sigma_{jt}^2 : j = 1, \ldots, p\}_{t=1}^{\infty}$ are adapted. By Engle and Kozicki (1993), we say that $\{Y_t\}$ has a common CH feature if there exists some nonzero $\gamma \in \mathbf{R}^k$ such that $\mathrm{Var}(\gamma^\intercal Y_t|\mathcal{F}_t)$ is constant. The conditional covariance structure (1.51) has some attractive properties that help to understand, for example, asset excess returns in a parsimonious way (Engle et al., 1990). Thus, tests of common CH features can be used to detect the underlying common factor structures that simplify capturing interrelations of economic and financial variables under consideration.

With the help of instrumental variables, a common CH feature can be reformulated by unconditional moments that fit into the classical GMM framework. The following assumption is taken directly from Dovonon and Renault (2013).

Assumption 1.5.1. *(i)* Λ *is of full column rank; (ii)* $\mathrm{Var}(\sigma_t^2)$ *is nonsingular for* $\sigma_t^2 \equiv (\sigma_{1t}^2, \ldots, \sigma_{pt}^2)^\intercal$*; (iii)* $E[Y_{t+1}|\mathcal{F}_t] = 0$*; (iv)* Z_t *is an* $m \times 1$ $\mathcal{F}_t$*-measurable random vector*

[12] A feature has to satisfy three axioms (Engle and Kozicki, 1993): (i) if Y_t has (resp. does not have) the feature, then γY_t will have (resp. not have) the feature for any $\gamma \neq 0$; (ii) if neither X_t nor Y_t have the feature, then $X_t + Y_t$ does not have the feature; (iii) if X_t has the feature but Y_t does not, then $X_t + Y_t$ will have the feature.

such that $\mathrm{Var}(Z_t)$ *is nonsingular; (v)* $\mathrm{Cov}(Z_t, \sigma_t^2)$ *has full column rank p; (vi)* $\{Y_t, Z_t\}$ *is stationary and ergodic such that* $E[\|Z_t\|^2] < \infty$ *and* $E[\|Y_t\|^4] < \infty$.

Assumption 1.5.1(i)-(ii) ensure that there are exactly $k - p$ linearly independent vectors γ, spanning the null space of $\Lambda^\intercal$, such that $\mathrm{Var}(\gamma^\intercal Y_t|\mathcal{F}_t)$ is constant. In other words, the common CH features γ are nonzero solutions of the equation $\Lambda^\intercal\gamma = 0$. Assumption 1.5.1(iii) is a normalization condition that helps to simplify the exposition. Assumption 1.5.1(iv) defines the instrument Z_t formed from the information set $\mathcal{F}_t$, while Assumption 1.5.1(v) implicitly requires that the number of instruments is no less than that of factors. Assumption 1.5.1(vi) further specifies the data generating process. We refer the readers to Dovonon and Renault (2013) for further details of discussions on Assumption 1.5.1.

Assumption 1.5.1 allows us to characterize common CH features as nonzero γ satisfying the vector of unconditional moment equalities (Dovonon and Renault, 2013):

$$E[Z_t\{(\gamma^\intercal Y_{t+1})^2 - c(\gamma)\}] = 0 \ , \tag{1.52}$$

where $c(\gamma) = E[(\gamma^\intercal Y_{t+1})^2]$. It is then tempting to employ Hansen's J statistic to test the existence of common CH features (Engle and Kozicki, 1993). Unfortunately, as noted by Dovonon and Renault (2013), the Jacobian matrix evaluated at the truth is zero, rendering standard theory inapplicable. In fact, with the help of second order analysis, Dovonon and Renault (2013) showed that the asymptotic distribution of the J statistic is highly nonstandard. Nonetheless, Dovonon and Gonçalves (2014) developed a corrected bootstrap that can consistently estimate the limiting law when the bootstrap of Hall and Horowitz (1996) fails to do so.

However, a key assumption in previous studies on test of common CH features is that there exists a unique nonzero γ such that (1.52) is satisfied, ensured by normalization (Dovonon and Renault, 2013; Dovonon and Gonçalves, 2014; Lee and Liao, 2014). This is undesirable for the following reasons. First, it is unknown *a priori* how many CH features are common to the series under consideration. Second, as pointed out by Engle et al. (1990)

in the context of asset pricing, empirical work often considers large numbers of assets and the numbers of common CH features are expected to be large as well. Third, exclusion restrictions and normalization condition that are intended to ensure uniqueness of γ as in Dovonon and Renault (2013) might in fact lead to no γ satisfying (1.52) and hence misleading conclusions. For example, suppose that $k = 2$ and $\Lambda = [1, 1]^\intercal$. Then by Lemma 2.1 in Dovonon and Renault (2013), any common CH feature γ must satisfy $\gamma^{(1)} + \gamma^{(2)} = 0$, contradicting the linear normalization $\gamma^{(1)} + \gamma^{(2)} = 1$ proposed in Dovonon and Renault (2013). These arguments motivate us to modify the J-test in a way that accommodates partial identification as well as degenerate Jacobian matrices.

1.5.2 A Modified J Test

To exclude zero solution, we employ the following normalization

$$\gamma \in \mathbb{S}^k \equiv \{\gamma' \in \mathbf{R}^k : \|\gamma'\| = 1\} \ . \tag{1.53}$$

Note that if γ is a common feature, so is $-\gamma$. Thus, under normalization (1.53), the set of common CH features is never a singleton. Next, to map the current setup into our developed framework, define a function $\phi : \prod_{j=1}^m \ell^\infty(\mathbb{S}^k) \to \mathbf{R}$ by

$$\phi(\theta) \equiv \inf_{\gamma \in \mathbb{S}^k} \|\theta(\gamma)\|^2 \ . \tag{1.54}$$

Then in view of the moment conditions (1.52), the hypothesis that there exists at least one common CH feature can be reformulated as

$$\mathrm{H}_0 : \phi(\theta_0) = 0 \qquad \mathrm{H}_1 : \phi(\theta_0) > 0 \ , \tag{1.55}$$

where $\theta_0 : \mathbb{S}^k \to \mathbf{R}^m$ is defined as $\theta_0(\gamma) \equiv E[Z_t\{(\gamma^\intercal Y_{t+1})^2 - c(\gamma)\}]$. In this formulation, we have taken the identity matrix I_m as the weighting matrix not only for simplicity, but more importantly because, as pointed out by Dovonon and Renault (2013), the rate of

convergence of the GMM estimator varies as the weighting matrix changes and hence the conventional notion of optimal weighting does not make sense.

As expected, under the null ϕ is Hadamard differentiable with degenerate derivative, and second order Hadamard directionally differentiable at θ_0 tangentially to $\prod_{j=1}^m C(\mathbb{S}^k)$ with the derivative given by: for any $h \in \prod_{j=1}^m C(\mathbb{S}^k)$,

$$\phi''_{\theta_0}(h) = \min_{\gamma \in \Gamma_0} \min_{v \in \mathbf{R}^k} \|h(\gamma) + G \operatorname{vec}(vv^\intercal)\|^2 , \tag{1.56}$$

where $\Gamma_0 = \{\gamma \in \mathbb{S}^k : \theta_0(\gamma) = 0\}$ is the identified set of γ, and $G \in \mathbf{M}^{m \times k^2}$ with the jth row given by $\operatorname{vec}(\Delta_j)^\intercal$ and

$$\Delta_j = E[Z_t^{(j)}(Y_{t+1}Y_{t+1}^\intercal - E[Y_{t+1}Y_{t+1}^\intercal])] .$$

Note that partial identification of γ invites irregularity through the first minimization problem in (1.56), in addition to the irregularity caused by the second minimization having multiple minimizers.

Next, let $\hat{\theta}_T : \mathbb{S}^k \to \mathbf{R}^m$ be defined by $\hat{\theta}_T(\gamma) = \frac{1}{T}\sum_{t=1}^T Z_t\{(\gamma^\intercal Y_{t+1})^2 - \hat{c}(\gamma)\}$ with $\hat{c}(\gamma) = \frac{1}{T}\sum_{t=1}^T (\gamma^\intercal Y_{t+1})^2$. Given the established differentiability of ϕ, the asymptotic distribution of $\phi(\hat{\theta}_T)$ is then an immediate consequence of Theorem 1.2.1 provided $\hat{\theta}_T$ converges weakly. Towards this end, we impose the following assumption as in Dovonon and Renault (2013).

Assumption 1.5.2. *Z_t, $\operatorname{vec}(Y_tY_t^\intercal)$ and $\operatorname{vec}(Y_tY_t^\intercal) \otimes Z_t$ fulfill CLT.*[13]

Assumptions 1.5.1 and 1.5.2 together imply that

$$\sqrt{T}\{\hat{\theta}_T - \theta_0\} \xrightarrow{L} \mathbb{G} \text{ in } \prod_{j=1}^m \ell^\infty(\mathbb{S}^k) , \tag{1.57}$$

where $\mathbb{G}$ is a zero mean Gaussian process with the covariance functional satisfying: for any

[13]The symbol $\otimes$ denotes Kronecker product.

γ_1, $\gamma_2 \in \Gamma_0$ and $\mu_z \equiv E[Z_t]$,

$$E[\mathbb{G}(\gamma_1)\mathbb{G}(\gamma_2)] = E[(Z_t - \mu_z)(Z_t - \mu_z)^\intercal \{(\gamma_1^\intercal Y_{t+1})^2 - c(\gamma_1)\}\{(\gamma_2^\intercal Y_{t+1})^2 - c(\gamma_2)\}] \ .$$

The proposition below delivers the limiting distribution of test statistic $T\phi(\hat{\theta}_T)$.

Proposition 1.5.1. *Let Assumptions 1.5.1 and 1.5.2 hold. Then we have under* H_0

$$T \min_{\gamma \in \mathbb{S}^k} \|\hat{\theta}_T(\gamma)\|^2 \xrightarrow{L} \min_{\gamma \in \Gamma_0} \min_{v \in \mathbf{R}^k} \|\mathbb{G}(\gamma) + G \operatorname{vec}(vv^\intercal)\|^2 \ . \tag{1.58}$$

The asymptotic distribution in (1.58) is a nonlinear functional of the Gaussian process $\mathbb{G}$. As shown by Dovonon and Gonçalves (2014), however, the recentered bootstrap of Hall and Horowitz (1996) fails to consistently estimate this limit. The bootstrap proposed by Dovonon and Gonçalves (2014) is not directly applicable because their identification assumption is violated.

We next demonstrate how our bootstrap works. First, let $\{Y_{t+1}^*, Z_t^*\}_{t=1}^T$ be a bootstrap sample, which can be obtained by block bootstrap, nonoverlapping or overlapping (Carlstein, 1986; Kunsch, 1989). But the limiting process $\{\mathbb{G}(\gamma) : \gamma \in \Gamma_0\}$ is determined by a martingale difference sequence indexed by $\gamma \in \Gamma_0$, the dependence structure of the data does not enter into the limit and we may thus employ Efron (1979)'s nonparametric bootstrap or more general bootstrap schemes. In any case, we set

$$\hat{\theta}_T^*(\gamma) = \frac{1}{T}\sum_{t=1}^T Z_t^* \{(\gamma^\intercal Y_{t+1}^*)^2 - \hat{c}^*(\gamma)\} \ , \ \hat{c}^*(\gamma) = \frac{1}{T}\sum_{t=1}^T (\gamma^\intercal Y_{t+1}^*)^2 \ . \tag{1.59}$$

To accommodate diverse resampling schemes, we simply impose the high level condition that $\hat{\theta}_T^*$ satisfies Assumptions 1.3.1 and 1.3.2 (Dehling et al., 2002).

It remains to estimate the derivative (1.56). The numerical differentiation approach

can be implemented as in the beginning of Section 1.3.5. That is, we estimate ϕ''_{θ_0} by

$$\hat{\phi}''_T(h) = \frac{\inf_{\gamma\in\mathbb{S}^k} \|\hat{\theta}_T(\gamma) + \kappa_T h(\gamma)\|^2 - \min_{\gamma\in\mathbb{S}^k} \|\hat{\theta}_T(\gamma)\|^2}{\kappa_T^2} \ , \tag{1.60}$$

where κ_T satisfies Assumption 1.3.5. We now describe how to estimate ϕ''_{θ_0} by exploiting its structure. Let $B_T = \{v \in \mathbf{R}^k : \|v\| \leq \kappa_T^{-1/4}\}$ where κ_T is to be specified. Then we may estimate $\phi''_{\theta_0}(h)$ by

$$\hat{\phi}''_T(h) = \inf_{\gamma\in\hat{\Gamma}_T} \min_{v\in B_T} \|h(\gamma) + \hat{G}\,\mathrm{vec}(vv^\intercal)\|^2 \ , \tag{1.61}$$

where $\hat{\Gamma}_T = \{\gamma \in \mathbb{S}^k : \|\hat{\theta}_T(\gamma)\|^2 - \phi(\hat{\theta}_T) \leq \kappa_T\}$,[14] and $\hat{G} \in \mathbf{M}^{m\times k^2}$ with its jth row given by $\mathrm{vec}(\hat{\Delta}_j)^\intercal$ for

$$\hat{\Delta}_j = \frac{1}{T}\sum_{t=1}^T Z_t^{(j)} Y_{t+1} Y_{t+1}^\intercal - \frac{1}{T}\sum_{t=1}^T Z_t^{(j)} \frac{1}{T}\sum_{t=1}^T Y_{t+1} Y_{t+1}^\intercal \ .$$

In fact, we may further restrict the bounded set B_T to reduce the computation burden for $\hat{\phi}''_T$; see Remark 1.5.1.

Remark 1.5.1. The derivative (1.56) can be rewritten as:

$$\phi''_{\theta_0}(h) = \min_{\gamma\in\Gamma_0} \min_{v\in\Gamma_0^\perp} \|h(\gamma) + G\,\mathrm{vec}(vv^\intercal)\|^2 \ , \tag{1.62}$$

where $\Gamma_0^\perp \equiv \{\lambda \in \mathbf{R}^k : \lambda^\intercal\gamma = 0 \ , \forall\, \gamma \in \Gamma_0\}$ denotes the orthogonal complement of Γ_0. Then for $\hat{\Gamma}_{T,\perp} = \{\gamma \in \mathbf{R}^k : \sup_{\lambda\in\hat{\Gamma}_T} |\gamma^\intercal\lambda| \leq \kappa_T^{1/4}\}$, we may estimate $\phi''_{\theta_0}(h)$ by

$$\hat{\phi}''_T(h) = \inf_{\gamma\in\hat{\Gamma}_T} \min_{v\in\hat{\Gamma}_{T,\perp}\cap B_T} \|h(\gamma) + \hat{G}\,\mathrm{vec}(vv^\intercal)\|^2 \ . \qquad \blacksquare$$

Clearly, the sequence $\{\kappa_T\}$ should tend to zero at a suitable rate as $T \to \infty$. This is made precise as follows.

Assumption 1.5.3. *$\{\kappa_T\}$ satisfies (i) $\kappa_T \downarrow 0$, and (ii) $\sqrt{T}\kappa_T \to \infty$ if $\hat{\phi}''_T$ is given by* (1.60)

[14] One can theoretically ignore $\phi(\hat{\theta}_T)$ in the expression of $\hat{\Gamma}_T$. As pointed out by Chernozhukov et al. (2007), however, such a modification helps avoid an empty set of solutions and improve power.

or $T\kappa_T \to \infty$ *if* $\hat{\phi}''_T$ *is given by* (1.61).

Combining the bootstrap $\hat{\theta}^*_T$ in (1.59) and the derivative estimator, we are then able to consistently estimate the law of the weak limit in (1.58) following Theorem 1.3.3, which in turn allows us to construct critical values. Specifically, let $\hat{c}_{1-\alpha}$ be the $1-\alpha$ quantile of $\hat{\phi}''_T(\sqrt{T}\{\hat{\theta}^*_T - \hat{\theta}_T\})$ conditional on the data:[15]

$$\hat{c}_{1-\alpha} \equiv \inf\{c \in \mathbf{R} : P_W(\hat{\phi}''_T(\sqrt{T}\{\hat{\theta}^*_T - \hat{\theta}_T\}) \leq c) \geq 1-\alpha\} \ . \tag{1.63}$$

The following proposition confirms that the test of rejecting existence of common CH features when $T\phi(\hat{\theta}_T) > \hat{c}_{1-\alpha}$ is valid.

Proposition 1.5.2. *Suppose Assumptions 1.3.1, 1.3.2, 1.5.1, 1.5.2, and 1.5.3 hold. If the cdf of the limit in* (1.58) *is continuous and strictly increasing at its* $1-\alpha$ *quantile for* $\alpha \in (0,1)$*, then we have under* H_0*,*

$$\lim_{T\to\infty} P(T \min_{\gamma \in \mathbb{S}^k} \|\hat{\theta}_T(\gamma)\|^2 > \hat{c}_{1-\alpha}) = \alpha \ .$$

Proposition 1.5.2 implies our test has pointwise asymptotic exact size α and thus is not conservative (in the pointwise sense). Establishing local size control, unfortunately, is challenging in this case, because ϕ''_{θ_0} fails to be subadditive in general when there exist more than one common CH features. In fact, the problem of developing (at least) locally valid *and* non-conservative overidentification tests is prevalent in the literature of partial identification (Chernozhukov et al., 2007; Andrews and Soares, 2010).

1.5.3 Simulation Studies

In this section, we examine the finite sample performance of our inference framework based on Monte Carlo simulations and show how the identification assumption in Dovonon

[15] As usual, P_W denotes the probability taken with respect to the bootstrap weights $\{W_T\}$, though in the current setup they are implicitly defined. Alternatively, one can think of P_W as the probability with respect to the bootstrap sample $\{Z^*_t, Y^*_{t+1}\}$ holding data fixed.

and Renault (2013) and Dovonon and Gonçalves (2014) can suffer from their imposed linear normalization.

As in Dovonon and Renault (2013) and Dovonon and Gonçalves (2014), we consider the following CH factor model:

$$Y_t = \Lambda F_t + u_t \ , \tag{1.64}$$

where Y_t is a $k \times 1$ vector that can be thought of asset returns, F_t is a $p \times 1$ vector of CH factors, Λ is a $k \times p$ matrix of factor loadings, and u_t is a vector of idiosyncratic shocks independent of F_t. The model (1.64) is essentially the factor model that underpins the arbitrage pricing theory (Ross, 1976).

Following Dovonon and Renault (2013) and Dovonon and Gonçalves (2014), we let $\{U_t\}$ be an i.i.d. sequence from $N(0, I_k/2)$, and the jth component $f_{j,t+1}$ of F_{t+1} follow a Gaussian-GARCH(1,1) model such that

$$f_{j,t+1} = \sigma_{j,t}\epsilon_{j,t+1} \ , \ \sigma^2_{j,t} = \omega_j + \alpha_j f^2_{j,t} + \beta_j \sigma^2_{j,t-1} \ ,$$

where $\omega_j, \alpha_j, \beta_j > 0$, $\{\epsilon_{k,t}\} \sim N(0,1)$ i.i.d. across both i and t, and $\{\sigma_{j0}\}$ are independent across j and of $\{\epsilon_{k,t}\}$. It follows that $\{f_{j,t}\}$ are independent across j for each t. The remaining specifications are detailed in Table 1.1. Our designs are the same as those in Dovonon and Renault (2013) and Dovonon and Gonçalves (2014) except that different values for Λ are used to illustrate the restrictiveness of the linear normalization. Designs D1 and D2 generate two assets while Designs D3, D4 and D5 generate three assets. In Designs D1, D3 and D4, the factor loading matrices Λ ensure the existence of common CH features and thus serves for investigation of size performance, while no common CH features exist in Designs D2 and D5, which help us to inspect power performance.

The tests are implemented with $m = 2$ and instruments $Z_t = (Y^2_{1,t}, Y^2_{2,t})^\intercal$ for Designs D1 and D2, and with $m = 3$ and $Z_t = (Y^2_{1,t}, Y^2_{2,t}, Y^2_{3,t})^\intercal$ for Designs D3, D4 and D5. For derivative estimation, we set the tuning parameters $\kappa_T = T^{-1/2}, T^{-2/3}, T^{-4/5}$ and

Table 1.1: Simulation Designs

Design	# of Assets	# of Factors	GARCH Parameters	Factor Loadings
D1	$k=2$	$p=1$	$(\omega_1,\alpha_1,\beta_1)=(0.2,0.2,0.6)$	$\Lambda=(1,1)^\intercal$
D2	$k=2$	$p=2$	$(\omega_1,\alpha_1,\beta_1)=(0.2,0.2,0.6)$ $(\omega_2,\alpha_2,\beta_2)=(0.2,0.4,0.4)$	$\Lambda=I_2$
D3	$k=3$	$p=1$	$(\omega_1,\alpha_1,\beta_1)=(0.2,0.2,0.6)$	$\Lambda=(1,1,1)^\intercal$
D4	$k=3$	$p=2$	$(\omega_1,\alpha_1,\beta_1)=(0.2,0.2,0.6)$ $(\omega_2,\alpha_2,\beta_2)=(0.2,0.4,0.4)$	$\Lambda=\begin{bmatrix}1 & 1 & 1\\ -1 & 0 & 1\end{bmatrix}^\intercal$
D5	$k=3$	$p=3$	$(\omega_1,\alpha_1,\beta_1)=(0.2,0.2,0.6)$ $(\omega_2,\alpha_2,\beta_2)=(0.2,0.4,0.4)$ $(\omega_3,\alpha_3,\beta_3)=(0.1,0.1,0.8)$	$\Lambda=I_3$

$\kappa_T = T^{-1/4}, T^{-1/3}, T^{-2/5}$ for the derivative estimator in (1.61) and the numerical derivative estimator as in (1.60) respectively. The corresponding results are denoted as CF1 and CF2, respectively. To show the restrictiveness of the linear normalization $\gamma \in \{\gamma' \in \mathbf{R}^k : \sum_{i=1}^k \gamma_i' = 1\}$ as in Dovonon and Renault (2013), Dovonon and Gonçalves (2014) and Lee and Liao (2014), we report the results based on Dovonon and Gonçalves (2014)'s corrected and continuously-corrected bootstrap, which are denoted as DG1 and DG2 respectively. The sample sizes are $T = 1,000$, $2,000$, $5,000$, $10,000$, $20,000$, $40,000$ and $50,000$ following Dovonon and Gonçalves (2014). To minimize the initial value effect, the data are obtained by generating $T + 100$ samples and dropping the first 100 samples. The results are based on $2,000$ Monte Carlo replications with 200 nonparametric bootstrap replications for each Monte Carlo. The nominal level is 5% throughout.

The results are summarized in Tables 1.2-1.5. Not surprisingly, Dovonon and Gonçalves (2014)'s resampling methods exhibit substantial size distortion, often times close to or over 50%, as shown by the columns labeled DG1 and DG2 in Tables 1.2, 1.3 and 1.4. This does not appear to be a finite sample issue as the distortion is especially severe in large samples. Rather, it is because the linear normalization excludes common CH features that actually exists in the data and in this way leads to wrong conclusions. Our tests

considerably reduce the null rejection rates for all the chosen tuning parameters, though both CF1 and CF2 exhibit some degrees of over- and under-rejection, due to the issue of tuning parameters. Tables 1.5 indicates that our tests are consistent, though CF2 seems to be superior than CF1 in small samples. Another interesting finding is that our bootstrap based on numerical differentiation appears to be more sensitive to the choice of tuning parameters, which is somewhat expected because the structural method (CF1) exploits more information of the derivative. We leave a thorough comparison between these two methods of derivative estimation for future study.

1.6 Conclusion

In this paper, we developed a general statistical framework for conducting inference on functionals exhibiting first order degeneracy, i.e., the first order derivative of the parameter is zero. Our first contribution implies that the standard bootstrap necessarily fails to work in these settings. In light of this failure, we provided two general solutions: one generalizes the Babu correction, and the other one is a modified bootstrap following Fang and Santos (2015). Our framework includes many existing results as special cases. To further demonstrate the applicability of our theory, we developed a test of common CH features studied by Dovonon and Renault (2013) but under weaker assumptions that allow the existence of more than one common CH features.

1.7 Acknowledgement

Chapter 1 “Inference on Functions under First Order Degeneracy,” in part, is currently being prepared for submission for publication of the material. Chen, Qihui; Fang, Zheng. The dissertation author was the primary investigator and author of this material.

Table 1.2: Rejection rates under the null: Design D1

$T\backslash$Tests	CF1			CF2			DG	
	$T^{-1/2}$	$T^{-2/3}$	$T^{-4/5}$	$T^{-1/4}$	$T^{-1/3}$	$T^{-2/5}$	DG1	DG2
1000	0.0850	0.0640	0.0420	0.0395	0.0185	0.0100	0.3975	0.4015
2000	0.0940	0.0715	0.0530	0.0550	0.0320	0.0120	0.5060	0.5045
5000	0.1010	0.0740	0.0515	0.0505	0.0290	0.0075	0.6215	0.6185
10000	0.1010	0.0820	0.0585	0.0550	0.0285	0.0090	0.6375	0.6270
20000	0.1005	0.0725	0.0525	0.0495	0.0285	0.0115	0.6750	0.6705
40000	0.1180	0.0900	0.0670	0.0700	0.0410	0.0165	0.6865	0.6845
50000	0.1070	0.0830	0.0660	0.0665	0.0410	0.0145	0.6895	0.6870

Table 1.3: Rejection rates under the null: Design D3

$T\backslash$Tests	CF1			CF2			DG	
	$T^{-1/2}$	$T^{-2/3}$	$T^{-4/5}$	$T^{-1/4}$	$T^{-1/3}$	$T^{-2/5}$	DG1	DG2
1000	0.0605	0.0390	0.0285	0.0660	0.0605	0.0430	0.2300	0.2400
2000	0.0645	0.0385	0.0280	0.0655	0.0570	0.0380	0.3425	0.3470
5000	0.0520	0.0385	0.0315	0.0505	0.0455	0.0275	0.3970	0.3965
10000	0.0690	0.0565	0.0450	0.0830	0.0665	0.0320	0.4385	0.4415
20000	0.0660	0.0600	0.0490	0.0850	0.0660	0.0335	0.4765	0.4790
40000	0.0520	0.0460	0.0390	0.0645	0.0475	0.0225	0.5030	0.5065
50000	0.0745	0.0670	0.0585	0.0920	0.0635	0.0395	0.5255	0.5290

Table 1.4: Rejection rates under the null: Design D4

$T\backslash$Tests	CF1			CF2			DG	
	$T^{-1/2}$	$T^{-2/3}$	$T^{-4/5}$	$T^{-1/4}$	$T^{-1/3}$	$T^{-2/5}$	DG1	DG2
1000	0.0715	0.0445	0.0265	0.1305	0.0915	0.0415	0.4795	0.4870
2000	0.0895	0.0515	0.0380	0.1485	0.0935	0.0330	0.6380	0.6515
5000	0.1055	0.0720	0.0545	0.1590	0.0960	0.0300	0.7810	0.7820
10000	0.1135	0.0615	0.0485	0.1440	0.0750	0.0290	0.8055	0.8030
20000	0.1155	0.0715	0.0555	0.1530	0.0960	0.0290	0.8495	0.8485
40000	0.1280	0.0810	0.0640	0.1655	0.0900	0.0300	0.8650	0.8670
50000	0.1150	0.0775	0.0660	0.1650	0.0855	0.0260	0.8610	0.8590

Table 1.5: Rejection rates under the alternative

$T\backslash$Tests	Design D2						Design D5					
	CF1			CF2			CF1			CF2		
	$T^{-1/2}$	$T^{-2/3}$	$T^{-4/5}$	$T^{-1/4}$	$T^{-1/3}$	$T^{-2/5}$	$T^{-1/2}$	$T^{-2/3}$	$T^{-4/5}$	$T^{-1/4}$	$T^{-1/3}$	$T^{-2/5}$
1000	0.6450	0.5915	0.5050	0.7255	0.6890	0.5570	0.1240	0.0740	0.0630	0.3990	0.3645	0.3000
2000	0.9410	0.9185	0.8805	0.9530	0.9365	0.8785	0.3520	0.2710	0.2300	0.6975	0.6675	0.5570
5000	0.9975	0.9975	0.9960	0.9995	0.9990	0.9950	0.8250	0.7710	0.7255	0.9610	0.9460	0.8885
10000	0.9980	0.9980	0.9975	0.9985	0.9985	0.9985	0.9865	0.9850	0.9755	0.9995	0.9985	0.9955
20000	0.9985	0.999	0.9985	0.9995	0.9995	0.9985	0.9980	0.9970	0.9955	1.0000	1.0000	1.0000
40000	0.9995	0.9995	0.9995	1.0000	1.0000	1.0000	1.0000	1.0000	0.9985	1.0000	1.0000	1.0000
50000	0.9995	0.9995	0.9995	0.9995	0.9995	0.9995	0.9995	0.9995	0.9990	1.0000	1.0000	1.0000

1.8 Appendix

1.8.1 Proofs of Main Results

PROOF OF THEOREM 1.2.1: For each $n \in \mathbf{N}$, let $\mathbb{D}_n \equiv \{h \in \mathbb{D} : \theta_0 + h/r_n \in \mathbb{D}_\phi\}$ and define $g_n : \mathbb{D}_n \to \mathbb{E}$ by

$$g_n(h_n) \equiv r_n^2\{\phi(\theta_0 + r_n^{-1}h_n) - \phi(\theta_0) - r_n^{-1}\phi'_{\theta_0}(h_n)\} \text{ for any } h_n \in \mathbb{D}_n \ .$$

By Assumption 1.2.1, $\|g_n(h_n) - \phi''_{\theta_0}(h)\|_{\mathbb{E}} \to 0$ whenever $h_n \to h \in \mathbb{D}_0$. Moreover, $\mathbb{G} \in \mathbb{D}_0$ (almost surely) is separable since it is tight by Assumption 1.2.2(ii). The first claim of the theorem then follows by Theorem 1.11.1(i) in van der Vaart and Wellner (1996a).

As for the second claim, define $f_n : \mathbb{D}_n \times \mathbb{D} \to \mathbb{E} \times \mathbb{E}$ by

$$f_n(h_n, h) \equiv (g_n(h_n), \phi''_{\theta_0}(h)) \text{ for any } (h_n, h) \in \mathbb{D}_n \times \mathbb{D} \ .$$

Assumption 1.2.1 and 1.2.3(i) allow us to conclude again by Theorem 1.11.1(i) in van der Vaart and Wellner (1996a) that

$$\begin{bmatrix} r_n^2\{\phi(\hat{\theta}_n) - \phi(\theta_0) - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)\} \\ \\ \phi''_{\theta_0}(r_n\{\hat{\theta}_n - \theta_0\}) \end{bmatrix} \xrightarrow{L} \begin{bmatrix} \phi''_{\theta_0}(\mathbb{G}) \\ \\ \phi''_{\theta_0}(\mathbb{G}) \end{bmatrix} \text{ in } \mathbb{E} \times \mathbb{E} \ . \tag{1.65}$$

By the continuous mapping theorem applied to result (1.65), we have

$$r_n^2\{\phi(\hat{\theta}_n) - \phi(\theta_0) - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)\} - \phi''_{\theta_0}(r_n\{\hat{\theta}_n - \theta_0\}) \xrightarrow{L} 0 \ . \tag{1.66}$$

The second claim then follows from result (1.66) and Lemma 1.10.2(iii) in van der Vaart and Wellner (1996a). ■

PROOF OF THEOREM 1.3.1: Inspecting the structure of the problem, we see that the

bootstrap consistency (1.27) is equivalent to $\phi''_{\theta_0}(\mathbb{G}+h)-\phi''_{\theta_0}(h) \overset{d}{=} \phi''_{\theta_0}(\mathbb{G})$ for all $h \in \text{Supp}(\mathbb{G})$ by exactly the same arguments as the proof of Theorem A.1 in Fang and Santos (2015). Thus, it boils down to showing that $\phi''_{\theta_0}(\mathbb{G}+h)-\phi''_{\theta_0}(h) \overset{d}{=} \phi''_{\theta_0}(\mathbb{G})$ for all $h \in \text{Supp}(\mathbb{G})$ if and only if $\phi''_{\theta_0}(h)=0$ for $\mathbb{G}$-almost h in $\mathbb{D}_0$. One direction is immediate since if latter holds, then both $\phi''_{\theta_0}(\mathbb{G}+h)-\phi''_{\theta_0}(h)$ and $\phi''_{\theta_0}(\mathbb{G})$ are degenerate at 0 for all $h \in \text{Supp}(\mathbb{G})$, and hence are equal in distribution.

The converse consists of two steps. To begin with, note that by Assumption 1.2.2(ii), $\mathbb{G}$ being centered Gaussian and Lemma A.7 in Fang and Santos (2015), we may assume without loss of generality that the support of $\mathbb{G}$ is $\mathbb{D}$ and that $\mathbb{D}$ is separable. Since $\mathbb{D}$ is separable, it follows that the Borel σ-algebra, the σ-algebra generated by the weak topology, and the cylindrical σ-algebra coincide by Theorem 2.1 in Vakhania et al. (1987). Furthermore, by Theorem 7.1.7 in Bogachev (2007), P is Radon with respect to the Borel σ-algebra, and hence also with respect to the cylindrical σ-algebra. Finally, let P be the probability measure on $\mathbb{D}$ induced by $\mathbb{G}$.

STEP 1: Show that ϕ''_{θ_0} corresponds to a bilinear map if $\phi''_{\theta_0}(\mathbb{G}+h)-\phi''_{\theta_0}(h) \overset{d}{=} \phi''_{\theta_0}(\mathbb{G})$ for all $h \in \text{Supp}(\mathbb{G})$.

For completeness, we introduce additional notation following Section 3.7 in Davydov et al. (1998). First, let $\mathbb{D}^*$ denote the dual space of $\mathbb{D}$, and $\langle x, x^* \rangle_{\mathbb{D}} = x^*(x)$ for any $x \in \mathbb{D}$ and $x^* \in \mathbb{D}^*$. Similarly denote the dual space of $\mathbb{E}$ by $\mathbb{E}^*$ and the corresponding bilinear form by $\langle \cdot, \cdot \rangle_{\mathbb{E}}$. Since $\mathbb{G}$ is Gaussian, $\mathbb{D}^* \subset L^2(P)$ (Bogachev, 1998, p.42). We may thus embed $\mathbb{D}^*$ into $L^2(P)$. Denote by $\mathbb{D}'_P$ the closure of $\mathbb{D}^*$, viewed as a subset of $L^2(P)$. By some abuse of notation write $x'(x) = \langle x', x \rangle_{\mathbb{D}}$ for any $x' \in \mathbb{D}'_P$ and $x \in \mathbb{D}$. Finally, for each $h \in \mathbb{D}$ we let P^h denote the law of $\mathbb{G}+h$, write $P^h \ll P$ whenever P^h is absolutely continuous with respect to P, and define the set:

$$\mathbb{H}_P \equiv \{h \in \mathbb{D} : P^{rh} \ll P \text{ for all } r \in \mathbf{R}\} .$$

Since P is Radon with respect to the cylindrical σ-algebra of $\mathbb{D}$, it follows by Theorem

7.1 in Davydov et al. (1998) that there exists a continuous linear map $I : \mathbb{H}_P \to \mathbb{D}'_P$ satisfying for every $h \in \mathbb{H}_P$:

$$\frac{dP^h}{dP}(x) = \exp\left\{\langle x, Ih\rangle_{\mathbb{D}} - \frac{1}{2}\sigma^2(h)\right\} \qquad \sigma^2(h) \equiv \int_{\mathbb{D}} \langle x, Ih\rangle^2_{\mathbb{D}}\, P(dx)\ . \tag{1.67}$$

Fix an arbitrary $e^* \in \mathbb{E}^*$ and $h \in \mathbb{H}_P$. Since $\phi''_{\theta_0}(\mathbb{G}+h) - \phi''_{\theta_0}(h) \stackrel{d}{=} \phi''_{\theta_0}(\mathbb{G})$ for all $h \in \text{Supp}(\mathbb{G})$, it follows that $\langle e^*, \phi''_{\theta_0}(\mathbb{G}+rh) - \phi''_{\theta_0}(rh)\rangle_{\mathbb{E}}$ and $\langle e^*, \phi''_{\theta_0}(\mathbb{G})\rangle_{\mathbb{E}}$ must be equal in distribution for all $r \in \mathbf{R}$.[16] In particular, their characteristic functions must equal each other, and hence for all $r \geq 0$ and $t \in \mathbf{R}$:

$$\begin{aligned} E[\exp\{it\langle e^*, \phi''_{\theta_0}(\mathbb{G})\rangle_{\mathbb{E}}\}] &= E[\exp\{it\{\langle e^*, \phi''_{\theta_0}(\mathbb{G}+rh) - \phi''_{\theta_0}(rh)\rangle_{\mathbb{E}}\}\}] \\ &= \exp\{-itr^2\langle e^*, \phi''_{\theta_0}(h)\rangle_{\mathbb{E}}\}E[\exp\{it\langle e^*, \phi''_{\theta_0}(\mathbb{G}+rh)\rangle_{\mathbb{E}}\}]\ , \end{aligned} \tag{1.68}$$

where in the second equality we have exploited ϕ''_{θ_0} being positively homogenous of degree two. Setting $C(t) \equiv E[\exp\{it\langle e^*, \phi''_{\theta_0}(\mathbb{G})\rangle_{\mathbb{E}}\}]$, we have by (1.68) that

$$\exp\{itr^2\langle e^*, \phi''_{\theta_0}(h)\rangle_{\mathbb{E}}\}C(t) = E[\exp\{it\langle e^*, \phi''_{\theta_0}(\mathbb{G}+rh)\rangle_{\mathbb{E}}\}]\ , \tag{1.69}$$

for all $r \geq 0$ and $t \in \mathbf{R}$.

We next aim to equate second order right derivatives of both sides in the identity (1.69). The second order right derivative of the left hand side at $r = 0$ is given by

$$2itC(t)\langle e^*, \phi''_{\theta_0}(h)\rangle_{\mathbb{E}}\ . \tag{1.70}$$

On the other hand, exploiting result (1.67), linearity of $I : \mathbb{H}_P \to \mathbb{D}'_P$ and that $h \in \mathbb{H}_P$ implies $rh \in \mathbb{H}_P$ for all $r \in \mathbf{R}$ and in particular for all $r \in [0,1]$, we may rewrite the right

[16] The proof of Lemma A.3 in Fang and Santos (2015) never exploits that ϕ'_{θ_0} is a first order derivative beyond continuity of ϕ'_{θ_0} and $\phi'_{\theta_0}(0) = 0$ which are satisfied by ϕ''_{θ_0}.

hand side of (1.69) as

$$E[\exp\{it\langle e^*, \phi''_{\theta_0}(\mathbb{G}+rh)\rangle_{\mathbb{E}}\}] = \int_{\mathbb{D}} \exp\{it\langle e^*, \phi''_{\theta_0}(x)\rangle_{\mathbb{E}}\}\frac{dP^{rh}}{dP}(x)P(dx)$$
$$= \int_{\mathbb{D}} \exp\left\{it\langle e^*, \phi''_{\theta_0}(x)\rangle_{\mathbb{E}} + r\langle x, Ih\rangle_{\mathbb{D}} - \frac{r^2}{2}\sigma^2(h)\right\}P(dx)\ . \qquad (1.71)$$

The integrand on the right hand side of (1.71) is differentiable with respect to r for all $r \in [0,1]$ and the resulting derivative is dominated by $\exp\{|\langle x, Ih\rangle_{\mathbb{D}}|\}\times\{|\langle x, Ih\rangle_{\mathbb{D}}|+\sigma^2(h)\}$ which is integrable against P since $\langle \mathbb{G}, Ih\rangle_{\mathbb{D}} \sim N(0, \sigma^2(h))$ by Proposition 2.10.3 in Bogachev (1998) and $Ih \in \mathbb{D}'_P$. Thus by Theorem 2.27(ii) in Folland (1999), the first order derivative of the right hand side in (1.71) at $r \in [0,1]$ exists and is given by

$$\int_{\mathbb{D}} \exp\left\{it\langle e^*, \phi''_{\theta_0}(x)\rangle_{\mathbb{E}} + r\langle x, Ih\rangle_{\mathbb{D}} - \frac{r^2}{2}\sigma^2(h)\right\}\{\langle x, Ih\rangle_{\mathbb{D}} - r\sigma^2(h)\}P(dx)\ . \qquad (1.72)$$

In turn, result (1.72) allows us to conclude that the second order right derivative of the right hand side in (1.71) at $r = 0$ exists and is given by

$$\int_{\mathbb{D}} \exp\{it\langle e^*, \phi''_{\theta_0}(x)\rangle_{\mathbb{E}}\}[\langle x, Ih\rangle^2_{\mathbb{D}} - \sigma^2(h)]P(dx)\ . \qquad (1.73)$$

Since the equation (1.69) holds for all $r \geq 0$ and $t \in \mathbf{R}$, it follows from (1.70) and (1.73) that for all $t \in \mathbf{R}$:

$$2itC(t)\langle e^*, \phi''_{\theta_0}(h)\rangle_{\mathbb{E}} = \int_{\mathbb{D}} \exp\{it\langle e^*, \phi''_{\theta_0}(x)\rangle_{\mathbb{E}}\}[\langle x, Ih\rangle^2_{\mathbb{D}} - \sigma^2(h)]P(dx)\ . \qquad (1.74)$$

Note that $t \mapsto C(t)$ is the characteristic function of $\langle e^*, \phi''_{\theta_0}(\mathbb{G}_0)\rangle_{\mathbb{E}}$ and hence it is continuous. Thus, since $C(0) = 1$ there exists a $t_0 > 0$ such that $C(t_0)t_0 \neq 0$. For such t_0 it follows from (1.74) that

$$\langle e^*, \phi''_{\theta_0}(h)\rangle_{\mathbb{E}} = -\frac{iE[\exp\{it_0\langle e^*, \phi''_{\theta_0}(\mathbb{G})\rangle_{\mathbb{E}}\}\{\langle \mathbb{G}, Ih\rangle^2_{\mathbb{D}} - \sigma^2(h)\}]}{2t_0C(t_0)}\ . \qquad (1.75)$$

Define a map $\Phi''_{\theta_0} : \mathbb{D} \times \mathbb{D} \to \mathbb{E}$ by

$$\Phi''_{\theta_0}(h, g) \equiv \frac{1}{4}[\phi''_{\theta_0}(h + g) - \phi''_{\theta_0}(h - g)] \ . \tag{1.76}$$

It then follows from (1.75) that, for any $e^* \in \mathbb{E}^*$ and any $g, h \in \mathbb{D}$,

$$\langle e^*, \Phi''_{\theta_0}(g, h)\rangle_{\mathbb{E}} = -\frac{iE[\exp\{it_0\langle e^*, \phi''_{\theta_0}(\mathbb{G})\rangle_{\mathbb{E}}\}\{\langle \mathbb{G}, Ig\rangle_{\mathbb{D}}\langle \mathbb{G}, Ih\rangle_{\mathbb{D}} - \sigma(g, h)\}]}{2t_0C(t_0)} \ , \tag{1.77}$$

where $\sigma(g, h) \equiv E[\langle \mathbb{G}, Ig\rangle\langle \mathbb{G}, Ih\rangle]$. Since $I : \mathbb{H}_P \to \mathbb{D}'_P$ is linear, $(h, g) \mapsto \langle e^*, \Phi''_{\theta_0}(g, h)\rangle_{\mathbb{E}}$ is bilinear on $\mathbb{H}_P \times \mathbb{H}_P$. Moreover, $(h, g) \mapsto \langle e^*, \Phi''_{\theta_0}(g, h)\rangle_{\mathbb{E}}$ is continuous on $\mathbb{H}_P \times \mathbb{H}_P$ due to continuity of ϕ''_{θ_0} (and hence Φ''_{θ_0}) and $e^* \in \mathbb{E}^*$. We thus conclude from $\mathbb{H}_P$ being a dense subspace of $\mathbb{D}$ by Proposition 7.4(ii) in Davydov et al. (1998) that $(h, g) \mapsto \langle e^*, \Phi''_{\theta_0}(g, h)\rangle_{\mathbb{E}}$ is continuous and bilinear on $\mathbb{D} \times \mathbb{D}$. Since $e^* \in \mathbb{E}^*$ is arbitrary, it follows from Lemma A.2 in van der Vaart (1991) that $\Phi''_{\theta_0} : \mathbb{D} \times \mathbb{D} \to \mathbb{E}$ is bilinear and continuous. By identity (1.76), we have $\phi''_{\theta_0}(h) = \Phi''_{\theta_0}(h, h)$ for all $h \in \mathbb{D}$. Hence, ϕ''_{θ_0} is a quadratic form corresponding to the bilinear map Φ''_{θ_0}.

STEP 2: Conclude that $\phi''_{\theta_0} = 0$ on the support of $\mathbb{G}$. Note that if ϕ is second order Hadamard differentiable, then one can directly start with Step 2.

By Lemma A.3 in Fang and Santos (2015), for all $h \in \mathbb{D}$,

$$\begin{aligned}
\phi''_{\theta_0}(\mathbb{G}) &\stackrel{d}{=} \phi''_{\theta_0}(\mathbb{G} + h) - \phi''_{\theta_0}(h) \\
&= \Phi''_{\theta_0}(\mathbb{G} + h, \mathbb{G} + h) - \Phi''_{\theta_0}(h, h) \\
&= \Phi''_{\theta_0}(\mathbb{G}, \mathbb{G}) + 2\Phi''_{\theta_0}(\mathbb{G}, h) = \phi''_{\theta_0}(\mathbb{G}) + 2\Phi''_{\theta_0}(\mathbb{G}, h) \ ,
\end{aligned} \tag{1.78}$$

where the third equality exploited bilinearity of Φ''_{θ_0}. Fix an arbitrary $e^* \in \mathbb{E}^*$. By result (1.78), we have for all $r \in \mathbf{R}$ and $h \in \mathbb{D}$,

$$\begin{aligned}
E[\exp\{it\langle e^*, \phi''_{\theta_0}(\mathbb{G})\rangle_{\mathbb{E}}\}] &= E[\exp\{it\langle e^*, \phi''_{\theta_0}(\mathbb{G}) + 2\Phi''_{\theta_0}(\mathbb{G}, rh)\rangle_{\mathbb{E}}\}] \\
&= E[\exp\{it\langle e^*, \phi''_{\theta_0}(\mathbb{G})\rangle_{\mathbb{E}}\}\exp\{2irt\langle e^*, \Phi''_{\theta_0}(\mathbb{G}, h)\rangle_{\mathbb{E}}\}] \ ,
\end{aligned} \tag{1.79}$$

where the last step used linearity of Φ''_{θ_0} in its second argument. We now equate second derivatives of both sides at $r = 0$. The second derivative of the left hand side is trivially zero, while that of the right hand side, by the recursive use of dominated convergence arguments, is given by $E[\exp\{it\langle e^*, \phi''_{\theta_0}(\mathbb{G})\rangle_{\mathbb{E}}\}\{2it\langle e^*, \Phi''_{\theta_0}(\mathbb{G}, h)\rangle_{\mathbb{E}}\}^2]$. Thus we have for all $t \in \mathbf{R}$,

$$E[\exp\{it\langle e^*, \phi''_{\theta_0}(\mathbb{G})\rangle_{\mathbb{E}}\}\{2it\langle e^*, \Phi''_{\theta_0}(\mathbb{G}, h)\rangle_{\mathbb{E}}\}^2] = 0 \ ,$$

which in turn implies that for all $t \in \mathbf{R} \setminus \{0\}$,

$$E[\exp\{it\langle e^*, \phi''_{\theta_0}(\mathbb{G})\rangle_{\mathbb{E}}\}\langle e^*, \Phi''_{\theta_0}(\mathbb{G}, h)\rangle^2_{\mathbb{E}}] = 0 \ . \tag{1.80}$$

Picking a sequence $t_n \downarrow 0$, replacing t with t_n in (1.80) and letting $n \to \infty$ leads to, by the dominated convergence theorem: for all $e^* \in \mathbb{E}^*$ and all $h \in \mathbb{D}$,

$$E[\langle e^*, \Phi''_{\theta_0}(\mathbb{G}, h)\rangle^2_{\mathbb{E}}] = 0 \ . \tag{1.81}$$

Consequently, $\langle e^*, \Phi''_{\theta_0}(g, h)\rangle_{\mathbb{E}} = 0$ for all $h \in \mathbb{D}$ and P-almost surely $g \in \mathbb{D}$. Since e^* is arbitrary, we conclude by Lemma 6.10 in Aliprantis and Border (2006) that $\Phi''_{\theta_0}(g, h) = 0$ for all $h \in \mathbb{D}$ and P-almost $g \in \mathbb{D}$. Hence, $\phi''_{\theta_0}(h) = 0$ for P-almost $h \in \mathbb{D}$.

Finally, denote by Ω the collection of all $h \in \mathbb{D}$ such that $\phi''_{\theta_0}(h) = 0$. Then we have $P(\Omega) = 1$ by Assumption 1.2.2(ii) and the above discussion. We claim that Ω is dense in the support of P. To see this, suppose otherwise and then there must exist some $h_0 \in \text{Supp}(P)$ and some $\delta > 0$ such that $B(h_0, \delta) \cap \Omega = \emptyset$. Note that i) $P(B(h_0, \delta)) > 0$ since $h_0 \in \text{Supp}(P)$, and ii) $\phi''_{\theta_0}(h) \neq 0$ for all $h \in B(h_0, \delta)$ by the definition of Ω. These contradict the fact $P(\Omega) = 1$. Since ϕ''_{θ_0} is continuous on $\overline{\mathbb{D}_0}$ by Assumption 1.2.3(i), it is also continuous on the support of P as a subset of $\overline{\mathbb{D}_0}$ by Assumption 1.2.2(ii) and Theorem II.2.1 in Parthasarathy (1967). In turn, we may conclude from Ω being dense in $\text{Supp}(P)$ and $\phi''_{\theta_0} = 0$ on Ω that $\phi''_{\theta_0} = 0$ on $\text{Supp}(P)$. ■

Proof of Theorem 1.3.2: Let $\mathbb{D}_n \equiv \{h \in \mathbb{D} : \theta_0 + h/r_n \in \mathbb{D}_\phi\}$ and define for each $n \in \mathbf{N}$

the map $\Psi_n : \mathbb{D}_n \times \mathbb{D}_n \to \mathbb{E}$ by

$$\Psi_n(g_n, h_n) \equiv \frac{\phi(\theta_0 + r_n^{-1}h_n) - \phi(\theta_0 + r_n^{-1}g_n) - \phi'_{\theta_0 + r_n^{-1}g_n}(r_n^{-1}\{h_n - g_n\})}{r_n^{-2}} \ .$$

If $\{g_n, h_n\}_{n=1}^{\infty} \subset \mathbb{D}_n$ satisfies $(g_n, h_n) \to (g, h) \in \mathbb{D}_0 \times \mathbb{D}_0$ as $n \to \infty$, then Assumption 1.3.3 allows us to conclude that

$$\begin{aligned}
\Psi_n(g_n, h_n) &\equiv \frac{\phi(\theta_0 + r_n^{-1}h_n) - \phi(\theta_0 + r_n^{-1}g_n) - \phi'_{\theta_0 + r_n^{-1}g_n}(r_n^{-1}\{h_n - g_n\})}{r_n^{-2}} \\
&= \frac{\{\phi(\theta_0 + r_n^{-1}h_n) - \phi(\theta_0) - r_n^{-1}\phi'_{\theta_0}(h_n)\} - \{\phi(\theta_0 + r_n^{-1}g_n) - \phi(\theta_0) - r_n^{-1}\phi'_{\theta_0}(g_n)\}}{r_n^{-2}} \\
&\quad - \frac{\{\phi'_{\theta_0 + r_n^{-1}g_n}(h_n) - \phi'_{\theta_0}(h_n)\} - \{\phi'_{\theta_0 + r_n^{-1}g_n}(g_n) - \phi'_{\theta_0}(g_n)\}}{r_n^{-1}} \\
&\to \Phi''_{\theta_0}(h, h) - \Phi''_{\theta_0}(g, g) - 2\Phi''_{\theta_0}(g, h) + 2\Phi''_{\theta_0}(g, g) \\
&= \Psi(g, h) \equiv \Phi''_{\theta_0}(h, h) + \Phi''_{\theta_0}(g, g) - 2\Phi''_{\theta_0}(g, h) \ . \qquad (1.82)
\end{aligned}$$

Since ϕ''_{θ_0} admitting a continuous extension on $\mathbb{D}$, by corresponding extension of Φ''_{θ_0} according to equation 1.76, it follows from (1.82) that

$$\Psi_n(g_n, h_n) - \Psi(g_n, h_n) = \Psi_n(g_n, h_n) - \Psi(g, h) - \{\Psi(g_n, h_n) - \Psi(g, h)\} \to 0 \ . \qquad (1.83)$$

Next, for notational simplicity, let $\mathbb{G}_n \equiv r_n\{\hat{\theta}_n - \theta_0\}$, $\mathbb{G}_n^* \equiv r_n\{\hat{\theta}_n^* - \hat{\theta}_n\}$ and $\mathbb{G}_n^\dagger \equiv r_n\{\hat{\theta}_n^* - \theta_0\} = \mathbb{G}_n^* + \mathbb{G}_n$. By Assumption 1.2.1, 1.2.2, 1.2.3(ii), 1.3.1 and 1.3.2(i), it follows from Lemma A.2 in Fang and Santos (2015) that for $\mathbb{G}_1, \mathbb{G}_2$ independent distributed according to $\mathbb{G}$,

$$(\mathbb{G}_n, \mathbb{G}_n^*) \xrightarrow{L} (\mathbb{G}_1, \mathbb{G}_2) \ . \qquad (1.84)$$

By the continuous mapping theorem and result (1.84) we have

$$(\mathbb{G}_n, \mathbb{G}_n^\dagger) = (\mathbb{G}_n, \mathbb{G}_n^* + \mathbb{G}_n) \xrightarrow{L} (\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2) \ . \qquad (1.85)$$

Combining the separability of $\mathbb{G}_1$ and $\mathbb{G}_2$ by Assumption 1.2.2(ii), results (1.83) and (1.85), we conclude by Theorem 1.11.1(i) in van der Vaart and Wellner (1996a) that

$$\Psi_n(\mathbb{G}_n, \mathbb{G}_n^\dagger) - \Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger) \xrightarrow{L} 0 \ . \tag{1.86}$$

By Lemma 1.10.2 in van der Vaart and Wellner (1996a) we have from (1.86) that

$$\Psi_n(\mathbb{G}_n, \mathbb{G}_n^\dagger) - \Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger) = o_p(1) \ . \tag{1.87}$$

Now fix $\epsilon > 0$. Note that

$$\begin{aligned} \sup_{f \in \mathrm{BL}_1(\mathbb{E})} & |E_W^*[f(\Psi_n(\mathbb{G}_n, \mathbb{G}_n^\dagger))] - E_W^*[f(\Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger))]| \\ & \leq \epsilon + 2P_W^*(\|\Psi_n(\mathbb{G}_n, \mathbb{G}_n^\dagger) - \Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger)\|_{\mathbb{E}} > \epsilon) \ . \end{aligned} \tag{1.88}$$

By Lemma 1.2.6 in van der Vaart and Wellner (1996a),

$$E_X^*[P_W^*(\|\Psi_n(\mathbb{G}_n, \mathbb{G}_n^\dagger) - \Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger)\|_{\mathbb{E}} > \epsilon)] \leq P^*(\|\Psi_n(\mathbb{G}_n, \mathbb{G}_n^\dagger) - \Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger)\|_{\mathbb{E}} > \epsilon) \ . \tag{1.89}$$

Results (1.87), (1.88) and (1.89), together with ϵ being arbitrary, then yield

$$\sup_{f \in \mathrm{BL}_1(\mathbb{E})} |E_W^*[f(\Psi_n(\mathbb{G}_n, \mathbb{G}_n^\dagger))] - E_W^*[f(\Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger))]| = o_p(1) \ . \tag{1.90}$$

Result (1.85) and Assumption 1.2.2(ii) implies that $(\mathbb{G}_n, \mathbb{G}_n^\dagger)$ is asymptotically measurable and asymptotically tight. In turn, Lemmas 1.4.3 and 1.4.4 in van der Vaart and Wellner (1996a) implies that $(\mathbb{G}_n, \mathbb{G}_n^\dagger, \mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2)$ is asymptotically tight and asymptotically measurable. Fix an arbitrary subsequence $\{n_k\}$. Then Theorem 1.3.9 in van der Vaart and Wellner (1996a) implies that $(\mathbb{G}_n, \mathbb{G}_n^\dagger, \mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2)$ converges weakly along a further subsequence of $\{n_k\}$ to a tight Borel law in $\prod_{j=1}^4 \mathbb{D}$, which is equal to $(\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2, \mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2)$ by marginal convergence. This is a weak limit where the dependence structure between the

first two components and last two components is known and in fact unique. Since n_k is arbitrary, it follows that

$$(\mathbb{G}_n, \mathbb{G}_n^\dagger, \mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2) \xrightarrow{L} (\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2, \mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2) \ . \tag{1.91}$$

Since $\Psi : \mathbb{D} \times \mathbb{D} \to \mathbb{E}$ and hence $(\Psi, \Psi) : \prod_{j=1}^4 \mathbb{D} \to \prod_{j=1}^2 \mathbb{E}$ is continuous, it follows from result (1.91) and the continuous mapping theorem that

$$(\Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger), \Psi(\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2)) \xrightarrow{L} (\Psi(\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2), \Psi(\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2)) \ . \tag{1.92}$$

Combination of the continuous mapping theorem and Lemma 1.10.2(iii) in van der Vaart and Wellner (1996a) yields that

$$\Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger) - \Psi(\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2) = o_p(1) \ . \tag{1.93}$$

By the triangle inequality, we have

$$\begin{aligned} \sup_{f \in \mathrm{BL}_1(\mathbb{E})} & |E_W^*[f(\Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger))] - E[f(\Psi(\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2))]| \\ & \leq \epsilon + 2P_W^*(\|\Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger) - \Psi(\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2)\|_{\mathbb{E}} > \epsilon) \ . \end{aligned} \tag{1.94}$$

By Lemma 1.2.6 in van der Vaart and Wellner (1996a) and result (1.93)

$$\begin{aligned} E_X^* P_W^*(\|\Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger) & - \Psi(\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2)\|_{\mathbb{E}} > \epsilon) \\ & \leq P^*(\|\Psi(\mathbb{G}_n, \mathbb{G}_n^\dagger) - \Psi(\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2)\|_{\mathbb{E}} > \epsilon) = o(1) \ . \end{aligned} \tag{1.95}$$

Combination of (1.90), (1.94), (1.95) and the triangle inequality leads to

$$\sup_{f \in \mathrm{BL}_1(\mathbb{E})} |E_W^*[f(\Psi_n(\mathbb{G}_n, \mathbb{G}_n^\dagger))] - E[f(\Psi(\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2))]| = o_p(1) \ . \tag{1.96}$$

The theorem follows by combining (1.90) and (1.96) and noticing that

$$\Psi_n(\mathbb{G}_n, \mathbb{G}_n^\dagger) = r_n^2\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n) - \phi'_{\hat{\theta}_n}(\hat{\theta}_n^* - \hat{\theta}_n)\} \text{ and } \Psi(\mathbb{G}_1, \mathbb{G}_1 + \mathbb{G}_2) = \phi''_{\theta_0}(\mathbb{G}_2) \ ,$$

where the second equality is due to bilinearity of Φ''_{θ_0}. ■

Proof of Theorem 1.3.3: Inspecting the proof of Theorem 3.2 in Fang and Santos (2015), we see that ϕ'_{θ_0} being a first order derivative is actually never exploited there. The conclusion of the theorem then follows in view of Lemma 1.8.2 when combined with exactly the same arguments in Fang and Santos (2015). ■

Proof of Proposition 1.3.1: Let $\{h_n\} \subset \mathbb{D}$ and $h \in \mathbb{D}_0$ such that $h_n \to h$. By Assumption 1.2.3(iii) $\phi'_{\theta_0} = 0$, so we may rewrite $\hat{\phi}''_n(h_n)$:

$$\begin{aligned}\hat{\phi}''_n(h_n) &= \frac{\phi(\hat{\theta}_n + t_n h_n) - \phi(\hat{\theta}_n) - t_n \phi'_{\theta_0}(h_n)}{t_n^2} \\ &= \frac{\phi(\theta_0 + t_n g_n) - \phi(\theta_0) - t_n \phi'_{\theta_0}(g_n)}{t_n^2} - \frac{r_n^2\{\phi(\hat{\theta}_n) - \phi(\theta_0) - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)\}}{(r_n t_n)^2} \ , \end{aligned} \tag{1.97}$$

where $g_n \equiv (t_n r_n)^{-1} r_n\{\hat{\theta}_n - \theta_0\} + h_n$. By Assumptions 1.2.2(i), 1.3.5, Lemma 1.10.2 in van der Vaart and Wellner (1996a) and $h_n \to h$, we have $g_n \xrightarrow{p} h$. By Assumptions 1.2.1, 1.2.2(ii) and Theorem 1.11.1(ii) in van der Vaart and Wellner (1996a), it then follows that

$$\frac{\phi(\theta_0 + t_n g_n) - \phi(\theta_0) - t_n \phi'_{\theta_0}(g_n)}{t_n^2} \xrightarrow{p} \phi''_{\theta_0}(h) \ . \tag{1.98}$$

By Assumption 1.2.1 and 1.2.2, it follows from Theorem 1.2.1 and $r_n t_n \to \infty$ that

$$\frac{r_n^2\{\phi(\hat{\theta}_n) - \phi(\theta_0) - \phi'_{\theta_0}(\hat{\theta}_n - \theta_0)\}}{(r_n t_n)^2} \xrightarrow{p} 0 \ . \tag{1.99}$$

Combining results (1.97), (1.98) and (1.99) we thus arrive at the desired conclusion. ■

Proof of Lemma 1.4.1: By Assumptions 1.2.2, 1.4.1 and 1.4.2, we have for $P_n \equiv P_{1/r_n}$,

$$r_n\{\hat{\theta}_n - \theta(P)\} = r_n\{\hat{\theta}_n - \theta(P_n)\} + r_n\{\theta(P_n) - \theta(P)\} \xrightarrow{L_n} \mathbb{G} + \theta'_0(h) \ . \tag{1.100}$$

Combination of Assumptions 1.2.1, 1.2.3(ii), $\phi(\theta(P)) = \phi'_{\theta_0} = 0$, and result (1.100) allows us to invoke the second order Delta method to conclude that

$$r_n^2 \phi(\hat{\theta}_n) = r_n^2 \{\phi(\hat{\theta}_n) - \phi(\theta(P)) - \phi'_{\theta_0}(\hat{\theta}_n - \theta(P))\} \xrightarrow{L_n} \phi''_{\theta_0}(\mathbb{G} + \theta'_0(h)) \ . \tag{1.101}$$

This completes the proof of the lemma. ■

Proof of Theorem 1.4.1: Under the assumptions in Theorem 1.3.3 and Assumptions 1.4.3(i)(ii), we can show following the proof of Corollary 3.2 in Fang and Santos (2015) that $\hat{c}_{1-\alpha} \xrightarrow{p} c_{1-\alpha}$ under P^n. By Theorem 12.2.3 and Corollary 12.3.1 in Lehmann and Romano (2005), P_n^n and P^n are mutually contiguous. It follows that

$$\hat{c}_{1-\alpha} \xrightarrow{p} c_{1-\alpha} \text{ under } P_n^n \ . \tag{1.102}$$

Lemma 1.4.1, Assumption 1.4.3(i)(ii) and result (1.102) allow us to conclude by the portmanteau theorem that

$$\liminf_{n\to\infty} \pi_n(P_{\eta/r_n}) \geq P(\phi''_{\theta_0}(\mathbb{G} + \theta'_0(\eta h)) > c_{1-\alpha}) \ . \tag{1.103}$$

This establishes the first claim of the theorem.

For the second claim, note that if $\eta \leq 0$, then

$$0 = \lim_{n\to\infty} r_n^2 \{\phi(\theta(P_n)) - \phi(\theta(P))\} = \phi''_{\theta_0}(\theta'_0(\eta h)) \ , \tag{1.104}$$

where we exploited $\phi(\theta(P)) = \phi(\theta(P_n) = 0$ for all n and Assumption 1.2.3(iii). Hence,

$$\begin{aligned}
\limsup_{n\to\infty} \pi_n(P_{\eta/r_n}) &\equiv \limsup_{n\to\infty} P_n^n(r_n^2\phi(\hat{\theta}_n) > \hat{c}_{1-\alpha}) \\
&\leq \limsup_{n\to\infty} P_n^n(r_n^2\phi(\hat{\theta}_n) \geq \hat{c}_{1-\alpha}) \\
&\leq P(\phi''_{\theta_0}(\mathbb{G} + \theta'_0(\eta h)) \geq c_{1-\alpha}) \\
&= P(\tau \circ \phi''_{\theta_0}(\mathbb{G} + \theta'_0(\eta h)) \geq \tau(c_{1-\alpha})) \\
&\leq P(\tau \circ \phi''_{\theta_0}(\mathbb{G}) + \tau \circ \phi''_{\theta_0}(\theta'_0(\eta h)) \geq \tau(c_{1-\alpha})) \\
&= P(\phi''_{\theta_0}(\mathbb{G}) \geq c_{1-\alpha}) = \alpha \ , \qquad (1.105)
\end{aligned}$$

where the second inequality is due to the Lemma 1.4.1, result (1.102) and the portmanteau theorem, the second equality is by τ being strictly increasing, the third inequality is by $\tau \circ \phi''_{\theta_0}$ being subadditive, and the third equality is due to result (1.104), $\tau(0) = 0$ and τ being strictly increasing. This proves the second claim of the theorem. ■

Lemma 1.8.1. *Suppose that Assumptions 1.2.2 and 1.3.1(ii) hold, and that $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{E} \equiv \mathbf{R}$ is Hadamard differentiable at $\theta_0 \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$ with ϕ'_{θ_0} satisfying Assumption 1.2.3(iii). Then $\hat{c}_{1-\alpha} \xrightarrow{p} 0$, where for $\alpha \in (0,1)$,*

$$\hat{c}_{1-\alpha} \equiv \inf\{c \in \mathbf{R} : P_W(r_n\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\} \leq c) \geq 1 - \alpha\} \ .$$

Proof: This lemma is somewhat similar to Lemma 5 in Andrews and Guggenberger (2010) and we include the proof here only for completeness. Fix $\alpha \in (0,1)$ and let $c_{1-\alpha} \equiv \inf\{c \in \mathbf{R} : P(\phi'_{\theta_0}(\mathbb{G}) \leq c) \geq 1 - \alpha\}$. Note that $c_{1-\alpha} = 0$ for all $\alpha \in (0,1)$. Since ϕ is Hadamard differentiable at $\theta_0 \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$, it follows by Theorem 3.9.15 in van der Vaart and Wellner (1996a) that

$$\sup_{f \in \mathrm{BL}_1(\mathbb{D})} |E_W[f(r_n\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\})] - E[f(\phi'_{\theta_0}(\mathbb{G}))]| \xrightarrow{p} 0 \ . \qquad (1.106)$$

This, together with Lemma 10.11 in Kosorok (2008a), give us: for all $t \in \mathbf{R} \setminus \{0\}$,

$$P_W(r_n\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\} \leq t) \xrightarrow{p} P(\phi'_{\theta_0}(\mathbb{G}) \leq t) \ . \tag{1.107}$$

Fix $\epsilon > 0$. Clearly, $c_{1-\alpha} \pm \epsilon \in \mathbf{R} \setminus \{0\}$ for all $\epsilon > 0$ and all $\alpha \in (0,1)$. It follows from (1.107) that

$$\begin{aligned} &P_W(r_n\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\} \leq c_{1-\alpha} - \epsilon) \xrightarrow{p} P(\phi'_{\theta_0}(\mathbb{G}) \leq c_{1-\alpha} - \epsilon) = 0 < 1 - \alpha \ , \\ &P_W(r_n\{\phi(\hat{\theta}_n^*) - \phi(\hat{\theta}_n)\} \leq c_{1-\alpha} + \epsilon) \xrightarrow{p} P(\phi'_{\theta_0}(\mathbb{G}) \leq c_{1-\alpha} + \epsilon) = 1 > 1 - \alpha \ . \end{aligned} \tag{1.108}$$

By definition of $\hat{c}_{1-\alpha}$, it follows from (1.108) that

$$P(-\epsilon \leq \hat{c}_{1-\alpha} \leq \epsilon) = P(c_{1-\alpha} - \epsilon \leq \hat{c}_{1-\alpha} \leq c_{1-\alpha} + \epsilon) \to 1 \ . \tag{1.109}$$

Since ϵ is arbitrary, the conclusion of the lemma then follows from result (1.109). ■

Lemma 1.8.2. *Let Assumptions 1.2.1 and 1.2.3(i) hold, and $\hat{\phi}_n'' : \mathbb{D} \to \mathbb{E}$ be an estimator depending on $\{X_i\}_{i=1}^n$. Then the following are equivalent:*

(i) For every compact set $K \subset \mathbb{D}_0$ and every $\epsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P\Big(\sup_{h \in K^\delta} \|\hat{\phi}_n''(h) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} > \epsilon \Big) = 0 \ . \tag{1.110}$$

(ii) For every compact set $K \subset \mathbb{D}_0$, every $\delta_n \downarrow 0$ and every $\epsilon > 0$,

$$\limsup_{n \to \infty} P\Big(\sup_{h \in K^{\delta_n}} \|\hat{\phi}_n''(h) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} > \epsilon \Big) = 0 \ . \tag{1.111}$$

(iii) For every sequence $\{h_n\} \subset \mathbb{D}$ and every $h \in \mathbb{D}_0$ such that $h_n \to h$ as $n \to \infty$,

$$\hat{\phi}_n''(h_n) \xrightarrow{p} \phi_{\theta_0}''(h) \ . \tag{1.112}$$

Proof: The equivalence between (i) and (ii) is intuitive and straightforward to establish.

Suppose that (i) holds. Fix a compact set $K \subset \mathbb{D}_0$, a sequence $\{\delta_n\}$ with $\delta_n \downarrow 0$, and $\epsilon, \eta > 0$. We want to show that there exists some $N_0 > 0$ such that for all $n \geq N_0$,

$$P\Big(\sup_{h \in K^{\delta_n}} \|\hat{\phi}_n''(h) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} > \epsilon \Big) \leq \eta \ . \tag{1.113}$$

But from (i) we know that there is some $\delta_0 > 0$ such that

$$\limsup_{n \to \infty} P\Big(\sup_{h \in K^{\delta_0}} \|\hat{\phi}_n''(h) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} > \epsilon \Big) < \eta \ , \tag{1.114}$$

which in turn implies that there is some N_1 satisfying for all $n \geq N_1$

$$P\Big(\sup_{h \in K^{\delta_0}} \|\hat{\phi}_n''(h) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} > \epsilon \Big) < \eta \ . \tag{1.115}$$

Since $\delta_n \downarrow 0$, there exists some N_2 such that $\delta_n \leq \delta_0$ for all $n \geq N_2$ and hence

$$P\Big(\sup_{h \in K^{\delta_n}} \|\hat{\phi}_n''(h) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} > \epsilon \Big) \leq P\Big(\sup_{h \in K^{\delta_0}} \|\hat{\phi}_n''(h) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} > \epsilon \Big) \ . \tag{1.116}$$

Setting $N_0 \equiv \max\{N_1, N_2\}$, we see that (1.113) follows from (1.115) and (1.116).

Conversely, suppose that (ii) holds, fix a compact set $K \subset \mathbb{D}_0$ and $\epsilon > 0$, and we aim to establish (i) or equivalently, there exists some $\delta_0 > 0$ such that (1.115) holds. Pick a sequence $\delta_n \downarrow 0$. Then there exists some N_0 such that (1.113) holds with "$\leq$" replaced by "$<$". Setting $\delta_0 \equiv \delta_{N_0}$, we may then conclude (1.115) from (1.113).

Now suppose (ii) (and hence (i)) holds again and let $\{h_n\} \subset \mathbb{D}$ such that $h_n \to h \in \mathbb{D}_0$. Fix $\delta > 0$. There must be some N_1 such that $\|h_n - h\|_{\mathbb{D}} < \delta$ for all $n \geq N_1$. By the triangle inequality we have: for all $n \geq N_1$,

$$\begin{aligned} \|\hat{\phi}_n''(h_n) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} &\leq \|\hat{\phi}_n''(h_n) - \phi_{\theta_0}''(h_n)\|_{\mathbb{E}} + \|\phi_{\theta_0}''(h_n) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} \\ &\leq \sup_{h \in K^{\delta}} \|\hat{\phi}_n''(h) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} + \|\phi_{\theta_0}''(h_n) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} \ . \end{aligned} \tag{1.117}$$

Part (iii) then follows from (1.117), part (i) and Assumption 1.2.3(i).

Finally, suppose that (iii) holds. Fix a compact set $K \subset \mathbb{D}_0$ and $\epsilon > 0$. Let $\delta_n \downarrow 0$. Note that if $\sup_{h \in K^{\delta_n}} \|\hat{\phi}_n''(h) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} > \epsilon$, then there must exist some $h_n \in K^{\delta_n}$ such that $\|\hat{\phi}_n''(h_n) - \phi_{\theta_0}''(h_n)\|_{\mathbb{E}} > \epsilon$ and this is true for all $n \in \mathbf{N}$. It follows that

$$P\Big(\sup_{h \in K^{\delta_n}} \|\hat{\phi}_n''(h) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} > \epsilon \Big) \leq P(\|\hat{\phi}_n''(h_n) - \phi_{\theta_0}''(h_n)\|_{\mathbb{E}} > \epsilon) \ . \tag{1.118}$$

Note that $h_n \in K^{\delta_n}$ is possibly random and satisfies $d(h_n, K) \equiv \inf_{a \in K} \|h_n - a\|_{\mathbb{D}} \leq \delta_n \to 0$ as $n \to \infty$. Fix an arbitrary subsequence $\{n_k\}$. Since K is compact, it follows by Lemma A.6 in Fang (2016) that there exists a further subsequence $\{n_{k_j}\}$ and some deterministic $h \in K$ such that $h_{n_{k_j}} \xrightarrow{p} h$ as $j \to \infty$. By the triangle inequality,

$$\begin{aligned} P(\|\hat{\phi}_n''(h_n) - \phi_{\theta_0}''(h_n)\|_{\mathbb{E}} > \epsilon) \leq & P(\|\hat{\phi}_n''(h_n) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} > \frac{\epsilon}{2}) \\ & + P(\|\phi_{\theta_0}''(h_n) - \phi_{\theta_0}''(h)\|_{\mathbb{E}} > \frac{\epsilon}{2}) \ . \end{aligned} \tag{1.119}$$

Since $h_{n_{k_j}} \xrightarrow{p} h$ as $j \to \infty$, the first term on the right hand side above tends to zero along $\{n_{k_j}\}$ by (iii) and Lemma 1.8.3, while the second term tends to zero along $\{n_{k_j}\}$ by Theorem 1.9.5 in van der Vaart and Wellner (1996a) and Assumption 1.2.3(i). Since $\{n_k\}$ is arbitrary, combination of results (1.118) and (1.119) then leads to (ii). ■

Lemma 1.8.3 (Extended Continuous Mapping Theorem)**.** *Let $\mathbb{D}$ and $\mathbb{E}$ be metric spaces equipped with metrics d and ρ respectively, $g_n : \mathbb{D}_n \subset \mathbb{D} \to \mathbb{E}$ a possibly random map for each $n \in \mathbf{N}$, and $g : \mathbb{D}_0 \subset \mathbb{D} \to \mathbb{E}$ a nonrandom map. Suppose that $g_n(x_n) \xrightarrow{p} g(x)$ whenever $x_n \to x$ for $x_n \in \mathbb{D}_n$ and $x \in \mathbb{D}_0$. If $X_n \xrightarrow{p} X$ such that X is Borel measurable, separable and satisfies $P(X \in \mathbb{D}_0) = 1$, then $g_n(X_n) \xrightarrow{p} g(X)$.*

Proof: We closely follow the proof of Proposition A.8.6 in Bickel et al. (1998) (see also van der Vaart and Wellner (1990)). Fix $\epsilon > 0$ throughout. First, we show that $g : \mathbb{D}_0 \to \mathbb{E}$ is continuous. By assumption, for each $x \in \mathbb{D}_0$ we have

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P(\mathrm{Osc}_{g_n}(B(x, \delta)) > \epsilon) = 0 \ , \tag{1.120}$$

where $\mathrm{Osc}_{g_n}(B(x,\delta)) \equiv \sup_{y,z\in B(x,\delta)} \rho(g_n(y), g_n(z))$ for $B(x,\delta) \equiv \{y \in \mathbb{D}_n : d(y,x) < \delta\}$. This can be easily seen by the triangle inequality:

$$\begin{aligned}\lim_{\delta\downarrow 0}\limsup_{n\to\infty} P(\mathrm{Osc}_{g_n}(B(x,\delta)) > \epsilon) \leq & \lim_{\delta\downarrow 0}\limsup_{n\to\infty} P(\sup_{y\in B(x,\delta)} \rho(g_n(y), g(x)) > \frac{\epsilon}{2}) \\ & + \lim_{\delta\downarrow 0}\limsup_{n\to\infty} P(\sup_{z\in B(x,\delta)} \rho(g_n(z), g(x)) > \frac{\epsilon}{2}) \\ & = 0 \ .\end{aligned}$$

Notice that again by assumption, the triangle inequality and result (1.120) we have

$$\begin{aligned}\rho(g(y), g(x)) &\leq \rho(g(y), g_n(y)) + \rho(g(x), g_n(x)) + \rho(g_n(y), g_n(x)) \\ &\leq \rho(g(y), g_n(y)) + \rho(g(x), g_n(x)) + \mathrm{Osc}_{g_n}(B(x, d(x,y))) \\ &\xrightarrow{p} 0 \ , \end{aligned} \tag{1.121}$$

as $n \to \infty$ followed by $d(x,y) \to \infty$. Since g is a nonrandom function, we must have $\rho(g(y), g(x)) \to 0$ as $d(y,x) \to 0$ and hence g is continuous on $\mathbb{D}_0$.

Next, for $x \in \mathbb{D}_0$ define

$$k(x,\epsilon) \equiv \min\{k : \text{for } \forall\, y \text{ with } d(y,x) < \frac{1}{k} \text{ and all } n \geq k, P(\rho(g_n(y), g(x)) \leq \epsilon) \geq 1-\epsilon\} \ .$$

This is well defined by a simple *reductio ad absurdum* argument as in Bickel et al. (1998). We now show that $k(\cdot,\epsilon) : \mathbb{D}_0 \to \mathbf{N}$ is measurable. This is done by proving that $k(\cdot,\epsilon)$ is lower semicontinuous, i.e., $x_m \to x$ for $\{x, x_m\} \subset \mathbb{D}_0$ implies

$$\liminf_{m\to\infty} k(x_m,\epsilon) \geq k(x,\epsilon) \ . \tag{1.122}$$

Fix $x \in \mathbb{D}_0$ and $\{x_m\} \subset \mathbb{D}_0$ such that $x_m \to x$ as $m \to \infty$. Then there must exist some subsequence $\{m'\}$ of $\{m\}$ such that $\liminf_{m\to\infty} k(x_m,\epsilon) = \lim_{m'\to\infty} k(x_{m'},\epsilon)$. Since $k(\cdot,\epsilon)$ is integer valued, we further have $\liminf_{m\to\infty} k(x_m,\epsilon) = k(x_{m'},\epsilon) \equiv k'$ for all m' sufficiently large. If $k' = \infty$, then the inequality (1.122) follows trivially. Otherwise, suppose that

$k' < \infty$. For any y with $d(x,y) < 1/k'$, there exists an m_0 such that $d(x_{m'}, y) < 1/k'$ for all $m' \geq m_0$. By definition of $k(x,\epsilon)$, it follows that for all $n \geq k'$,

$$P(\rho(g_n(y), g(x_{m'})) \leq \epsilon) \geq 1 - \epsilon \ . \tag{1.123}$$

Letting $m' \uparrow \infty$, we have by $x_{m'} \to x$ and continuity of g and P that for all $n \geq k'$,

$$P(\rho(g_n(y), g(x)) \leq \epsilon) \geq 1 - \epsilon \ . \tag{1.124}$$

Thus we must have $k(x,\epsilon) \leq k' = \liminf_{m\to\infty} k(x_m, \epsilon)$ and hence $k(\cdot,\epsilon)$ is Borel measurable.

Since $P(X \in \mathbb{D}_0) = 1$, we may assume without loss of generality that X takes values in $\mathbb{D}_0$. In turn, it follows that $k(X,\epsilon)$ is a Borel $\mathbf{N}$-valued random variable. Thus there exists some $k_0 \equiv k_0(\epsilon)$ such that

$$P(k(X,\epsilon) > k_0) < \epsilon \ . \tag{1.125}$$

Since $X_n \xrightarrow{p} X$, there exists some $n_0 \equiv n_0(\epsilon)$ such that for all $n \geq n_0(\epsilon)$,

$$P(d(X_n, X) > \frac{1}{k_0}) < \epsilon \ . \tag{1.126}$$

Now define

$$B_n \equiv \{\rho(g_n(X_n), g(X)) > \epsilon\} \ , \ C_n \equiv \{d(X_n, X) > \frac{1}{k_0}\} \ , \ D \equiv \{k(X,\epsilon) > k_0\} \ .$$

It follows that for all $n \geq \max\{n_0, k_0\}$,

$$\begin{aligned} P(B_n) &\leq P(B_n \cap (C_n^c \cap D^c)) + P(B_n \cap (C_n^c \cap D^c)^c) \\ &\leq P(B_n \cap (C_n^c \cap D^c)) + P(C_n) + P(D) \leq 3\epsilon \ , \end{aligned}$$

by definition of $k(x,\epsilon)$, results (1.125) and (1.126), and we are done since ϵ is arbitrary. ■

1.8.2 Results for Examples 1.2.1 - 1.2.6

Example 1.2.2: Moment Inequalities

In this example, it is a simple exercise to show that

$$\phi'_\theta(h) = \begin{cases} 2\theta h & \text{if } \theta > 0 \\ \\ 0 & \text{if } \theta \leq 0 \end{cases}, \; \phi''_\theta(h) = \begin{cases} h^2 & \text{if } \theta > 0 \\ \\ (\max\{h,0\})^2 & \text{if } \theta = 0 \\ \\ 0 & \text{if } \theta < 0 \end{cases} . \tag{1.127}$$

Thus, ϕ is Hadamard differentiable with the derivative ϕ'_θ degenerate at $\theta \leq 0$ and in particular at the "least favorable point" $\theta = 0$. Moreover, ϕ_θ is second order Hadamard directionally differentiable. The derivative ϕ''_θ is nondegenerate at 0, though degenerate whenever $\theta < 0$.

Exploiting the structure in (1.127), we may easily estimate the derivative by

$$\hat{\phi}''_n(h) = \begin{cases} h^2 & \text{if } \overline{X}_n > \kappa_n \\ \\ (\max\{h,0\})^2 & \text{if } |\overline{X}_n| \leq \kappa_n \\ \\ 0 & \text{if } \overline{X}_n < \kappa_n \end{cases} , \tag{1.128}$$

where $\kappa_n \downarrow 0$ satisfies $\sqrt{n}\kappa_n \uparrow \infty$, and $\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^n X_i$. Interestingly, construction of $\hat{\phi}''_n$ as above amounts to the generalized moment selection procedure as in Andrews and Soares (2010) for conducting inference in moment inequalities models.

Example 1.2.3: Cramer-von Mises Functionals

Cramer-von Mises functionals can be viewed as generalized Wald functionals. It is straightforward to show that ϕ is first and second Hadamard differentiable at any $\theta \in$

$\ell^\infty(\mathbf{R}^{d_x})$ with derivatives satisfying:

$$\phi'_\theta(h) = 2\int (\theta - F_0)h\, dF_0 \ , \ \phi''_\theta(h) = \int h^2\, dF_0 \ ,$$

for all $h \in \ell^\infty(\mathbf{R}^{d_x})$. Note that first order derivative ϕ'_θ is degenerate when $\theta = F_0$, while second order derivative ϕ''_θ is nowhere degenerate. The corresponding bilinear map $\Phi''_\theta : \ell^\infty(\mathbf{R}^{d_x}) \times \ell^\infty(\mathbf{R}^{d_x}) \to \mathbf{R}$ is given by $\Phi''_\theta(h, g) = \int hg\, dF_0$. In this example, there is no need for derivative estimation because ϕ''_{θ_0} is a known map.

Example 1.2.4: Stochastic Dominance

Lemma 1.8.4. *Let $w : \mathbf{R} \to \mathbf{R}^+$ satisfy $\int_{\mathbf{R}} w(u)du < \infty$ and $\phi : \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R}) \to \mathbf{R}$ be given by $\phi(\theta) = \int_{\mathbf{R}} \max\{\theta^{(1)}(u) - \theta^{(2)}(u), 0\}^2 w(u)du$ for any $\theta = (\theta^{(1)}, \theta^{(2)}) \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$. Then it follows that*

(i) ϕ is first order Hadamard differentiable at any $\theta \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ with $\phi'_\theta : \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R}) \to \mathbf{R}$ satisfying for any $h = (h^{(1)}, h^{(2)}) \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$

$$\phi'_\theta(h) = 2\int_{B_+(\theta)} [\theta^{(1)}(u) - \theta^{(2)}(u)][h^{(1)}(u) - h^{(2)}(u)]w(u)du \ ,$$

where $B_+(\theta) \equiv \{u \in \mathbf{R} : \theta^{(1)}(u) > \theta^{(2)}(u)\}$.

(ii) ϕ is second order Hadamard directionally differentiable at any $\theta \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ and the derivative $\phi''_\theta : \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R}) \to \mathbf{R}$ is given by: for any $h = (h^{(1)}, h^{(2)}) \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$

$$\phi''_\theta(h) = \int_{B_0(\theta)} \max\{h^{(1)}(u) - h^{(2)}(u), 0\}^2 w(u)du + \int_{B_+(\theta)} [h^{(1)}(u) - h^{(2)}(u)]^2 w(u)du \ ,$$

where $B_0(\theta) \equiv \{u \in \mathbf{R} : \theta^{(1)}(u) = \theta^{(2)}(u)\}$.

PROOF: Fix $\theta \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$. Further, let $t_n \downarrow 0$, $\{h_n\} = \{(h_n^{(1)}, h_n^{(2)})\}$ be a sequence in $\ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ satisfying $\|h_n^{(1)} - h^{(1)}\|_\infty \vee \|h_n^{(2)} - h^{(2)}\|_\infty = o(1)$ for some $h = (h^{(1)}, h^{(2)}) \in$

$\ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$, and

$$B_-(\theta) \equiv \{u \in \mathbf{R} : \theta^{(1)}(u) < \theta^{(2)}(u)\} \ .$$

Observe that since $\theta^{(1)}(u) - \theta^{(2)}(u) < 0$ for all $u \in B_-(\theta)$, and $\|h_n^{(1)} - h_n^{(2)}\|_\infty = O(1)$ due to $\|h^{(1)} - h^{(2)}\|_\infty < \infty$, the dominated convergence theorem yields that:

$$\lim_{n\to\infty} \frac{1}{t_n} \int_{B_-(\theta)} \max\{(\theta^{(1)}(u) - \theta^{(2)}(u)) + t_n(h_n^{(1)}(u) - h_n^{(2)}(u)), 0\}^2 w(u)du = 0 \ , \quad (1.129)$$

$$\lim_{n\to\infty} \frac{1}{t_n} \int_{B_0(\theta)} \max\{(\theta^{(1)}(u) - \theta^{(2)}(u)) + t_n(h_n^{(1)}(u) - h_n^{(2)}(u)), 0\}^2 w(u)du = 0 \ , \quad (1.130)$$

and

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{t_n} \Big[\int_{B_+(\theta)} \max\{(\theta^{(1)}(u) - \theta^{(2)}(u)) + t_n(h_n^{(1)}(u) - h_n^{(2)}(u)), 0\}^2 w(u)du \\
&\qquad\qquad - \int_{B_+(\theta)} (\theta^{(1)}(u) - \theta^{(2)}(u))^2 w(u)du \Big] \\
&= \lim_{n\to\infty} \int_{B_+(\theta)} \frac{1}{t_n} \big[\max\{(\theta^{(1)}(u) - \theta^{(2)}(u)) + t_n(h_n^{(1)}(u) - h_n^{(2)}(u)), 0\}^2 \\
&\qquad\qquad - (\theta^{(1)}(u) - \theta^{(2)}(u))^2 \big] w(u)du \\
&= 2 \int_{B_+(\theta)} [\theta^{(1)}(u) - \theta^{(2)}(u)][h^{(1)}(u) - h^{(2)}(u)] w(u)du \ . \qquad (1.131)
\end{aligned}$$

Combining results (1.129) - (1.131) yields

$$\phi'_\theta(h) \equiv \lim_{n\to\infty} \frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} = 2 \int_{B_+(\theta)} [\theta^{(1)}(u) - \theta^{(2)}(u)][h^{(1)}(u) - h^{(2)}(u)] w(u)du \ ,$$

which establishes the first claim of the lemma.

Next fix $\theta \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ and let $\{h_n\}$ and $\{t_n\}$ be as before. Therefore, by the

dominated convergence theorem we have

$$\lim_{n\to\infty}\int_{B_-(\theta)}\frac{1}{t_n^2}\max\{(\theta^{(1)}(u)-\theta^{(2)}(u))+t_n(h_n^{(1)}(u)-h_n^{(2)}(u)),0\}^2w(u)du=0\ , \quad (1.132)$$

$$\lim_{n\to\infty}\int_{B_0(\theta)}\frac{1}{t_n^2}\max\{(\theta^{(1)}(u)-\theta^{(2)}(u))+t_n(h_n^{(1)}(u)-h_n^{(2)}(u)),0\}^2w(u)du$$
$$=\int_{B_0(\theta)}\max\{h^{(1)}(u)-h^{(2)}(u),0\}^2w(u)du\ , \quad (1.133)$$

and

$$\lim_{n\to\infty}\frac{1}{t_n^2}\Big[\int_{B_+(\theta)}\max\{(\theta^{(1)}(u)-\theta^{(2)}(u))+t_n(h_n^{(1)}(u)-h_n^{(2)}(u)),0\}^2w(u)du$$
$$-\int_{B_+(\theta)}(\theta^{(1)}(u)-\theta^{(2)}(u))^2w(u)du-t_n2\int_{B_+(\theta)}[\theta^{(1)}(u)-\theta^{(2)}(u)][h_n^{(1)}(u)-h_n^{(2)}(u)]w(u)du\Big]$$
$$\to\int_{B_+(\theta)}[h^{(1)}(u)-h^{(2)}(u)]^2w(u)du\ . \quad (1.134)$$

It follows from results (1.132)-(1.134) that

$$\phi''_\theta(h)\equiv\lim_{n\to\infty}\frac{\phi(\theta+t_nh_n)-\phi(\theta)-t_n\phi'_\theta(h_n)}{t_n^2}$$
$$=\int_{B_0(\theta)}\max\{h^{(1)}(u)-h^{(2)}(u),0\}^2w(u)du+\int_{B_+(\theta)}[h^{(1)}(u)-h^{(2)}(u)]^2w(u)du\ .$$

This competes the proof of the second claim and we are done. ■

Example 1.2.5: Conditional Moment Inequalities

Lemma 1.8.5. *Let $\phi:\ell^\infty(\mathcal{F})\times\ell^\infty(\mathcal{F})\to\mathbf{R}$ be given by $\phi(\theta)=\sup_{f\in\mathcal{F}}\{[\max(\theta^{(1)}(f),0)]^2+[\theta^{(2)}(f)]^2\}$ where $\mathcal{F}$ is compact under some metric d. Then it follows that:*

(i) ϕ is Hadamard differentiable at any $\theta\in\ell^\infty(\mathcal{F})\times\ell^\infty(\mathcal{F})$ satisfying $\theta^{(1)}\le 0$ and $\theta^{(2)}=0$, and its derivative $\phi'_\theta(h)=0$ for any $h\in\ell^\infty(\mathcal{F})\times\ell^\infty(\mathcal{F})$

(ii) ϕ is second order Hadamard directionally differentiable at any $\theta\in C(\mathcal{F})\times C(\mathcal{F})$ satisfying $\theta^{(1)}\le 0$ and $\theta^{(2)}=0$ tangentially to $C(\mathcal{F})\times C(\mathcal{F})$, and the derivative is

given by: for any $h \in C(\mathcal{F}) \times C(\mathcal{F})$,

$$\phi''_\theta(h) = \max\{ \sup_{f \in \mathcal{F}_0} \{\max(h^{(1)}(f), 0)^2 + [h^{(2)}(f)]^2\}, \sup_{f \in \mathcal{F} \setminus \mathcal{F}^0} [h^{(2)}(f)]^2\} \ ,$$

where $\mathcal{F}_0 \equiv \{f \in \mathcal{F} : \theta^{(1)}(f) = 0\}$, *and* $\sup \emptyset \equiv 0$.

Remark 1.8.1. Note that if $\mathcal{F}_0 = \emptyset$, then ϕ''_θ simplifies to $\phi''_\theta(h) = \sup_{f \in \mathcal{F}} [h^{(2)}(f)]^2$. ■

PROOF: Let $\theta \in \ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$ satisfying $\theta^{(1)} \leq 0$ and $\theta^{(2)} = 0$, $\{h_n\} \subset \ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$ such that $h_n \to h \in \ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$, and $t_n \downarrow 0$. Combining $\theta^{(1)} \leq 0$, $\theta^{(2)} = 0$ and the triangle inequality, we have

$$\begin{aligned} |\phi(\theta + t_n h_n) - \phi(\theta)| &= \big| \sup_{f \in \mathcal{F}} \{[\max(\theta^{(1)}(f) + t_n h_n^{(1)}(f), 0)]^2 + [\theta^{(2)}(f) + t_n h_n^{(2)}(f)]^2\} \big| \\ &\leq \sup_{f \in \mathcal{F}} [\max(\theta^{(1)}(f) + t_n h_n^{(1)}(f), 0)]^2 + t_n^2 \sup_{f \in \mathcal{F}} [h_n^{(2)}(f)]^2 \\ &\leq \sup_{f \in \mathcal{F}} [\max(t_n h_n^{(1)}(f), 0)]^2 + t_n^2 \sup_{f \in \mathcal{F}} [h_n^{(2)}(f)]^2 = o(t_n) \ , \end{aligned} \tag{1.135}$$

as desired in part (i), where in the last step we used the fact that $h_n^{(1)} = h_n^{(2)} = O(1)$.

As for the second claim, let $\theta \in C(\mathcal{F}) \times C(\mathcal{F})$ satisfying $\theta^{(1)} \leq 0$ and $\theta^{(2)} = 0$, $\{h_n\} \subset \ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$ such that $h_n \to h \in C(\mathcal{F}) \times C(\mathcal{F})$, and $t_n \downarrow 0$. By $\theta^{(1)} \leq 0$ and $\theta^{(2)} = 0$, Lipschtiz continuity of the sup operator and the triangle inequality we have

$$\begin{aligned} &|\phi(\theta + t_n h_n) - \phi(\theta + t_n h)| \\ &\qquad = |\sup_{f \in \mathcal{F}} \{\max(\theta^{(1)}(f) + t_n h_n^{(1)}(f), 0)^2 + [t_n h_n^{(2)}(f)]^2\} \\ &\qquad\quad - \sup_{f \in \mathcal{F}} \{\max(\theta^{(1)}(f) + t_n h^{(1)}(f), 0)^2 + [t_n h^{(2)}(f)]^2\}| \\ &\qquad \leq \sup_{f \in \mathcal{F}} |\max(\theta^{(1)}(f) + t_n h_n^{(1)}(f), 0)^2 - \max(\theta^{(1)}(f) + t_n h^{(1)}(f), 0)^2| \\ &\qquad\quad + \sup_{f \in \mathcal{F}} |[t_n h_n^{(2)}(f)]^2 - [t_n h^{(2)}(f)]^2| \ . \end{aligned} \tag{1.136}$$

Since $\|h_n - h\|_\infty = o(1)$ and $\theta^{(1)} \leq 0$, it follows that

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} |\max(\theta^{(1)}(f) + t_n h_n^{(1)}(f), 0)^2 - \max(\theta^{(1)}(f) + t_n h^{(1)}(f), 0)^2| \\
&\quad \leq \sup_{f\in\mathcal{F}} |\max(\theta^{(1)}(f) + t_n h_n^{(1)}(f), 0) - \max(\theta^{(1)}(f) + t_n h^{(1)}(f), 0)| \\
&\qquad \times \sup_{f\in\mathcal{F}} |\max(\theta^{(1)}(f) + t_n h_n^{(1)}(f), 0) + \max(\theta^{(1)}(f) + t_n h^{(1)}(f), 0)| \\
&\quad \leq \sup_{f\in\mathcal{F}} |t_n h_n^{(1)}(f) - t_n h^{(1)}(f)| \sup_{f\in\mathcal{F}} \{\max(t_n h_n^{(1)}(f), 0) + \max(t_n h^{(1)}(f), 0)\} \\
&\quad = o(t_n)O(t_n) = o(t_n^2) \ , \qquad (1.137)
\end{aligned}$$

and that

$$\sup_{f\in\mathcal{F}} |[t_n h_n^{(2)}(f)]^2 - [t_n h^{(2)}(f)]^2| = o(t_n^2) \ . \qquad (1.138)$$

Combination of results (1.136), (1.137) and (1.138) leads to

$$|\phi(\theta + t_n h_n) - \phi(\theta + t_n h)| = o(t_n^2) \ . \qquad (1.139)$$

Next, fix $\delta > 0$. By definition of $\mathcal{F}_0^\delta$, compactness of $\mathcal{F}$ and continuity of $\theta^{(1)}$, we see that $\sup_{f\in\mathcal{F}\setminus\mathcal{F}_0^\delta} \theta^{(1)}(f) < 0$. Since also $t_n h^{(1)} = o(1)$ and $h^{(1)} \in C(\mathcal{F})$, it follows that $\theta^{(1)}(f) + t_n h^{(1)}(f) < 0$ for all $f \in f \in \mathcal{F} \setminus \mathcal{F}_0^\delta$ and for all n large. In turn we have

$$\begin{aligned}
&\lim_{\delta\downarrow 0} \lim_{n\to\infty} t_n^{-2} \sup_{f\in\mathcal{F}\setminus\mathcal{F}_0^\delta} \{\max(\theta^{(1)}(f) + t_n h^{(1)}(f), 0)^2 + [t_n h^{(2)}(f)]^2\} \\
&\qquad = \lim_{\delta\downarrow 0} \sup_{f\in\mathcal{F}\setminus\mathcal{F}_0^\delta} [h^{(2)}(f)]^2 = \sup_{f\in\mathcal{F}\setminus\mathcal{F}_0} [h^{(2)}(f)]^2 \ , \qquad (1.140)
\end{aligned}$$

where the last step is due to $h^{(2)} \in C(\mathcal{F})$. On the other hand, we have,

$$\begin{aligned}
&\lim_{\delta\downarrow 0}\limsup_{n\to\infty} t_n^{-2}\Big|\sup_{f\in\mathcal{F}_0^\delta}\{\max(\theta^{(1)}(f)+t_nh^{(1)}(f),0)^2+[t_nh^{(2)}(f)]^2\}\\
&\qquad\qquad -t_n^2\sup_{f\in\mathcal{F}_0}\{\max(h^{(1)}(f),0)^2+[h^{(2)}(f)]^2\}\Big|\\
&\le \lim_{\delta\downarrow 0}\limsup_{n\to\infty} t_n^{-2}\sup_{f\in\mathcal{F}_0^\delta}\{\max(t_nh^{(1)}(f),0)^2+[t_nh^{(2)}(f)]^2\}\\
&\qquad\qquad -\sup_{f\in\mathcal{F}_0}\{\max(t_nh^{(1)}(f),0)^2+[t_nh^{(2)}(f)]^2\}\\
&\le \lim_{\delta\downarrow 0}\limsup_{n\to\infty} t_n^{-2}\sup_{f,g\in\mathcal{F}:d(f,g)\le\delta}|\max(t_nh^{(1)}(f),0)^2-\max(t_nh^{(1)}(g),0)^2|\\
&\le \lim_{\delta\downarrow 0}\sup_{f,g\in\mathcal{F}:d(f,g)\le\delta}|\max(h^{(1)}(f),0)^2-\max(h^{(1)}(g),0)^2|=0\ ,
\end{aligned}\tag{1.141}$$

where the first inequality is due to $\theta(f)=0$ for all $f\in\mathcal{F}_0$ and $\theta^{(1)}\le 0$, the second inequality exploits the definition and compactness of $\mathcal{F}_0^\delta$, and the equality is due to uniform continuity of $h^{(1)}$ on $\mathcal{F}$ since $h^{(1)}\in C(\mathcal{F})$ and $\mathcal{F}$ is compact.

Finally, combining results (1.140), (1.141), and $\phi(\theta)=0$ we have:

$$\begin{aligned}
\limsup_{n\to\infty} t_n^{-2}\{\phi(\theta+t_nh)-\phi(\theta)\} &= \limsup_{n\to\infty} t_n^{-2}\phi(\theta+t_nh)\\
&= \limsup_{n\to\infty} t_n^{-2}\sup_{f\in\mathcal{F}}\{\max(\theta^{(1)}(f)+t_nh^{(1)}(f),0)^2+[t_nh^{(2)}(f)]^2\}\\
&= \lim_{\delta\downarrow 0}\limsup_{n\to\infty} t_n^{-2}\max\Big\{\sup_{f\in\mathcal{F}_0^\delta}\{\max(\theta^{(1)}(f)+t_nh^{(1)}(f),0)^2+[t_nh^{(2)}(f)]^2\},\\
&\qquad\qquad \sup_{f\in\mathcal{F}\setminus\mathcal{F}_0^\delta}\{\max(\theta^{(1)}(f)+t_nh^{(1)}(f),0)^2+[t_nh^{(2)}(f)]^2\}\Big\}\\
&= \max\Big\{\sup_{f\in\mathcal{F}_0}\{\max(h^{(1)}(f),0)^2+[h^{(2)}(f)]^2\},\sup_{f\in\mathcal{F}\setminus\mathcal{F}_0}[h^{(2)}(f)]^2\Big\}\ .
\end{aligned}\tag{1.142}$$

It follows from $\phi'_\theta=0$, (1.139) and (1.142) that

$$\begin{aligned}
&\lim_{n\to\infty}\frac{\phi(\theta+t_nh_n)-\phi(\theta)-t_n\phi'_\theta(h_n)}{t_n^2}\\
&\qquad = \max\Big\{\sup_{f\in\mathcal{F}\setminus\mathcal{F}_0}[h^{(2)}(f)]^2,\sup_{f\in\mathcal{F}_0}\{\max(h^{(1)}(f),0)^2+[h^{(2)}(f)]^2\}\Big\}\ ,
\end{aligned}\tag{1.143}$$

as desired for the second claim of the lemma. ■

Suppose that $\hat{\mathcal{F}}_0$ and $\hat{\mathcal{F}}_{0,c}$ are estimators of $\mathcal{F}_0 \equiv \{f \in \mathcal{F} : \theta_0^{(1)}(f) = 0\}$ and $\mathcal{F} \setminus \mathcal{F}^0$ that satisfy[17]

$$d_H(\hat{\mathcal{F}}_0, \mathcal{F}_0; L^2(W)) = o_p(1) \text{ and } d_H(\hat{\mathcal{F}}_{0,c}, \mathcal{F} \setminus \mathcal{F}_0; L^2(W)) = o_P(1) \ . \tag{1.144}$$

Based on $\hat{\mathcal{F}}_0$ and $\hat{\mathcal{F}}_{0,c}$ and in view of Lemma B.3 in Fang and Santos (2015), we may estimate the derivative as follows:

$$\hat{\phi}_n''(h) = \max\{\sup_{f \in \hat{\mathcal{F}}_0} \{\max(h^{(1)}(f), 0)^2 + [h^{(2)}(f)]^2\}, \sup_{f \in \hat{\mathcal{F}}_{0,c}} [h^{(2)}(f)]^2\} \ . \tag{1.145}$$

The estimation of $\mathcal{F}_0$ and $\mathcal{F} \setminus \mathcal{F}^0$ is in accordance with the generalized moment selection in Andrews and Shi (2013); see also Kaido and Santos (2014).

Example 1.2.6: Overidentification Test

Lemma 1.8.6. *Let $\Gamma \subset \mathbf{R}^k$ be a compact set, and $\phi : \prod_{j=1}^m \ell^\infty(\Gamma) \to \mathbf{R}$ be given by $\phi(\theta) = \inf_{\gamma \in \Gamma} \theta(\gamma)^\intercal W \theta(\gamma)$ where $\theta \in \prod_{j=1}^m \ell^\infty(\Gamma)$ and W is a $m \times m$ symmetric positive definite matrix. Then we have*

(i) ϕ is Hadamard differentiable at any $\theta \in \prod_{j=1}^m \ell^\infty(\Gamma)$ satisfying $\theta(\gamma) = 0$ for some $\gamma \in \Gamma$ with the derivative given by $\phi'_\theta(h) = 0$ for all $h \in \prod_{j=1}^m \ell^\infty(\Gamma)$.

(ii) If $\Gamma_0(\theta) \equiv \{\gamma \in \Gamma : \theta(\gamma) = 0\}$ is in the interior of Γ, $\theta \in \prod_{j=1}^m C^1(\Gamma)$ satisfies $\phi(\theta) = 0$, and for some small $\epsilon > 0$, $\inf_{\gamma \in \Gamma \setminus \Gamma_0(\theta)^\epsilon} \|\theta(\gamma)\| \geq C\epsilon^\kappa$ for some $\kappa \in (0, 1]$ and some $C > 0$, then ϕ is second order Hadamard directionally differentiable at θ tangentially to $\prod_{j=1}^m C(\Gamma)$ with the derivative given by: for any $h \in \prod_{j=1}^m C(\Gamma)$

$$\phi''_\theta(h) = \min_{\gamma_0 \in \Gamma_0(\theta)} \min_{v \in \mathbf{R}^k} \{h(\gamma_0) - J(\gamma_0)v\}^\intercal W \{h(\gamma_0) - J(\gamma_0)v\} \ ,$$

[17] We note that for two generic sets A and B in a metric space, neither $d_H(A, B)$ controls $d_H(A^c, B^c)$ nor $d_H(A^c, B^c)$ controls $d_H(A, B)$ (Lemenant et al., 2014).

where $J : \Gamma_0(\theta) \to \mathbf{M}^{m\times k}$ *is the Jacobian matrix defined by* $J(\gamma_0) \equiv \frac{d\theta(\gamma)}{d\gamma^\intercal}\big|_{\gamma=\gamma_0}$.

PROOF: Fix $\theta \in \prod_{j=1}^m \ell^\infty(\Gamma)$ and let $t_n \downarrow 0$ and $\{h_n, h\} \subset \prod_{j=1}^m \ell^\infty(\Gamma)$ such that $h_n \to h$. For a vector $a \in \mathbf{R}^m$, define the norm $\|a\|_W = \sqrt{a^\intercal W a}$. It follows that

$$\begin{aligned}|\phi(\theta + t_n h_n) - \phi(\theta)| &= \inf_{\gamma\in\Gamma} \|\theta(\gamma) + t_n h_n(\gamma)\|_W^2 \\ &\leq \inf_{\gamma\in\Gamma_0(\theta)} \|\theta(\gamma) + t_n h_n(\gamma)\|_W^2 \leq t_n^2 \inf_{\gamma\in\Gamma_0(\theta)} \|h_n(\gamma)\|_W^2 = o(t_n) \ , \end{aligned} \tag{1.146}$$

where the second inequality is because $\theta(\gamma) = 0$ for all $\gamma \in \Gamma_0(\theta)$ and the last step is due to $h_n = O(1)$ by assumption. This establishes part (i).

For part (ii), fix $\theta \in \prod_{j=1}^m C^1(\Gamma)$ with $\phi(\theta) = 0$ and let $t_n \downarrow 0$ and $\{h_n\} \subset \prod_{j=1}^m \ell^\infty(\Gamma)$ such that $h_n \to h \in \prod_{j=1}^m C(\Gamma)$. First of all, note that for $\gamma_0 \in \Gamma_0(\theta)$,

$$\begin{aligned}|\phi(\theta + t_n h_n) - \phi(\theta + t_n h)| &= \big| \inf_{\gamma\in\Gamma} \|\theta(\gamma) + t_n h_n(\gamma)\|_W^2 - \inf_{\gamma\in\Gamma} \|\theta(\gamma) + t_n h(\gamma)\|_W^2 \big| \\ &= \big| \inf_{\gamma\in\Gamma} \|\theta(\gamma) + t_n h_n(\gamma)\|_W - \inf_{\gamma\in\Gamma} \|\theta(\gamma) + t_n h(\gamma)\|_W \big| \\ &\quad \times \big| \inf_{\gamma\in\Gamma} \|\theta(\gamma) + t_n h_n(\gamma)\|_W + \inf_{\gamma\in\Gamma} \|\theta(\gamma) + t_n h(\gamma)\|_W \big| \\ &\leq t_n \|h_n - h\|_\infty \{\|\theta(\gamma_0) + t_n h_n(\gamma_0)\|_W + \|\theta(\gamma_0) + t_n h(\gamma_0)\|_W\} \\ &\leq t_n^2 \|h_n - h\|_\infty \{\|h_n(\gamma_0)\|_W + \|h(\gamma_0)\|_W\} = o(t_n^2) \ , \end{aligned} \tag{1.147}$$

where the first inequality is due to the Lipschitz continuity of the inf operator and the triangle inequality, and the last inequality follows from $h_n \to h$ and $\theta(\gamma_0) = 0$ for $\gamma_0 \in \Gamma_0(\theta)$.

Next, for each fixed $a \geq (2C^{-1} \max_{\gamma\in\Gamma} \|h(\gamma)\|_W)^{1/\kappa}$, by assumption and the triangle

inequality we have: for all n sufficiently large so that $t_n^\kappa \geq t_n$,

$$\begin{aligned}\inf_{\gamma\in\Gamma\setminus\Gamma_0(\theta)^{at_n}} \|\theta(\gamma)+t_nh(\gamma)\|_W &\geq \inf_{\gamma\in\Gamma\setminus\Gamma_0(\theta)^{at_n}} \|\theta(\gamma)\|_W - t_n \sup_{\gamma\in\Gamma\setminus\Gamma_0(\theta)^{at_n}} \|h(\gamma)\|_W \\ &\geq C(at_n)^\kappa - t_n \max_{\gamma\in\Gamma}\|h(\gamma)\|_W \geq 2t_n^\kappa \max_{\gamma\in\Gamma}\|h(\gamma)\|_W - t_n\max_{\gamma\in\Gamma}\|h(\gamma)\|_W \\ &\geq 2t_n \max_{\gamma\in\Gamma}\|h(\gamma)\|_W - t_n\max_{\gamma\in\Gamma}\|h(\gamma)\|_W \geq t_n \min_{\gamma\in\Gamma_0(\theta)}\|h(\gamma)\|_W \\ &= \min_{\gamma\in\Gamma_0(\theta)}\|\theta(\gamma)+t_nh(\gamma)\|_W \geq \sqrt{\phi(\theta+t_nh)}\ , \end{aligned} \tag{1.148}$$

which implies that for all n large,

$$\phi(\theta+t_nh) = \min_{\gamma\in\Gamma_0(\theta)^{at_n}} \|\theta(\gamma)+t_nh(\gamma)\|_W^2\ . \tag{1.149}$$

Now for $\gamma_0 \in \Gamma_0(\theta)$, set $V_{n,\gamma_0}(a) \equiv \{v \in \mathbf{R}^k : \gamma_0 + t_n v \in \Gamma, \|v\| \leq a\}$ and $V(a) \equiv \{v \in \mathbf{R}^k : \|v\| \leq a\}$. Note that $\bigcup_{\gamma_0\in\Gamma_0(\theta)} V_{n,\gamma_0}(a) = \Gamma_0(\theta)^{at_n}$. Since θ and h are continuous, it then follows that

$$\phi(\theta+t_nh) = \min_{\gamma_0\in\Gamma_0(\theta)}\min_{v\in V_{n,\gamma_0}(a)} \|\theta(\gamma_0+t_nv)+t_nh(\gamma_0+t_nv)\|_W^2\ . \tag{1.150}$$

In turn, notice that

$$\begin{aligned}\big|\phi(\theta+t_nh) - &\min_{\gamma_0\in\Gamma_0(\theta)}\min_{v\in V_{n,\gamma_0}(a)} \|\theta(\gamma_0+t_nv)+t_nh(\gamma_0)\|_W^2\big| \\ &\leq 2t_n\|h(\gamma_0)\|_W \cdot t_n \max_{\gamma_0\in\Gamma_0(\theta)}\max_{v\in V_{n,\gamma_0}(a)} \|h(\gamma_0+t_nv)-h(\gamma_0)\|_W \\ &\leq 2t_n^2 \max_{\gamma_1,\gamma_2\in\Gamma:\|\gamma_1-\gamma_2\|\leq at_n} \|h(\gamma_1)-h(\gamma_2)\|_W = o(t_n^2)\ , \end{aligned} \tag{1.151}$$

where the first inequality follows from the formula $|b^2-c^2| \leq |b+c||b-c|$ and that γ_0 is any fixed element in $\Gamma_0(\theta)$, and the last step follows from uniform continuity of h on Γ because h is continuous on Γ and Γ is compact.

Since $\theta \in \prod_{j=1}^{m} C^1(\Gamma)$, we further have,

$$\begin{aligned}
&\Big| \min_{\gamma_0 \in \Gamma_0(\theta)} \min_{v \in V_{n,\gamma_0}(a)} \|\theta(\gamma_0 + t_n v) + t_n h(\gamma_0)\|_W^2 \\
&\qquad - \min_{\gamma_0 \in \Gamma_0(\theta)} \min_{v \in V_{n,\gamma_0}(a)} \|\theta(\gamma_0) + J(\gamma_0) t_n v + t_n h(\gamma_0)\|_W^2 \Big| \\
&\leq 2t_n \max_{\gamma_0 \in \Gamma_0(\theta)} \|h(\gamma_0)\|_W \cdot \max_{\gamma_0 \in \Gamma_0(\theta)} \max_{v \in V_{n,\gamma_0}(a)} \|\theta(\gamma_0 + t_n v) - \theta(\gamma_0) - J(\gamma_0) t_n v\|_W \ . \qquad (1.152)
\end{aligned}$$

By the mean value theorem applied entry-wise to $\theta(\gamma_0 + t_n v) - \theta(\gamma_0)$, there exist $\tilde{\gamma}_n^{(1)}(\gamma_0, v)$, $\ldots, \tilde{\gamma}_n^{(m)}(\gamma_0, v)$ all between θ_0 and $\theta_0 + t_n v$ such that

$$\|\theta(\gamma_0 + t_n v) - \theta(\gamma_0) - J(\gamma_0) t_n v\| = \|J(\tilde{\gamma}_n) t_n v - J(\gamma_0) t_n v\| \ , \qquad (1.153)$$

where by abuse of notation we write

$$J(\tilde{\gamma}_n) \equiv \begin{bmatrix} \frac{d\theta^{(1)}}{d\gamma^\intercal}\big|_{\gamma = \tilde{\gamma}_n^{(1)}(\gamma_0, v)} \\ \vdots \\ \frac{d\theta^{(m)}}{d\gamma^\intercal}\big|_{\gamma = \tilde{\gamma}_n^{(m)}(\gamma_0, v)} \end{bmatrix} \ .$$

Since $\theta \in \prod_{j=1}^{m} C^1(\Gamma)$ and Γ is compact, $J(\cdot)$ is uniformly continuous on Γ and hence

$$\begin{aligned}
&\max_{\gamma_0 \in \Gamma_0(\theta)} \max_{v \in V_{n,\gamma_0}(a)} \|J(\tilde{\gamma}_n) t_n v - J(\gamma_0) t_n v\| \\
&\qquad \leq t_n \max_{\gamma_0 \in \Gamma_0(\theta)} \max_{v \in V_{n,\gamma_0}(a)} \{\|J(\tilde{\gamma}_n) - J(\gamma_0)\| \|v\|\} = o(t_n) \ . \qquad (1.154)
\end{aligned}$$

Since all norms in finite dimensional spaces are equivalent, it follows from results (1.151), (1.152), (1.153), (1.154) and $\theta(\gamma_0) = 0$ for all $\gamma_0 \in \Gamma_0(\theta)$ that

$$\Big|\phi(\theta + t_n h) - \min_{\gamma_0 \in \Gamma_0(\theta)} \min_{v \in V_{n,\gamma_0}(a)} \|J(\gamma_0) t_n v + t_n h(\gamma_0)\|_W^2 \Big| = o(t_n^2) \ . \qquad (1.155)$$

By assumption, $\Gamma_0(\theta)$ is in the interior of Γ and so $V_{n,\gamma_0}(a) = V(a)$ for all n suffi-

ciently large. It follows that

$$\min_{\gamma_0\in\Gamma_0(\theta)}\min_{v\in V_{n,\gamma_0}(a)}\|J(\gamma_0)t_nv+t_nh(\gamma_0)\|_W^2=t_n^2\min_{\gamma_0\in\Gamma_0(\theta)}\min_{v\in V(a)}\|h(\gamma_0)-J(\gamma_0)v\|_W^2\ . \quad (1.156)$$

For each $\gamma_0\in\Gamma_0(\theta)$, by the projection theorem there is some $v^*\in\mathbf{R}^k$ such that

$$\min_{v\in\mathbf{R}^k}\|h(\gamma_0)-J(\gamma_0)v\|_W^2=\|h(\gamma_0)-J(\gamma_0)v^*\|_W^2\ . \quad (1.157)$$

Thus, by choosing a large if necessary so that $v^*\in V(a)$, we have from results (1.155), (1.156) and (1.157) that

$$\left|\phi(\theta+t_nh)-t_n^2\min_{\gamma_0\in\Gamma_0(\theta)}\min_{v\in\mathbf{R}^k}\|h(\gamma_0)-J(\gamma_0)v\|_W^2\right|=o(t_n^2)\ . \quad (1.158)$$

Combining (1.158), $\phi(\theta)=0$ and part (i), we then arrive at part (ii). ■

1.8.3 Proofs for Section 1.5

Lemma 1.8.7. *Let $\phi:\prod_{j=1}^m\ell^\infty(\mathbb{S}^k)\to\mathbf{R}$ be given by $\phi(\theta)=\inf_{\gamma\in\mathbb{S}^k}\|\theta(\gamma)\|^2$. Then*

(i) ϕ is Hadamard differentiable at any $\theta\in\prod_{j=1}^m\ell^\infty(\mathbb{S}^k)$ satisfying $\theta(\gamma_0)=0$ for some $\gamma_0\in\mathbb{S}^k$ and the derivative satisfies $\phi'_\theta(h)=0$ for all $h\in\prod_{j=1}^m\ell^\infty(\mathbb{S}^k)$.

(ii) ϕ is second order Hadamard directionally differentiable at any $\theta_0(\gamma)\equiv E[Z_t\{(\gamma^\intercal Y_{t+1})^2-c(\gamma)\}]$ under Assumption 1.5.1 tangentially to $\prod_{j=1}^m C(\mathbb{S}^k)$ with the derivative given by: for all $h\in\prod_{j=1}^m C(\mathbb{S}^k)$,

$$\phi''_{\theta_0}(h)=\min_{\gamma\in\Gamma_0}\min_{v\in\mathbf{R}^k}\|h(\gamma)+G\,\mathrm{vec}(vv^\intercal)\|^2\ , \quad (1.159)$$

where $\Gamma_0=\{\gamma\in\mathbb{S}^k:\theta_0(\gamma)=0\}$ is the (nonempty) identified set of γ, and $G\in\mathbf{M}^{m\times k^2}$ with the jth row given by $\mathrm{vec}(\Delta_j)^\intercal$ and

$$\Delta_j=E[Z_t^{(j)}(Y_{t+1}Y_{t+1}^\intercal-E[Y_{t+1}Y_{t+1}^\intercal])]\ .$$

PROOF: Fix $\theta \in \prod_{j=1}^m \ell^\infty(\mathbb{S}^k)$ satisfying $\theta(\gamma_0) = 0$ for some $\gamma_0 \in \mathbb{S}^k$, $\{h_n\} \subset \prod_{j=1}^m \ell^\infty(\mathbb{S}^k)$ such that $h_n \to h \in \prod_{j=1}^m \ell^\infty(\mathbb{S}^k)$, and $t_n \downarrow 0$. It follows that

$$\begin{aligned}|\phi(\theta + t_n h_n) - \phi(\theta)| &= \inf_{\gamma\in\mathbb{S}^k} \|\theta(\gamma) + t_n h_n(\gamma)\|^2 \\ &\leq \|\theta(\gamma_0) + t_n h_n(\gamma_0)\|^2 = t_n^2 \|h_n(\gamma_0)\|^2 = o(t_n).\end{aligned}$$

where in the last step we used the fact that $\sup_{\gamma\in\mathbb{S}^k} \|h_n(\gamma)\| = O(1)$. So $\phi'_\theta(h) = 0$ for any $h \in \prod_{j=1}^m \ell^\infty(\mathbb{S}^k)$, as desired for the first claim of the lemma.

Now consider $\theta_0(\gamma) \equiv E[Z_t\{(\gamma^\intercal Y_{t+1})^2 - c(\gamma)\}]$ and suppose that Assumption 1.5.1 holds. Pick $\{h_n\} \subset \prod_{j=1}^m \ell^\infty(\mathbb{S}^k)$ such that $h_n \to h \in \prod_{j=1}^m C(\mathbb{S}^k)$, and $t_n \downarrow 0$. Note that $\phi(\theta_0) = 0$ under Assumption 1.5.1. Then first, we have

$$\begin{aligned}|\phi(\theta_0 + t_n h_n) - \phi(\theta_0 + t_n h)| &= |\inf_{\gamma\in\mathbb{S}^k} \|\theta_0(\gamma) + t_n h_n(\gamma)\|^2 - \inf_{\gamma\in\mathbb{S}^k} \|\theta_0(\gamma) + t_n h(\gamma)\|^2| \\ &\leq |\inf_{\gamma\in\mathbb{S}^k} \|\theta_0(\gamma) + t_n h_n(\gamma)\| + \inf_{\gamma\in\mathbb{S}^k} \|\theta_0(\gamma) + t_n h(\gamma)\|| \\ &\quad \cdot t_n \sup_{\gamma\in\mathbb{S}^k} \|h_n(\gamma) - h(\gamma)\| \\ &\leq t_n |\inf_{\gamma\in\Gamma_0} \|\theta_0(\gamma) + t_n h_n(\gamma)\| + \inf_{\gamma\in\Gamma_0} \|\theta_0(\gamma) + t_n h(\gamma)\|| \\ &\quad \cdot \sup_{\gamma\in\mathbb{S}^k} \|h_n(\gamma) - h(\gamma)\| \\ &= o(t_n^2)\ . \end{aligned} \tag{1.160}$$

Next, let $\Gamma_0^\epsilon \equiv \{\gamma \in \mathbb{S}^k : \min_{s\in\Gamma_0} \|s - \gamma\| \leq \epsilon\}$ and $\Gamma_1^\epsilon \equiv \{\gamma \in \mathbb{S}^k : \min_{s\in\Gamma_0} \|s - \gamma\| \geq \epsilon\}$. By Equation (7) in Dovonon and Renault (2013), $\theta_0(\gamma) = \mathrm{Cov}(Z_t, \sigma_t^2)\mathrm{Diag}(\Lambda^\intercal\gamma\gamma^\intercal\Lambda))$, where for a $p \times p$ matrix A, $\mathrm{Diag}(A)$ denotes the $p \times 1$ vector consisting of diagonal entries. Also, let $\lambda_{\min}(\cdot)$ and $\lambda^+_{\min}(\cdot)$ denote the smallest and the smallest positive singular values, respectively. We then have: for $C \equiv p^{-1/2}\lambda^+_{\min}(\Lambda^\intercal)\lambda_{\min}(\mathrm{Cov}(Z_t, \sigma_t^2))/2$,

$$\begin{aligned}\min_{\gamma\in\Gamma_1^\epsilon} \|\theta_0(\gamma)\| &\geq \min_{\gamma\in\Gamma_1^\epsilon} \|\mathrm{Diag}(\Lambda^\intercal\gamma\gamma^\intercal\Lambda)\|\lambda_{\min}(\mathrm{Cov}(Z_t, \sigma_t^2)) \\ &\geq \min_{\gamma\in\Gamma_1^\epsilon} \|\Lambda^\intercal\gamma\|^2 p^{-1/2}\lambda_{\min}(\mathrm{Cov}(Z_t, \sigma_t^2)) \geq C\epsilon^2\ ,\end{aligned}$$

where the first inequality follows from a simple application of the singular value decomposition of $\mathrm{Cov}(Z_t, \sigma_t^2)$, the second inequality exploits the generalized mean inequality, and last inequality is by Lemma 1.8.10. Note that $\lambda_{\min}(\mathrm{Cov}(Z_t, \sigma_t^2)) > 0$ by Assumption 1.5.1(v). Let $\Delta \equiv [3C^{-1} \max_{\gamma \in \mathbb{S}^k} \|h(\gamma)\|]^{1/2} > 0$ for the nontrivial case $\max_{\gamma \in \mathbb{S}^k} \|h(\gamma)\| > 0$. Then it follows by the triangle inequality that: for n sufficient large such that $t_n \leq \sqrt{t_n}$,

$$\begin{aligned} \min_{\gamma \in \Gamma_1^{\sqrt{t_n}\Delta}} \|\theta_0(\gamma) + t_n h(\gamma)\| &\geq \min_{\gamma \in \Gamma_1^{\sqrt{t_n}\Delta}} \|\theta_0(\gamma)\| - t_n \max_{\gamma \in \mathbb{S}^k} \|h(\gamma)\| \\ &\geq 3t_n \max_{\gamma \in \mathbb{S}^k} \|h(\gamma)\| - t_n \max_{\gamma \in \mathbb{S}^k} \|h(\gamma)\| \\ &> t_n \min_{\gamma \in \Gamma_0} \|h(\gamma)\| \geq \sqrt{\phi(\theta_0 + t_n h)} \ , \end{aligned}$$

and therefore

$$\phi(\theta_0 + t_n h) = \min_{\gamma \in \Gamma_0^{\sqrt{t_n}\Delta}} \|\theta_0(\gamma) + t_n h(\gamma)\|^2 \ .$$

For $\gamma \in \Gamma_0$, let $V_{n,\gamma}^{\Delta} \equiv \{v \in \mathbf{R}^k : \gamma + \sqrt{t_n} v \in \mathbb{S}^k \text{ and } \|v\| \leq \Delta\}$ and $V_{\gamma}^{\Delta} \equiv \{v \in \mathbf{R}^k : \gamma^\intercal v = 0 \text{ and } \|v\| \leq \Delta\}$. Then we have

$$\begin{aligned} \phi(\theta_0 + t_n h) &= \min_{\gamma \in \Gamma_0} \min_{v \in V_{n,\gamma}^{\Delta}} \|\theta_0(\gamma + \sqrt{t_n} v) + t_n h(\gamma + \sqrt{t_n} v)\|^2 \\ &= \min_{\gamma \in \Gamma_0} \min_{v \in V_{n,\gamma}^{\Delta}} \|\theta_0(\gamma + \sqrt{t_n} v) + t_n h(\gamma)\|^2 + o(t_n^2) \ , \end{aligned} \tag{1.161}$$

where the first equality follows by the definition of $\Gamma_0^{\sqrt{t_n}\Delta}$ and the second equality follows by noting that

$$\begin{aligned} &|\min_{\gamma \in \Gamma_0} \min_{v \in V_{n,\gamma}^{\Delta}} \|\theta_0(\gamma + \sqrt{t_n} v) + t_n h(\gamma + \sqrt{t_n} v)\|^2 - \min_{\gamma \in \Gamma_0} \min_{v \in V_{n,\gamma}^{\Delta}} \|\theta_0(\gamma + \sqrt{t_n} v) + t_n h(\gamma)\|^2| \\ &\leq 2t_n \|h(\gamma_0)\| \cdot t_n \max_{\gamma \in \Gamma_0} \max_{v \in V_{n,\gamma}^{\Delta}} \|h(\gamma + \sqrt{t_n} v) - h(\gamma)\| \\ &\leq 2t_n^2 \|h(\gamma_0)\| \max_{\gamma_1, \gamma_2 \in \mathbb{S}^k, \|\gamma_1 - \gamma_2\| \leq \sqrt{t_n}\Delta} \|h(\gamma_1) - h(\gamma_2)\| = o(t_n^2) \ , \end{aligned}$$

where γ_0 in the first inequality is any fixed element in Γ_0, the last equality follows by the uniform continuity of h over $\mathbb{S}^k$. By $\theta_0(\gamma) = G \operatorname{vec}(\gamma\gamma^\intercal)$ (Dovonon and Renault, 2013) and

the definition of Γ_0, we have

$$\begin{aligned}\min_{\gamma\in\Gamma_0}\min_{v\in V_{n,\gamma}^{\Delta}}\|\theta_0(\gamma+\sqrt{t_n}v)+t_nh(\gamma)\|^2 &= t_n^2\min_{\gamma\in\Gamma_0}\min_{v\in V_{n,\gamma}^{\Delta}}\|G\,\mathrm{vec}(vv^{\intercal})+h(\gamma)\|^2\\ &= t_n^2\min_{\gamma\in\Gamma_0}\min_{v\in V_{\gamma}^{\Delta}}\|G\,\mathrm{vec}(vv^{\intercal})+h(\gamma)\|^2+o(t_n^2)\\ &= t_n^2\min_{\gamma\in\Gamma_0}\min_{v\in\mathbf{R}^k}\|G\,\mathrm{vec}(vv^{\intercal})+h(\gamma)\|^2+o(t_n^2)\ , \qquad (1.162)\end{aligned}$$

where the second equality follows by the fact that $V_{n,\gamma}^{\Delta}$ converges to V_{γ}^{Δ} uniformly in $\gamma\in\Gamma_0$ with respect to the Hausdorff metric by Lemma 1.8.11 and Lemma B.3 in Fang and Santos (2015), and the third equality by the facts that $G\,\mathrm{vec}(vu^{\intercal})=0$ for all $v\in\Gamma_0$ and all $u\in\mathbf{R}^k$ and that the inside minimum can be attained in V_{γ}^{Δ} for all Δ large enough. Combining (1.160), (1.161) and (1.162) yields

$$\phi''_{\theta_0}(h)=\lim_{n\to\infty}\frac{\phi(\theta_0+t_nh_n)}{t_n^2}=\min_{\gamma\in\Gamma_0}\min_{v\in\mathbf{R}^k}\|h(\gamma)+G\,\mathrm{vec}(vv^{\intercal})\|^2,$$

which establishes the second result. ■

Lemma 1.8.8. *Under Assumptions 1.5.1 and 1.5.2, we have*

$$\sqrt{T}\{\hat{\theta}_T-\theta_0\}\xrightarrow{L}\mathbb{G}\ \textit{in}\ \prod_{j=1}^{m}\ell^{\infty}(\mathbb{S}^k)\ ,$$

where $\mathbb{G}$ is a zero mean Gaussian process with the covariance functional satisfying: for any $\gamma_1,\ \gamma_2\in\Gamma_0$ and $\mu_z=E[Z_t]$,

$$E[\mathbb{G}(\gamma_1)\mathbb{G}(\gamma_2)]=E[(Z_t-\mu_z)(Z_t-\mu_z)^{\intercal}\{(\gamma_1^{\intercal}Y_{t+1})^2-c(\gamma_1)\}\{(\gamma_2^{\intercal}Y_{t+1})^2-c(\gamma_2)\}]\ .$$

Proof: By elementary rearrangements we have

$$\sqrt{T}\{\hat{\theta}_T(\gamma)-\theta_0(\gamma)\}=\sqrt{T}G_T(\gamma)-\sqrt{T}(\hat{\mu}_z-\mu_z)\{\hat{c}(\gamma)-c(\gamma)\}\ ,$$

where $\hat{\mu}_z = \frac{1}{T}\sum_{t=1}^{T} Z_t$, $\hat{c}(\gamma) = \frac{1}{T}\sum_{t=1}^{T}(\gamma^\intercal Y_{t+1})^2$, and

$$G_T(\gamma) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(Z_t - \mu_z)\{(\gamma^\intercal Y_{t+1})^2 - c(\gamma)\} - E[(Z_t - \mu_z)\{(\gamma^\intercal Y_{t+1})^2 - c(\gamma)\}] \ .$$

By Assumptions 1.5.1(vi) and 1.5.2, and the law of large numbers for stationary and ergodic sequences and the compactness of $\mathbb{S}^k$, we have

$$\sqrt{T}(\hat{\mu}_z - \mu_z)(\hat{c} - c) = o_p(1) \text{ in } \prod_{j=1}^{m} \ell^\infty(\mathbb{S}^k) \ .$$

Once again by Assumptions 1.5.1(vi) and 1.5.2, together with $G_T(\gamma) = \sqrt{T}\tilde{G}\,\mathrm{vec}(\gamma\gamma^\intercal)$ where $\tilde{G} \in \mathbf{M}^{m\times k^2}$ having its jth row given by $(\mathrm{vec}(\tilde{\Delta}_j))^\intercal$ for

$$\begin{aligned}\tilde{\Delta}_j = \frac{1}{T}\sum_{t=1}^{T}(Z_t^{(j)} - \mu_z^{(j)})\{Y_{t+1}Y_{t+1}^\intercal - E(Y_{t+1}Y_{t+1}^\intercal)\} \\ - E[(Z_t^{(j)} - \mu_z^{(j)})\{Y_{t+1}Y_{t+1}^\intercal - E(Y_{t+1}Y_{t+1}^\intercal)\}] \ ,\end{aligned}$$

we have by the compactness of $\mathbb{S}^k$,

$$G_T \xrightarrow{L} \mathbb{G} \text{ in } \prod_{j=1}^{m} \ell^\infty(\mathbb{S}^k)$$

for some Gaussian process $\mathbb{G}(\gamma)$. In particular, for $\gamma \in \Gamma_0$ the summand in $G_T(\gamma)$ is a martingale difference sequence, so for any γ_1, $\gamma_2 \in \Gamma_0$, the covariance functional satisfies

$$E[\mathbb{G}(\gamma_1)\mathbb{G}(\gamma_2)] = E[(Z_t - \mu_z)(Z_t - \mu_z)^\intercal\{(\gamma_1^\intercal Y_{t+1})^2 - c(\gamma_1)\}\{(\gamma_2^\intercal Y_{t+1})^2 - c(\gamma_2)\}] \ .$$

This completes the proof of the lemma. ∎

Lemma 1.8.9. *Suppose Assumptions 1.5.1, 1.5.2 and 1.5.3 hold. Let $\hat{\phi}_T''$ be constructed as in* (1.61)*. Then we have: whenever $h_T \to h$ as $T \to \infty$ for a sequence $\{h_T\} \subset \prod_{j=1}^{m} \ell^\infty(\mathbb{S}^k)$*

and $h \in \prod_{j=1}^m C(\mathbb{S}^k)$*, it follows that*

$$\hat{\phi}''_T(h_T) \xrightarrow{p} \phi''_{\theta_0}(h) \ .$$

PROOF: Pick a sequence $\{h_T\} \subset \prod_{j=1}^m \ell^\infty(\mathbb{S}^k)$ and $h \in \prod_{j=1}^m C(\mathbb{S}^k)$ such that $h_T \to h$ as $T \to \infty$. Define

$$\tilde{\phi}''_T(h) = \min_{\gamma \in \hat{\Gamma}_n} \min_{v \in B_T} \|h(\gamma) + G \operatorname{vec}(vv^\intercal)\|^2 \ .$$

Then we have

$$\begin{aligned}
&|\hat{\phi}''_T(h_T) - \tilde{\phi}''_T(h)| \\
&\quad \le \Big| \inf_{\gamma \in \hat{\Gamma}_n} \min_{v \in B_T} \|h_T(\gamma) + \hat{G} \operatorname{vec}(vv^\intercal)\| + \inf_{\gamma \in \hat{\Gamma}_n} \min_{v \in B_T} \|h(\gamma) + G \operatorname{vec}(vv^\intercal)\| \Big| \\
&\qquad \cdot \Big| \inf_{\gamma \in \hat{\Gamma}_n} \min_{v \in B_T} \|h_T(\gamma) + \hat{G} \operatorname{vec}(vv^\intercal)\| - \inf_{\gamma \in \hat{\Gamma}_n} \min_{v \in B_T} \|h(\gamma) + G \operatorname{vec}(vv^\intercal)\| \Big| \\
&\quad \le (\sup_{\gamma \in \mathbb{S}^k} \|h_T(\gamma)\| + \sup_{\gamma \in \mathbb{S}^k} \|h(\gamma)\|) \sup_{\gamma \in \mathbb{S}^k} \|h_T(\gamma) - h(\gamma)\| \sup_{v \in B_T} \|\operatorname{vec}(vv^\intercal)\| \|\hat{G} - G\| \\
&\quad \lesssim \sup_{v \in B_T} T^{-1/2} \|v\|^2 \|\sqrt{T}\{\hat{G} - G\}\| \le T^{-1/2} \kappa_T^{-1/2} \|\sqrt{T}\{\hat{G} - G\}\| \\
&\quad = o_p(1) \ , \qquad (1.163)
\end{aligned}$$

where "$\lesssim$" follows from $h_T \to h$, and the last step is by Assumptions 1.5.2 and 1.5.3.

Next, under Assumptions 1.5.1, 1.5.2 and 1.5.3, we have by Theorem 3.1 in Chernozhukov et al. (2007) that $d_H(\hat{\Gamma}_n, \Gamma_0) \xrightarrow{p} 0$ as $T \to \infty$, with $a_T = T$, $b_T = \sqrt{T}$, and $\hat{c} = T\kappa_T$. Let

$$\bar{\phi}''_T(h) = \min_{\gamma \in \Gamma_0} \min_{v \in B_T} \|h(\gamma) + G \operatorname{vec}(vv^\intercal)\|^2 \ .$$

Since $h \in \prod_{j=1}^m C(\mathbb{S}^k)$ and $\mathbb{S}^k$ is compact, together with $d_H(\hat{\Gamma}_n, \Gamma_0) \xrightarrow{p} 0$, it follows that

$$\begin{aligned}
&|\tilde{\phi}''_T(h) - \bar{\phi}''_T(h)| \\
&\leq \sup_{\|\gamma_1-\gamma_2\|_{\mathbb{D}} \leq d_H(\hat{\Gamma}_n,\Gamma_0)} \left| \min_{v\in B_T} \|h(\gamma_1) + G\,\mathrm{vec}(vv^\intercal)\|^2 - \min_{v\in B_T} \|h(\gamma_2) + G\,\mathrm{vec}(vv^\intercal)\|^2 \right| \\
&\leq \sup_{\|\gamma_1-\gamma_2\|_{\mathbb{D}} \leq d_H(\hat{\Gamma}_n,\Gamma_0)} \|h(\gamma_1) - h(\gamma_2)\| = o_p(1) \ . \qquad (1.164)
\end{aligned}$$

Since $\bar{\phi}''_T(h)$ is monotonically decreasing as $T \uparrow \infty$, we further have

$$\bar{\phi}''_T(h) \to \min_{\gamma\in\Gamma_0} \min_{v\in\mathbf{R}^k} \|h(\gamma) + G\,\mathrm{vec}(vv^\intercal)\|^2 = \phi''_{\theta_0}(h) \ . \qquad (1.165)$$

The lemma then follows from results (1.163), (1.164) and (1.165). ■

Proof of Proposition 1.5.2: By Lemmas 1.8.8 and 1.8.9, Assumptions 1.3.1 and 1.3.2, and the cdf of the weak limit being strictly increasing at $c_{1-\alpha}$, we have $\hat{c}_{1-\alpha} \xrightarrow{p} c_{1-\alpha}$ following exactly the same proof of Corollary 3.2 in Fang and Santos (2015).[18] Then under H_0, the conclusion follows from combining Proposition 1.5.1, Slutsky thoerem, $c_{1-\alpha}$ being a continuity point of the weak limit and the portmanteau theorem. ■

Lemma 1.8.10. *Let Λ and Γ_1^ϵ be given as in the proof of Lemma 1.8.7. Then under Assumption 1.5.1 and H_0, for all sufficiently small $\epsilon > 0$, we have*

$$\min_{\gamma\in\Gamma_1^\epsilon} \|\Lambda^\intercal\gamma\| \geq \frac{\epsilon}{\sqrt{2}} \sigma^+_{\min}(\Lambda^\intercal) \ ,$$

where $\sigma^+_{\min}(\Lambda^\intercal)$ denotes the smallest positive singular value of $\Lambda^\intercal$.

Proof: To begin with, note that i) $\Gamma_0 = \arg\min_{\gamma\in\mathbb{S}^k} \|\Lambda^\intercal\gamma\|$ by Assumption 1.5.1, ii) $\Gamma_0 \neq \emptyset$ under the null, iii) $\sigma^+_{\min}(\Lambda^\intercal)$ is well-defined by Assumption 1.5.1(i) so that $\Gamma_0 \subsetneqq \mathbb{S}^k$. Let $\Lambda^\intercal = P\Sigma Q^\intercal$ be the singular value decomposition of $\Lambda^\intercal$, where $P \in \mathbf{M}^{p\times p}$ and $Q \in \mathbf{M}^{k\times k}$ are orthonormal, and $\Sigma \in \mathbf{M}^{p\times k}$ is a diagonal matrix with diagonal entries in descending

[18] Note that ϕ''_{θ_0} trivially admits a continuous extension on $\prod_{j=1}^m \ell^\infty(\mathbb{S}^k)$ with the first min replaced by inf.

order. Since Λ is of full column rank, $\sigma^+_{\min}(\Lambda^\intercal)$ is equal to the pth diagonal entry of Σ with $p < k$.

Fix $\gamma \in \Gamma_1^\epsilon$. Let $a_\gamma \equiv Q^\intercal\gamma$ and write $a_\gamma = [a_\gamma^{(1)\intercal}, a_\gamma^{(2)\intercal}]^\intercal$ for $a_\gamma^{(1)} \in \mathbf{R}^p$ and $a_\gamma^{(2)} \in \mathbf{R}^{k-p}$. Suppose first that $\|a_\gamma^{(2)}\| \neq 0$. Then we have

$$\|[0, a_\gamma^{(2)\intercal}]^\intercal/\|a_\gamma^{(2)}\| - a_\gamma\| = \|Q[0, a_\gamma^{(2)\intercal}]^\intercal/\|a_\gamma^{(2)}\| - \gamma\| \geq \min_{s\in\Gamma_0}\|s-\gamma\| \geq \epsilon \ , \tag{1.166}$$

since $Q[0, a_\gamma^{(2)\intercal}]^\intercal/\|a_\gamma^{(2)}\| \in \Gamma_0$ by direct calculations. In turn, result (1.166) implies

$$\|a_\gamma^{(1)}\|^2 + (1 - \|a_\gamma^{(2)}\|)^2 \geq \epsilon^2 \ . \tag{1.167}$$

Moreover, we know from $Q \in \mathbf{M}^{k\times k}$ being orthonormal and $\gamma \in \mathbb{S}^k$ that

$$\|a_\gamma^{(1)}\|^2 + \|a_\gamma^{(2)}\|^2 = 1 \ . \tag{1.168}$$

Combining results (1.166) and (1.167) we may thus conclude that

$$2\|a_\gamma^{(1)}\|^2 = \|a_\gamma^{(1)}\|^2 + 1 - \|a_\gamma^{(2)}\|^2 \geq \|a_\gamma^{(1)}\|^2 + (1 - \|a_\gamma^{(2)}\|)^2 \geq \epsilon^2 \ , \tag{1.169}$$

implying that $\|a_\gamma^{(1)}\| \geq \frac{\epsilon}{\sqrt{2}}$. This also holds for all sufficiently small $\epsilon > 0$ when $\|a_\gamma^{(2)}\| = 0$ in which case $\|a_\gamma^{(1)}\| = 1$ in view of (1.168). Consequently, we have

$$\begin{aligned}\min_{\gamma\in\Gamma_1^\epsilon}\|\Lambda^\intercal\gamma\| &= \min_{\gamma\in\Gamma_1^\epsilon}\|P\Sigma Q^\intercal\gamma\| = \min_{\gamma\in\Gamma_1^\epsilon}\|\Sigma a_\gamma\| \\ &\geq \lambda^+_{\min}(\Lambda^\intercal)\min_{\gamma\in\Gamma_1^\epsilon}\|a_\gamma^{(1)}\| \geq \lambda^+_{\min}(\Lambda^\intercal)\frac{\epsilon}{\sqrt{2}} \ ,\end{aligned} \tag{1.170}$$

for all sufficiently small $\epsilon > 0$. This completes the proof of the lemma. ■

Lemma 1.8.11. *Let $V^\Delta_{n,\gamma}$ and V^Δ_γ be defined as in the proof of Lemma 1.8.7. Then $d_H(V^\Delta_{n,\gamma}, V^\Delta_\gamma) \to 0$ uniformly in $\gamma \in \Gamma_0$ as $n \to \infty$.*

PROOF: First, note that $V^\Delta_{n,\gamma} = \{v \in \mathbf{R}^k : \gamma + \sqrt{t_n}v \in \mathbb{S}^k \text{ and } \|v\| \leq \Delta\}$. For $u \in V^\Delta_{n,\gamma}$, set

$u^* \equiv u - (\gamma^\intercal u)\gamma$. It is a simple exercise to verify that $u^* \in V_\gamma^\Delta$. It follows that

$$\min_{v \in V_\gamma^\Delta} \|u - v\| \leq \|u - u^*\| \leq \frac{1}{2}\sqrt{t_n}\Delta^2 \ . \tag{1.171}$$

In turn, result (1.171) implies that: for all $\gamma \in \Gamma_0$,

$$\max_{u \in V_{n,\gamma}^\Delta} \min_{v \in V_\gamma^\Delta} \|u - v\| \leq \frac{1}{2}\sqrt{t_n}\Delta^2 \ . \tag{1.172}$$

On the other hand, for $v \in V_\gamma^\Delta$, set $v^* = v - b_n\gamma$ for $b_n = (1 - \sqrt{1 - t_n\|v\|})/\sqrt{t_n}$ if $\|v\| < \Delta$, and $v^* = a_n v - b_n\gamma$ for $a_n = 1 - \sqrt{t_n}$ and $b_n = (1 - \sqrt{1 - t_n(1 - \sqrt{t_n})^2\|v\|^2})/\sqrt{t_n}$ if $\|v\| = \Delta$. In any case, $v^* \in V_{n,\gamma}^\Delta$ by direct calculations. Therefore,

$$\min_{v \in V_\gamma^\Delta} \max_{u \in V_{n,\gamma}^\Delta} \|u - v\| \leq \min_{v \in V_\gamma^\Delta} \|v - v^*\| = O(\sqrt{t_n}) \ , \tag{1.173}$$

uniformly in $\gamma \in \Gamma_0$, where we exploited the facts that $b_n = O(\sqrt{t_n})$ uniformly in $\gamma \in \Gamma_0$ and that V_γ^Δ is bounded. The lemma then follows from results (1.172) and (1.173). ■

Chapter 2

Improved Inference on the Rank of a Matrix with Applications to IV and Cointegration Models

Abstract

This chapter develops new methods for examining a "no greater than" inequality of the rank of a matrix and for rank determination in a general setup, which improve upon existing methods. Existing rank tests assume *a priori* that the rank is no less than the hypothesized value, which is often unrealistic. These tests when directly applied may fail to control the asymptotic null rejection rate, and the multiple testing method based on them can be conservative with the asymptotic null rejection rate strictly below the nominal level whenever the rank is less than the hypothesized value. We prove that our proposed tests have the asymptotic null rejection rate that is exactly equal to the nominal level under minimal assumptions regardless of whether the rank is less than or equal to the hypothesized value. As our simulation results show, these characteristics lead to an improved power property in general. In application to a context with stationary and nonstationary data,

respectively, our tests yield improved tests for identification in linear IV models and for the existence of stochastic trend and/or cointegration with or without VAR specification. In addition, our simulation results show that the improved power property of our tests leads to an improved accuracy of the sequential testing procedure for rank determination.

2.1 Introduction

The rank of a matrix plays a fundamental role in numerous economic and statistical settings, including identification of structural parameters (Fisher, 1966), existence of common features (Engle and Kozicki, 1993) with the canonical example being that of cointegration (Engle and Granger, 1987), the rank of a (consumer) demand system (Gorman, 1981; Lewbel, 1991), specification of factor models (Ross, 1976), dimension reduction in regression analysis (Li, 1991; Bura and Yang, 2011), and model specification in time series (Aoki, 1990; Gill and Lewbel, 1992). These problems reduce to examining the following hypotheses: for an unknown matrix Π_0 of size $m \times k$ with $m \geq k$,

$$\mathrm{H}_0 : \mathrm{rank}(\Pi_0) \leq r \quad \text{v.s.} \quad \mathrm{H}_1 : \mathrm{rank}(\Pi_0) > r \ , \tag{2.1}$$

where $r \in \{0, \ldots, k-1\}$ is some prespecified value and $\mathrm{rank}(\Pi_0)$ denotes the rank of Π_0. If $r = k - 1$, then (2.1) is simply a testing problem of whether Π_0 has full rank.

Despite a rich set of results in the literature, previous studies instead focus on the following hypotheses

$$\mathrm{H}_0^{(r)} : \mathrm{rank}(\Pi_0) = r \quad \text{v.s.} \quad \mathrm{H}_1^{(r)} : \mathrm{rank}(\Pi_0) > r \ . \tag{2.2}$$

In effect, this is a different testing problem and assumes *a priori* that $\mathrm{rank}(\Pi_0)$ is no less than r. Unfortunately, in the aforementioned problems, it is unrealistic to make such an assumption. As shown in Section 2.2.2, when in fact $\mathrm{rank}(\Pi_0) < r$, directly applying existing rank tests to (2.1) may fail to control the asymptotic null rejection rate, since the asymptotic distributions of test statistics can be very different from those when $\mathrm{rank}(\Pi_0) = r$. As we shall prove (see Lemma 2.7.4), when $\mathrm{rank}(\Pi_0) < r$, the problem (2.1) becomes irregular in the sense that a functional characterizing the problem admits a degenerate first order derivative and is second order nondifferentiable. A general inferential framework for such functionals was not available until very recently (Fang and Santos, 2015; Chen and Fang,

2015). To the best of our knowledge, no direct tests for (2.1) exist in the literature.

Our method builds on the insight that (2.1) can be equivalently reformulated as

$$\mathrm{H}_0 : \phi(\Pi_0) = 0 \quad \text{v.s.} \quad \mathrm{H}_1 : \phi(\Pi_0) > 0 \ , \tag{2.3}$$

where $\phi(\Pi_0) \equiv \sum_{j=r+1}^{k} \sigma_j^2(\Pi_0)$ is the sum of the $k - r$ smallest squared singular values $\sigma_j^2(\Pi_0)$ of Π_0 (i.e., the sum of the $k-r$ smallest eigenvalues of $\Pi_0^\intercal \Pi_0$). For a given estimator $\hat{\Pi}_n$ of Π_0, we then employ the plug-in estimator $\tau_n^2 \phi(\hat{\Pi}_n)$ as our test statistic, where τ_n is the rate at which $\hat{\Pi}_n$ admits an asymptotic distribution. Towards invoking the Delta method, we prove, however, that the first order derivative of the map $\Pi \mapsto \phi(\Pi)$ is null at $\Pi = \Pi_0$ under H_0, necessitating a second order analysis. Since the asymptotic distributions (under the composite null) implied by the second order Delta method (Shapiro, 2000) are highly nonstandard, we appeal to the bootstrap procedure recently developed by Fang and Santos (2015) and Chen and Fang (2015) in order to obtain valid critical values and conduct inference. We also extend the results to accommodate the case when the convergence rates of $\hat{\Pi}_n$ are not homogenous across its columns as in VAR models with stochastic trend and cointegration (see Appendix 2.7.2).

There are several attractive features of our tests. First, since we rely on the Delta method, the theory is conceptually simple and requires minimal assumptions. Essentially, all we need are a matrix estimator $\hat{\Pi}_n$ that converges weakly and a consistent bootstrap analog $\hat{\Pi}_n^*$. As a matter of fact, our tests apply to various data generating processes. Second, implementation of the procedure is computationally easy, only involving calculation of singular value decompositions. Finally, since construction of the critical values is based on bootstrapping the asymptotic distributions pointwise in Π_0, the resulting tests have the asymptotic null rejection rate that is exactly equal to the nominal level regardless of whether $\mathrm{rank}(\Pi_0) = r$ or $\mathrm{rank}(\Pi_0) < r$. As our simulation results show, these characteristics lead to good power properties of our tests in general. In application to a context with stationary and nonstationary data, respectively, our tests yield new and powerful tests for identification in linear IV models (Fisher, 1966) and for the existence of stochastic trend and/or cointegration

with or without VAR specification (Engle and Granger, 1987).

As an alternative to the direct application, one may instead adapt existing rank tests into multiple testing procedures, since H_0 holds if and only if $H_0^{(q)}$ holds for some $0 \leq q \leq r$. Specifically, the multiple testing method rejects H_0 if and only if all $H_0^{(q)}$ are rejected and otherwise fails to reject. However, as demonstrated in Sections 2.2.2 and 2.4.1, the method can be severely conservative when $\text{rank}(\Pi_0) > r$ and Π_0 is close to a matrix with rank strictly less than r, with the asymptotic null rejection rate strictly below the nominal level when $\text{rank}(\Pi_0) < r$. This is in sharp contrast to our tests, which by design achieve asymptotic null rejection rates exactly equal to the nominal level and hence improve the power properties. In an application to testing for identification in stochastic discount factor models, compared to the multiple testing method based the Kleibergen and Paap (2006) test, our tests suggest much weaker evidence of non-identification of the risk premia parameters.

In some settings such as the rank of a demand system, specification of factor models and model specification in time series, the main concern boils down to determining the true rank of a matrix. To determine $\text{rank}(\Pi_0)$, one may implement the sequential testing procedure, following Johansen (1995), based on rank tests for (2.1) or (2.2). Interestingly, efficient rank determination does not require the ability of detecting whether $\text{rank}(\Pi_0)$ is strictly less than a hypothesized value. This explains the prevalence of existing rank tests in rank determination. Nevertheless, the power of detecting whether $\text{rank}(\Pi_0)$ is strictly greater than hypothesized values plays an important role in the procedure. Our simulation results show that the improved power property of our tests leads to an improved accuracy of the sequential testing procedure for rank determination.

As mentioned previously, the literature has been mostly concerned with the hypotheses (2.2). In the context of multivariate regression, Anderson (1951) proposed a likelihood ratio test based on canonical correlations. This test is restrictive in the sense that it crucially depends on a Kronecker product structure of the covariance matrix of a matrix estimator. Building on the LDU decomposition approach in Gill and Lewbel (1992), Cragg and Donald

(1996) proposed a test with the test statistic being a quadratic form of the vectorization of a submatrix in the LDU decomposition that is sensitive to variable ordering. In Cragg and Donald (1997), the authors provided a test based on a constrained minimum χ^2 distance criterion, which is computationally intensive because it involves minimization over the set of all matrices with rank r. Moreover, both tests rely on the condition that the asymptotic covariance matrix of the vectorization of the matrix estimator is nonsingular, which we do not require in our analysis. Motivated by the need to relax this nonsingularity condition, Robin and Smith (2000) developed a test based on functionals of the characteristics of a suitably transformed matrix. However, their test depends on a rank condition that is "empirically nonverifiable". All these rank tests may fail to control the asymptotic null rejection rate when directly applied to the hypotheses (2.1).

Moreover, Kleibergen and Paap (2006) proposed a test based on singular value decomposition of a transformed matrix with the test statistic having the $\chi^2((m-r)(k-r))$ asymptotic distribution under $\mathrm{H}_0^{(r)}$. Despite overcoming many of the deficiencies of previous tests, this test still requires some covariance matrix nonsingular because it is based on a Wald statistic, which we do not require in our analysis. More importantly, this rank test also has the aforementioned drawback when directly applied to the hypotheses (2.1). There are, nonetheless, a few exceptions that study (2.1), notably Cragg and Donald (1993) who considered a special case of Cragg and Donald (1997). However, the asymptotic distribution of the test statistic when $\mathrm{rank}(\Pi_0) < r$ is not available, though Cragg and Donald (1993) established that the asymptotic null distribution when $\mathrm{rank}(\Pi_0) = r$ is least favorable under somewhat restrictive conditions. Thus, when $\mathrm{rank}(\Pi_0) > r$ and Π_0 is close to a matrix with rank strictly less than r, their test can be conservative. We refer the reader to Camba-Mendez and Kapetanios (2009a), Portier and Delyon (2014) and Al-Sadoon (2015) for further discussions of the literature.

The remainder of the chapter is organized as follows. Section 2.2 presents related examples to illustrate the importance of the problem, and demonstrates the drawback of existing rank tests and the conservativeness of the multiple testing method. Section 2.3

develops the test statistic, establishes its asymptotic distribution, and proposes a bootstrap procedure for inference. Section 2.4 presents Monte Carlo studies, applies our method to study identification in stochastic discount factor models, and demonstrates the accuracy improvement of the sequential testing procedure for rank determination based on our tests. Section 2.5 briefly concludes. All the proofs are collected in the appendices.

2.2 Examples and Motivation

In this section, we first present related examples in which the testing problem (2.1) is of importance. In order to motivate the development of our tests, we then demonstrate that existing rank tests when directly applied to (2.1) can fail to control the asymptotic null rejection rate, and that the multiple testing method can be conservative.

2.2.1 Examples

The first example is what motivated this paper in the first place.

Example 2.2.1 (Identification)**.** Let $Y \in \mathbf{R}$ and $Z \in \mathbf{R}^k$ be random variables satisfying

$$Y = Z^\intercal \beta_0 + u \ . \tag{2.4}$$

Let $W \in \mathbf{R}^m$ be instrument variables such that $E[Wu] = 0$ with $m \geq k$. Then identification of the coefficient β_0 reduces to whether $E[WZ^\intercal]$ is of full rank. Thus, testing for identification of β_0 reduces to examining the hypotheses (2.1) with

$$\Pi_0 = E[WZ^\intercal] \text{ and } r = k - 1 \ . \tag{2.5}$$

We cannot restrict ourselves to examine the hypotheses (2.2), since it is unrealistic to assume $\text{rank}(\Pi_0) \geq k - 1$ unless $k = 1$. More generally, (local) identification in parametric, semiparametric and nonparametric models can often be expressed in terms of some matrices being of full rank (Fisher, 1961; Rothenberg, 1971; Roehrig, 1988; Chesher, 2003; Matzkin,

2008; Chen et al., 2014b). For identification in DSGE models, see, for example, Canova and Sala (2009) and Komunjer and Ng (2011). In addition, when $W = Z$, then Π_0 is a positive semidefinite matrix and the concern becomes the existence of perfect multicolinearity among Z. ■

The next example concerns the existence of stochastic trend and/or cointegration in a vector autoregression (VAR) system (Engle and Granger, 1987; Johansen, 1991).

Example 2.2.2 (VAR Trend/Cointegration)**.** Let $\{Y_t\}$ be a $k \times 1$ time series such that each component of Y_t is integrated of order 0 or 1, that is, each component is a stationary or unit root process. Assume the entire vector is a VAR(1) process

$$Y_t = \Phi_0 Y_{t-1} + u_t \ , \tag{2.6}$$

where u_t are white noise with nonsingular covariance matrix Σ. The error-correction representation of (2.6) is given by (Hamilton, 1994, p.580):

$$\Delta Y_t = (\Phi_0 - I_k) Y_{t-1} + u_t \ . \tag{2.7}$$

Then the existence of stochastic trend for Y_t means that $\Phi_0 - I_k$ is not of full rank. Thus, testing for the existence of stochastic trend reduces to examining the hypotheses (2.1) with

$$\Pi_0 = \Phi_0 - I_k \text{ and } r = k - 1 \ . \tag{2.8}$$

It is unrealistic to assume that there is at most one linearly independent stochastic trend (i.e., $\text{rank}(\Pi_0) \geq k-1$) unless $k = 1$, so we cannot instead focus on examining the hypotheses (2.2). In addition, the existence of cointegrating relations for Y_t means that $\Phi_0 - I_k$ is nonzero.[1] Thus, testing for the existence of cointegration reduces to examining the hypotheses (2.1) with $r = 0$. We confine our attention to the class of VAR(1) models with white noise errors for simplicity, but our framework applies more broadly to VAR(p) processes

[1] Recall that Y_t is said to be cointegrated if there exists nonzero $\lambda \in \mathbf{R}^k$ such that $\lambda^\intercal Y_t$ is stationary.

with dependent and heteroskedastic errors. ■

Our results allow us to study stochastic trend and cointegration nonparametrically. The following example concerns the existence of stochastic trend and/or cointegration without a VAR specification (Engle and Granger, 1987; Bierens, 1997; Shintani, 2001).

Example 2.2.3 (Nonparametric Trend/Cointegration)**.** Let $\{Y_t\}$ be a $k \times 1$ time series such that each component of Y_t is integrated of order 0 or 1, that is, each component is a stationary or unit root process. Let the first difference of Y_t follow a linear process

$$\Delta Y_t = C(L)u_t \equiv \sum_{j=0}^{\infty} C_j u_{t-j} \ , \tag{2.9}$$

where u_t are white noise with nonsingular covariance matrix Σ, and $C_0 = I_k$. Since the long run covariance matrix of ΔY_t is equal to $C(1)\Sigma C(1)^{\intercal}$, then existence of cointegrating relations for Y_t means that the long run covariance matrix of ΔY_t is not of full rank. Thus, testing for the existence of cointegration reduces to examining the hypotheses (2.1) with

$$\Pi_0 = \sum_{t=-\infty}^{\infty} E[\Delta Y_t \Delta Y_0] \text{ and } r = k - 1 \ . \tag{2.10}$$

We cannot restrict ourselves to examine the hypotheses (2.2), since it is unrealistic to assume there is at most one linearly independent cointegration vectors (i.e., $\text{rank}(\Pi_0) \geq k-1$) unless $k = 1$. In addition, the existence of stochastic trend for Y_t means that $\Phi_0 - I_k$ is nonzero. Thus, testing for the existence of stochastic trend reduces to examining the hypotheses (2.1) with $r = 0$. ■

Cointegration is just one particular example of the more general notion of common features (Engle and Kozicki, 1993). Our fourth example pertains to the existence of general common features.

Example 2.2.4 (Common Features)**.** Let $\{Y_t\}$ be a $k \times 1$ time series. According to Engle and Kozicki (1993), a feature that is present in each component of Y_t is said to be common

to Y_t if there exists a nonzero linear combination of Y_t that fails to have the feature. Suppose that $\{Y_t\}$ is generated according to

$$Y_t = \Gamma_0^\intercal Z_t + \Xi_0^\intercal W_t + u_t \ , \tag{2.11}$$

where W_t can be thought of as control variables, and Z_t is an $m \times 1$ vector reflecting the feature under consideration with $m \geq k$. For example, testing for the existence of common serial correlation would set Z_t to be lags of Y_t, and testing for the existence of common conditionally heteroskedastic factors would set Z_t to be relevant factors. We refer to Engle and Kozicki (1993), Engle and Susmel (1993) and Dovonon and Renault (2013) for details of these and other examples. By the definition of common feature and the specification of (2.11), existence of common features means that Γ_0 is not of full rank. Thus, testing for the existence of common features reduces to examining the hypotheses (2.1) with

$$\Pi_0 = \Gamma_0 \text{ and } r = k - 1 \ . \tag{2.12}$$

Since the number of common features is unknown *a priori*, we cannot restrict ourself to examine the hypotheses (2.2) by assuming $\text{rank}(\Pi_0) \geq k - 1$ unless $k = 1$. ■

The concerns in the remaining examples reduce to determining the true rank of a matrix, which relies on examining a sequence of hypotheses (2.1) or (2.2). Our fifth example is directly related to the rank of demand systems, a notion developed by Gorman (1981) for exactly aggregable demand systems and generalized by Lewbel (1991) to all demand systems.

Example 2.2.5 (Consumer Demand)**.** An Engel curve is the function describing the allocation of an individual's consumption expenditures with the prices of all goods fixed, and the rank of a demand system is the dimension of the space spanned by the Engel curves of the system (Lewbel, 1991). Suppose that there are k goods in the system and the Engel

curve is given by

$$Y = \Gamma_0 G(Z) + u \ , \tag{2.13}$$

where Y is a $k \times 1$ vector of budget shares on the k goods, Z is total expenditure, $G(\cdot)$ is a $r_0 \times 1$ vector of unknown function with $r_0 \leq k$, and u is a vector zero mean random variables independent of Z. Assume Γ_0 is of full rank, then the rank r_0 of the demand system is equal to the rank of Γ_0. Let $Q(\cdot)$ be a $m \times 1$ vector of known functions with $m \geq k$. Then the rank of Γ_0 is equal to the rank of

$$\Pi_0 = E[Q(Z)Y^\intercal] \ , \tag{2.14}$$

if $E[Q(Z)G(Z)^\intercal]$ is of full rank. Thus, determining the rank r_0 of the demand system reduces to determining the rank of Π_0. The rank of the demand system provides evidence on consistency of consumer behaviors with utility maximization, and has implications for welfare comparisons and aggregation across goods and across consumers (Lewbel, 1991, 2006; Barnett and Serletis, 2008). ■

Factor analysis has been widely used in modeling variations, covariance and dynamics of time series (Anderson, 2003; Lam and Yao, 2012). Our next example shows the importance of matrix rank determination in identifying the number of factors in factor analysis.

Example 2.2.6 (Factor Analysis)**.** Let $Y \in \mathbf{R}^p$ be generated by the following model

$$Y = \mu_0 + \Lambda_0 F + u \ , \tag{2.15}$$

where F is a $r_0 \times 1$ vector of unobserved common factors with $E[F] = 0$ and $r_0 \leq p$, and u is an idiosyncratic error term with $E[u] = 0$. Assume $\mathrm{Var}(F)$ is of full rank, then the number r_0 of common factors is equal to the rank of $\mathrm{Var}(F)$. Let us write $Y = [Y_1^\intercal, Y_2^\intercal, Y_3^\intercal]^\intercal$ for $Y_1 \in \mathbf{R}^m$, $Y_2 \in \mathbf{R}^k$ and $Y_2 \in \mathbf{R}^{p-k-m}$ for some $r_0 \leq k \leq m < p$ and $m + k \leq p$. Write

$\Lambda_0 = [\Lambda_{0,1}^\intercal, \Lambda_{0,2}^\intercal, \Lambda_{0,3}^\intercal]^\intercal$ with $\Lambda_{0,1}$ and $\Lambda_{0,2}$ having m and k rows. Given the mild condition that u is independent of F and $E[uu']$ is diagonal, the rank of $\text{Var}(F)$ is equal to the rank of

$$\Pi_0 = \text{Cov}(Y_1, Y_2) \ , \tag{2.16}$$

if $\Lambda_{0,1}$ and $\Lambda_{0,1}$ are of full rank. Thus, determining the number r_0 of these common factors reduces to determining the rank of Π_0. Such a question also arises in the interbattery factor analysis (Gill and Lewbel, 1992), the dynamic analysis of time series (Lam and Yao, 2012), and finance and macroeconomics (Bai and Ng, 2002, 2007). ■

Our final example is taken from Gill and Lewbel (1992), and manifests how matrix rank determination is useful in model selection for ARMA processes and state space models.

Example 2.2.7 (Model Selection)**.** Let $\{Y_t\}$ be a $p \times 1$ weakly stationary time series, which has the following state space representation:

$$Y_t = \Gamma_0 Z_t + u_t \ , \ Z_t = \Lambda_0 Z_{t-1} + \epsilon_t \ , \tag{2.17}$$

where Z_t is a $r_0 \times 1$ vector of state variables, and u_t and ϵ_t are error terms. It turns out that the number r_0 of state variables is equal to the rank of the Hankel matrix

$$\Pi_0 = E(\begin{bmatrix} Y_{t+1} \\ \vdots \\ Y_{t+b} \end{bmatrix} \begin{bmatrix} Y_t^\intercal & \cdots & Y_{t-b+1}^\intercal \end{bmatrix}) \ , \tag{2.18}$$

for b sufficiently large (Aoki, 1990, p.52). Consequently, determining the number of state variables r_0 to model Y_t reduces to determining the rank of Π_0. When Y_t is a scalar and follows an ARMA(p_1, p_2) model, then Y_t has a state space representation with the number r_0 of state variables equal to $\max(p_1, p_2)$ (Aoki, 1990). Thus, determining the rank of the

Hankel matrix is crucial for model specification in these contexts. ■

2.2.2 Motivation

To proceed, we let $\alpha \in (0,1)$ be the nominal level and $\phi_n^{(r)}$ be any one of the existing rank tests designed for the hypotheses (2.2), which are reviewed in the introduction.[2] It has been well established in the literature that $\lim_{n\to\infty} P(\phi_n^{(r)} = 1) = \alpha$ under $\mathrm{H}_0^{(r)}$ and $\lim_{n\to\infty} P(\phi_n^{(r)} = 1) = 1$ under $\mathrm{H}_1^{(r)}$.

When $\mathrm{rank}(\Pi_0) < r$, the asymptotic distributions of test statistics have not been established and can be very different from those when $\mathrm{rank}(\Pi_0) = r$. On the one hand, $\phi_n^{(r)}$ may fail to control the asymptotic rejection rate. In Appendix 2.7.3, we prove that this is true for the Kleibergen and Paap (2006) version of $\phi_n^{(r)}$. Therefore, $\phi_n^{(r)}$ cannot be directly applied to test for the hypotheses (2.1). On the other hand, the asymptotic rejection rate of $\phi_n^{(r)}$ can be strictly below the nominal level, i.e, $\lim_{n\to\infty} P(\phi_n^{(r)} = 1) < \alpha$. In Appendix 2.7.3, we also prove that this is true for the Kleibergen and Paap (2006) version of $\phi_n^{(r)}$. By Theorem 2 of Cragg and Donald (1993), this is also true for the Cragg and Donald (1997) version of $\phi_n^{(r)}$. In view of this, $\phi_n^{(r)}$ may alternatively be conservative when directly applied to the hypotheses (2.1). Thus, the critical value may be adjusted to improve the power of $\phi_n^{(r)}$ for detecting H_1 when Π_0 is close to a matrix with rank strictly less than r.

Given that H_0 being false is equivalent to $\mathrm{H}_0^{(q)}$ being false for all $0 \leq q \leq r$, one may then consider implementing multiple existing rank tests in order to obtain tests for the hypotheses (2.1) such that the asymptotic null rejection rate is controlled. The multiple testing method is based on the decision rule $\phi_n = \prod_{q=0}^{r} \phi_n^{(q)}$, which means that H_0 is rejected if and only if $\mathrm{H}_0^{(q)}$ is rejected for all $0 \leq q \leq r$. In VAR models (see, for instance, Example 2.2.2), Johansen (1995, Chapter 12) used this method to test for inequality of cointegration rank. In stochastic discount factor models, Kleibergen and Paap (2006) employed this method to test for identification of the risk premia parameters. Indeed, the asymptotic null

[2]Rejection means $\phi_n^{(r)} = 1$ and acceptance means $\phi_n^{(r)} = 0$.

rejection rate of this method is controlled, since under H_0,

$$\lim_{n\to\infty} P(\phi_n = 1) = \lim_{n\to\infty} P(\phi_n^{(0)} = 1, \ldots, \phi_n^{(r)} = 1) \leq \lim_{n\to\infty} P(\phi_n^{(\text{rank}(\Pi_0))} = 1) = \alpha \ , \quad (2.19)$$

where the first inequality holds since $P(A) \leq P(B)$ for $A \subset B$. Moreover, this method is consistent, since under H_1,

$$\lim_{n\to\infty} P(\phi_n = 1) = \lim_{n\to\infty} P(\phi_n^{(0)} = 1, \ldots, \phi_n^{(r)} = 1) \geq 1 - \sum_{q=0}^{r} (1 - \lim_{n\to\infty} P(\phi_n^{(q)} = 1)) = 1 \ ,$$

where the inequality holds by the Boole's inequality.

Unfortunately, the multiple testing method can be conservative. When $\text{rank}(\Pi_0) < r$, the inequality of (2.19) becomes strict whenever $\lim_{n\to\infty} P(\phi_n^{(r)} = 1) < \alpha$. This is because

$$\lim_{n\to\infty} P(\phi_n = 1) = \lim_{n\to\infty} P(\phi_n^{(0)} = 1, \ldots, \phi_n^{(r)} = 1) \leq \lim_{n\to\infty} P(\phi_n^{(r)} = 1) < \alpha \ , \quad (2.20)$$

where the first inequality holds since $P(A) \leq P(B)$ for $A \subset B$. As mentioned above, this is true for the Cragg and Donald (1997) and Kleibergen and Paap (2006) version of $\phi_n^{(r)}$. Thus, the critical value of each $\phi_n^{(q)}$ may be adjusted to improve the power of the multiple testing method for detecting H_1 when Π_0 is close to a matrix with rank strictly less than r. Furthermore, due to the dependence among $\{\phi_n^{(q)}\}_{q=0}^{r}$ the inequality in both (2.19) and (2.20) may become strict. In view of this, power loss may occur in a complicated way.

To show the drawback of existing rank tests and the conservativeness of the multiple testing method, we focus on the Kleibergen and Paap (2006) test and present some simulation evidence.[3] We assume that

$$Z_i^\intercal = W_i^\intercal \Pi_0 + u_i^\intercal \ , i = 1, \ldots, n \ , \quad (2.21)$$

[3]Two main reasons for the focus are: the Kleibergen and Paap (2006) test is preferred in terms of assumptions and computation, and has the most citations (over $1,000$) among the existing rank tests according to Google Scholar.

with $W_i \overset{i.i.d.}{\sim} N(0, I_6)$, $u_i \overset{i.i.d.}{\sim} N(0, I_6)$ and $n = 1,000$. Let

$$\Pi_0 = \text{diag}(\mathbf{1}_{6-d}, \mathbf{0}_d) + \delta I_6 \text{ for } \delta \geq 0 \text{ and } d = 1, \ldots, 6 \ , \tag{2.22}$$

where $\mathbf{1}_{6-d}$ denotes a $(6-d) \times 1$ vector of ones and $\mathbf{0}_d$ denotes a $d \times 1$ vector of zeros. We examine the hypotheses (2.1) with $r = 5$, that is, we test whether Π_0 has full rank. The design of Π_0 implies H_0 is true if and only if $\delta = 0$. In particular, $\text{rank}(\Pi_0) = 6 - d$ under H_0, so $\text{rank}(\Pi_0) < r$ when $d \neq 1$ and $\text{rank}(\Pi_0) = r$ when $d = 1$. Thus, $d \neq 1$ represents the case when Π_0 is close to a matrix with rank strictly less than r, while $d = 1$ represents the regular case. From the above argument, it shall be expected that when $d \neq 1$, the Kleibergen and Paap (2006) test may over-reject H_0 when $\delta = 0$ or may be inefficient in detecting H_1 when $\delta > 0$. Moreover, the multiple testing method may be inefficient in detecting H_1 when $\delta > 0$. The value of δ represents how strong H_1 deviates away from H_0.

To implement the Kleibergen and Paap (2006) test and the multiple testing method, we estimate Π_0 by $\hat{\Pi}_n = \frac{1}{n}\sum_{i=1}^{n} W_i Z_i^{\intercal}$. See Appendix 2.7.3 for a review on the Kleibergen and Paap (2006) test. By the central limit theorem, the asymptotic distribution of $\hat{\Pi}_n$ is zero mean Gaussian with convergence rate $\sqrt{n}$ and all assumptions in Kleibergen and Paap (2006) are satisfied. Let the nominal level be 5%. The rejection rates, which are based on $10,000$ simulation replications, are plotted in Figures 2.1 and 2.2. We use KP-D to denote the Kleibergen and Paap (2006) test when directly applied and KP-M to denote the multiple testing method. First, as expected, the rejection rates of KP-M are no greater than the 5% nominal level when $\delta = 0$ and tend to one as δ increases for all cases. When $d = 1$, the null rejection rate is close to the 5% nominal level. When $d \neq 1$, however, the null rejection rates are far below the 5% nominal level. This suggests that KP-M may be conservative when $d \neq 1$. Indeed, the power curve shifts to right and more parts fall below the 5% nominal level as d increases. This hints a method of power improvement by dragging the curves to the left such that all of them are above the 5% nominal level. Similarly, as Figure 2.2 shows, KP-D has the same issue under the considered model. Note that the difference between the two methods in Figure 2.2 is negligible, despite the fact that KP-D is more

powerful.

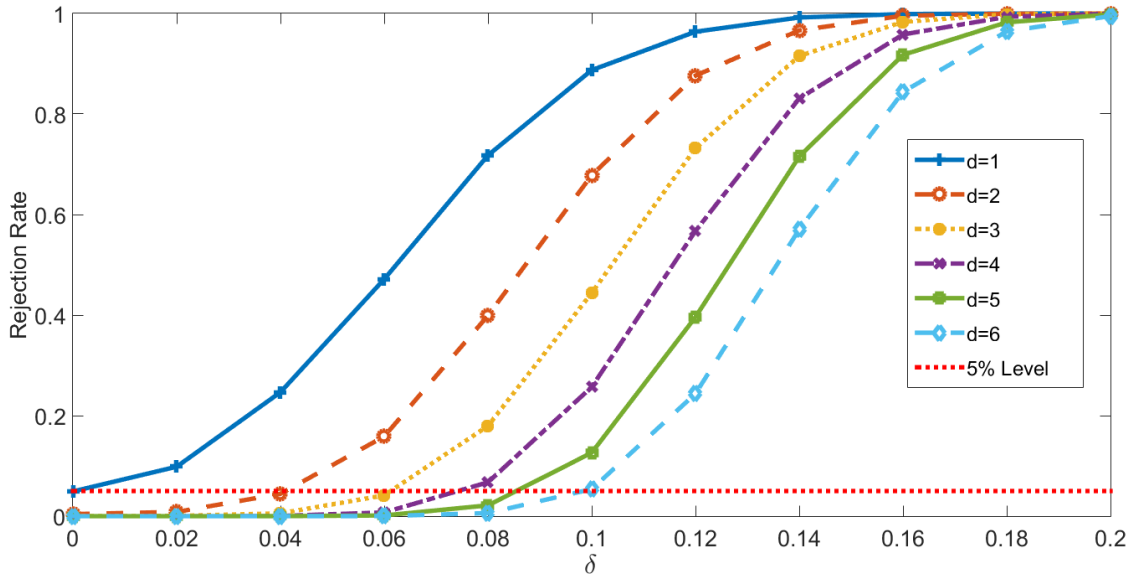


Figure 2.1: The rejection rate of the multiple testing method based on the Kleibergen and Paap (2006) test with 5% nominal level

2.3 Asymptotic Analysis

We can express the hypotheses (2.1) more tractably in terms of singular values. To see this, let $\sigma_1(\Pi_0) \geq \cdots \geq \sigma_k(\Pi_0) \geq 0$ be singular values of Π_0.[4] Then the rank of Π_0 is equal to the number of nonzero singular values of Π_0; see, for example, Problem 3.1.2 in Horn and Johnson (1991). It follows that the hypotheses (2.1) can be equivalently reformulated as

$$\mathrm{H}_0 : \sum_{j=r+1}^{k} \sigma_j^2(\Pi_0) = 0 \quad \text{v.s.} \quad \mathrm{H}_1 : \sum_{j=r+1}^{k} \sigma_j^2(\Pi_0) > 0 \ . \tag{2.23}$$

Given the reformulation in (2.23), it is convenient to study the differential properties of the map $\Pi_0 \mapsto \sum_{j=r+1}^{k} \sigma_j^2(\Pi_0)$. By leveraging the existing Delta method, we in turn establish the asymptotic distributions of the plug-in statistic $\sum_{j=r+1}^{k} \sigma_j^2(\hat{\Pi}_n)$ under the null for a given

[4] Recall that $\sigma_1^2(\Pi_0), \ldots, \sigma_k^2(\Pi_0)$ are numerically identical to eigenvalues of $\Pi_0^{\intercal}\Pi_0$.

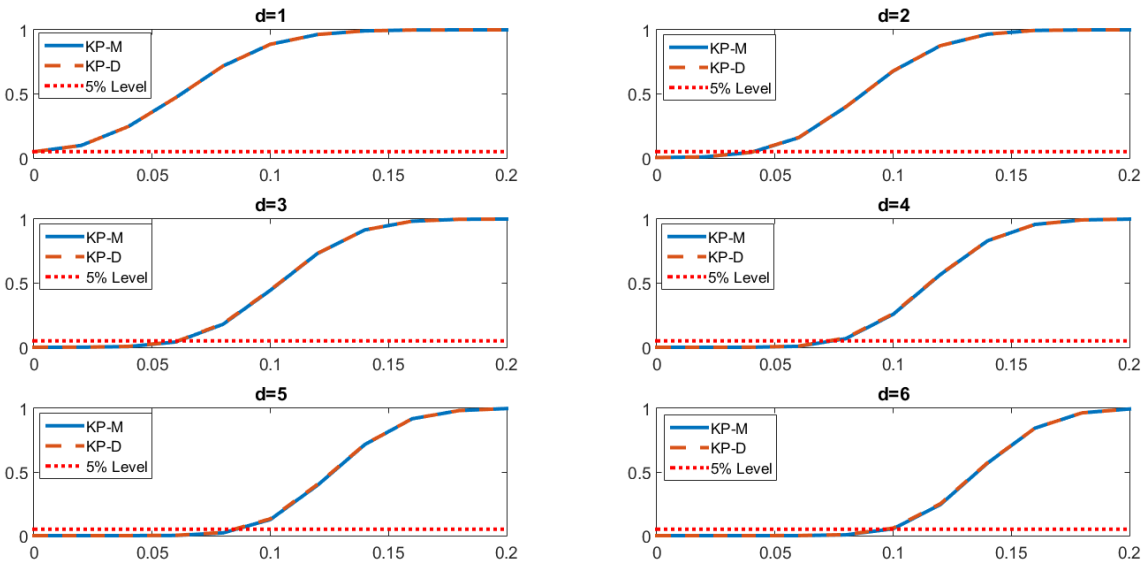


Figure 2.2: Comparison between the Kleibergen and Paap (2006) test and the multiple testing method based on it with 5% nominal level

estimator $\hat{\Pi}_n$ of Π_0. Since the resulting asymptotic distributions are highly nonstandard, we resort to the resampling procedure developed by Fang and Santos (2015) and Chen and Fang (2015) in order to obtain critical values.

2.3.1 Differential Properties

For ease of exposition, define $\phi : \mathbf{M}^{m\times k} \to \mathbf{R}$ by

$$\phi(\Pi) \equiv \sum_{j=r+1}^{k} \sigma_j^2(\Pi) \ , \tag{2.24}$$

where we recall that $\sigma_j(\Pi)$ is the jth largest singular value of Π. To derive the differentiability of ϕ, it shall prove useful to establish the following representation.

Lemma 2.3.1. *Let* $\mathbb{S}^{k\times q} \equiv \{U \in \mathbf{M}^{k\times q} : U^\intercal U = I_q\}$ *for* $q = 1, \ldots, k$. *Then we have*

$$\phi(\Pi) = \min_{U\in\mathbb{S}^{k\times(k-r)}} \|\Pi U\|^2 \ . \tag{2.25}$$

Lemma 2.3.1 shows that $\phi(\Pi)$ can be represented as a quadratic minimum over the

space of orthonormal matrices in $\mathbf{M}^{m\times(k-r)}$. The special case when $r = k-1$ – a test of Π having full rank – is a well known implication of the classical Courant-Fischer theorem, i.e., $\sigma_k^2(\Pi) = \min_{\|U\|=1} \|\Pi U\|^2$. Note that the minimum in (2.25) is achieved and hence well defined.

We are now in a position to analyze the differential properties of ϕ. It turns out that ϕ is not fully differentiable but belongs to a class of directionally differentiable maps. For completeness, we next introduce the appropriate notions of differentiability.

Definition 2.3.1. Let $\mathbf{M}^{m\times k}$ be equipped with the norm $\|\cdot\|$ and $\varphi : \mathbf{M}^{m\times k} \to \mathbf{R}$.

(i) The map φ is said to be *Hadamard directionally differentiable* at $\Pi \in \mathbf{M}^{m\times k}$ if there is a map $\varphi'_\Pi : \mathbf{M}^{m\times k} \to \mathbf{R}$ such that:

$$\lim_{n\to\infty} \frac{\varphi(\Pi + t_n M_n) - \varphi(\Pi)}{t_n} = \varphi'_\Pi(M) \ , \tag{2.26}$$

for all sequences $\{M_n\} \subset \mathbf{M}^{m\times k}$ and $\{t_n\} \subset \mathbf{R}_+$ such that $t_n \downarrow 0$, and $M_n \to M \in \mathbf{M}^{m\times k}$ as $n \to \infty$.

(ii) Suppose that $\varphi : \mathbf{M}^{m\times k} \to \mathbf{R}$ is Hadamard directionally differentiable at $\Pi \in \mathbf{M}^{m\times k}$. We say that φ is *second order Hadamard directionally differentiable* at $\Pi \in \mathbf{M}^{m\times k}$ if there is a map $\varphi''_\Pi : \mathbf{M}^{m\times k} \to \mathbf{R}$ such that:

$$\lim_{n\to\infty} \frac{\varphi(\Pi + t_n M_n) - \varphi(\Pi) - t_n \phi'_\Pi(M_n)}{t_n^2} = \varphi''_\Pi(M) \ , \tag{2.27}$$

for all sequences $\{M_n\} \subset \mathbf{M}^{m\times k}$ and $\{t_n\} \subset \mathbf{R}^+$ such that $t_n \downarrow 0$, and $M_n \to M \in \mathbf{M}^{m\times k}$ as $n \to \infty$.

Compared with Hadamard full differentiability (van der Vaart, 1998) which requires continuity and linearity of the derivative, the directional derivative is generally nonlinear though necessarily continuous. In fact, linearity is the exact gap between these two notions of differentiability. Remarkably, the Delta method remains valid even if ϕ is only Hadamard

directionally differentiable. We refer the readers to Shapiro (1990, 1991), Dümbgen (1993), and a recent review by Fang and Santos (2015) for further details. Unfortunately, as shall be proved, the asymptotic distribution of our statistic $\phi(\hat{\Pi}_n)$ implied by the Delta method is degenerate under the null, which creates substantial challenges for inference. This motivates the second order Hadamard directional differentiability. Compared with second order Hadamard full differentiability which requires a quadratic form of the derivative corresponding to a bilinear map, the directional derivative ϕ''_θ is generally nonquadratic though continuous. In fact, quadratic form structure is the exact gap between these two notions of differentiability. Similarly, the second order Delta method remains valid even if ϕ is only second order Hadamard directionally differentiable. We refer the readers to Shapiro (2000) and a recent review by Chen and Fang (2015) for further details.

The following proposition establishes the differentiability of ϕ.

Proposition 2.3.1. *Let $\phi : \mathbf{M}^{m\times k} \to \mathbf{R}$ be defined as in* (2.24).

(i) ϕ is first order Hadamard directionally differentiable at any $\Pi \in \mathbf{M}^{m\times k}$ with the derivative $\phi'_\Pi : \mathbf{M}^{m\times k} \to \mathbf{R}$ given by

$$\phi'_\Pi(M) = \min_{U\in\Psi(\Pi)} 2\mathrm{tr}(U^\intercal \Pi^\intercal M U) \ , \tag{2.28}$$

where $\Psi(\Pi) \equiv \arg\min_{U\in\mathbb{S}^{k\times(k-r)}} \|\Pi U\|^2$.

(ii) ϕ is second order Hadamard directionally differentiable at any $\Pi \in \mathbf{M}^{m\times k}$ satisfying $\phi(\Pi) = 0$ with the derivative $\phi''_\Pi : \mathbf{M}^{m\times k} \to \mathbf{R}$ given by

$$\phi''_\Pi(M) = \min_{U\in\Psi(\Pi)} \min_{V\in\mathbf{M}^{k\times(k-r)}} \|MU + \Pi V\|^2 \ . \tag{2.29}$$

Proposition 2.3.1 implies that ϕ is Hadamard directionally differentiable at any $\Pi \in \mathbf{M}^{m\times k}$. In particular, when $\mathrm{rank}(\Pi) \le r$, it exhibits a degenerate derivative, i.e., $\phi'_\Pi(\mathrm{M}) = 0$ for all $M \in \mathbf{M}^{m\times k}$. Moreover, Proposition 2.3.1 implies that ϕ is second order Hadamard directionally differentiable at any $\Pi \in \mathbf{M}^{m\times k}$ with $\mathrm{rank}(\Pi) \le r$. In general, ϕ

is not second order fully Hadamard differentiable at $\Pi \in \mathbf{M}^{m\times k}$ with $\text{rank}(\Pi) \leq r$ unless $\text{rank}(\Pi) = r$, see Lemma 2.7.4. Thus, the accommodation of $\text{rank}(\Pi) < r$ causes the irregularity of ϕ.

To conclude this section, we provide a simplified analytical expression for ϕ''_Π. Let $\Pi = P\Sigma Q^\intercal$ be a singular value decomposition of Π, where $P \in \mathbb{S}^{m\times m}$ and $Q \in \mathbb{S}^{k\times k}$, and $\Sigma \in \mathbf{M}^{m\times k}$ is diagonal with diagonal entries in descending order. Let $r^* \equiv \text{rank}(\Pi)$. Write $P = [P_1, P_2]$ and $Q = [Q_1, Q_2]$ for $P_1 \in \mathbf{M}^{m\times r^*}$ and $Q_1 \in \mathbf{M}^{k\times r^*}$, respectively. Thus, the columns of P_2 and Q_2 are the left-singular vectors and right-singular vectors of Π associated with the zero singular values, respectively. Then the following proposition gives a simplified analytical expression of ϕ''_Π.

Proposition 2.3.2. *Suppose $r^* \leq r$ and let $\phi''_\Pi : \mathbf{M}^{m\times k} \to \mathbf{R}$ be given as in Proposition 2.3.1. Then for $M \in \mathbf{M}^{m\times k}$,*

$$\phi''_\Pi(M) = \sum_{j=r-r^*+1}^{k-r^*} \sigma_j^2(P_2^\intercal M Q_2) \ . \tag{2.30}$$

Proposition 2.3.2 implies $\phi''_\Pi(M)$ is the sum of the $k - r$ smallest squared singular values of transformed matrix $P_2^\intercal M Q_2$. Observe that P_2 and Q_2 are from singular value decomposition, so calculation of the derivative requires no more than calculation of singular value decomposition as in the test statistic. As we will see later, this facilitates the computation of our test statistic and makes our test procedure attractive. Note P_2 and Q_2 can be chosen up to postmultiplication by $(m - r^*) \times (m - r^*)$ and $(k - r^*) \times (k - r^*)$ orthonormal matrices, respectively, but the term on the right hand side of (2.30) is invariant to the choice of P_2 and Q_2.

2.3.2 The Asymptotic Distributions

Given the established differentiability of ϕ and null first order derivative, the asymptotic distribution of $\phi(\hat{\Pi}_n)$ can be easily obtained by the second order Delta method

(Shapiro, 2000), provided $\hat{\Pi}_n$ converges weakly. Towards this end, we impose the following assumption.

Assumption 2.3.1. *Let* $\Pi_0 \in \mathbf{M}^{m\times k}$ *and there are* $\hat{\Pi}_n : \{X_i\}_{i=1}^n \to \mathbf{M}^{m\times k}$ *such that* $\tau_n(\hat{\Pi}_n - \Pi_0) \xrightarrow{L} \mathcal{M}$ *for some* $\tau_n \uparrow \infty$ *and random matrix* $\mathcal{M} \in \mathbf{M}^{m\times k}$.

Assumption 2.3.1 imposes that the estimator $\hat{\Pi}_n$ for Π_0 admits a weak limit $\mathcal{M} \in \mathbf{M}^{m\times k}$ at a scalar rate τ_n. The estimator $\hat{\Pi}_n$ is defined as a function of the data $\{X_i\}_{i=1}^n$ into $\mathbf{M}^{m\times k}$, and the weak convergence "$\xrightarrow{L}$" is understood with respect to the joint law of $\{X_i\}_{i=1}^n$, which need not be i.i.d.. In particular, τ_n is allowed to be any parametric or nonparametric rate that covers all the above examples.

Let $\Pi_0 = P_0\Sigma_0 Q_0^\intercal$ be a singular value decomposition of Π_0, where $P_0 \in \mathbb{S}^{m\times m}$ and $Q_0 \in \mathbb{S}^{k\times k}$, and $\Sigma_0 \in \mathbf{M}^{m\times k}$ is diagonal with diagonal entries in descending order. Let $r_0 \equiv \text{rank}(\Pi_0)$. Write $P_0 = [P_{0,1}, P_{0,2}]$ and $Q_0 = [Q_{0,1}, Q_{0,2}]$ for $P_{0,1} \in \mathbf{M}^{m\times r_0}$ and $Q_{0,1} \in \mathbf{M}^{k\times r_0}$, respectively. Thus, the columns of $P_{0,2}$ and $Q_{0,2}$ are the left-singular vectors and right-singular vectors of Π_0 associated with the zero singular values, respectively. The following proposition delivers the asymptotic distributions of $\phi(\hat{\Pi}_n)$.

Proposition 2.3.3. *Suppose Assumption 2.3.1 holds. Then we have*

$$\tau_n(\phi(\hat{\Pi}_n) - \phi(\Pi_0)) \xrightarrow{L} \min_{U\in\Psi(\Pi_0)} 2\text{tr}(U^\intercal \Pi_0^\intercal \mathcal{M} U) \ , \tag{2.31}$$

and under H_0*,*

$$\tau_n^2 \phi(\hat{\Pi}_n) \xrightarrow{L} \sum_{j=r-r_0+1}^{k-r_0} \sigma_j^2(P_{0,2}^\intercal \mathcal{M} Q_{0,2}) \ . \tag{2.32}$$

Proposition 2.3.3 implies that $\tau_n\phi(\hat{\Pi}_n)$ converges in distribution to a degenerate limit at 0 under H_0. This prevents us from making inference based on the first order framework (Chen and Fang, 2015). Proposition 2.3.3 also implies that $\tau_n^2\phi(\hat{\Pi}_n)$ converges in distribution to a generally nondegenerate limit under H_0. This enables us to make inference based on the second order framework. The limit is a nonlinear function of the weak limit

$\mathcal{M}$ and remarkably nonstandard especially when $r_0 < r$. In general, an analytical (pivotal) distribution is not available. Note that $P_{0,2}$ and $Q_{0,2}$ are identified up to postmultiplication by $(m - r_0) \times (m - r_0)$ and $(k - r_0) \times (k - r_0)$ orthonormal matrices, respectively, but the term on the right hand side of (2.32) is invariant to the choice of $P_{2,0}$ and $Q_{2,0}$.

In order to see how our results apply to various settings, we now turn to examples introduced in Section 2.2.1. We shall focus on Examples 2.2.1 and 2.2.3 exclusively for conciseness; Examples 2.2.2 and 2.2.4-2.2.7 will be treated in Appendix 2.7.2. In particular, Assumption 2.3.1 is not well satisfied in Example 2.2.2 since the convergence rates of $\hat{\Pi}_n$ are not homogenous across its columns, and we extend the result in Proposition 2.3.3 for it.

Example 2.2.1 (Continued). Suppose $\{W_i, Z_i\}_{i=1}^n$ is a sequence of data from Example 2.2.1. Let $\hat{\Pi}_n$ be the method of moment estimator

$$\hat{\Pi}_n = \frac{1}{n} \sum_{i=1}^{n} W_i Z_i^\intercal \ . \tag{2.33}$$

Under certain weak dependence and moment condition, the central limit theorem implies that Assumption 2.3.1 is satisfied with $\tau_n = \sqrt{n}$ and $\mathcal{M}$ being a zero mean Gaussian. When $r_0 < k - 1$, the asymptotic distribution of $n\phi(\hat{\Pi}_n)$ can be highly nonstandard. ■

Example 2.2.3 (Continued). Suppose $\{Y_t\}_{t=1}^n$ is a sequence of data from Example 2.2.3. Let $\hat{\Pi}_n$ be a kernel HAC estimator

$$\hat{\Pi}_n = \sum_{j=-n+1}^{n-1} k(\frac{j}{b_n}) \hat{\Gamma}_n(j) \ , \tag{2.34}$$

where $\hat{\Gamma}_n(j) \equiv \frac{1}{n} \sum_{t=1}^{n-j} \Delta Y_t \Delta Y_{t+j}^\intercal$ for $j \geq 0$, $\hat{\Gamma}_n(j) = \hat{\Gamma}_n(-j)^\intercal$ for $j < 0$, $k(\cdot)$ is a kernel function, and b_n is a bandwidth parameter. Under certain weak dependence and moment conditions, $\hat{\Pi}_n$ is asymptotically normal at the rate $\sqrt{n/b_n}$. For example, see Hannan (1970), Brillinger (1981), Priestley (1981) and Berkes et al. (2016). So, Assumption 2.3.1 is satisfied with $\tau_n = \sqrt{n/b_n}$ and $\mathcal{M}$ being a zero mean Gaussian. In testing for the existence of cointegration, when $r_0 < k - 1$, the asymptotic distribution of $n\phi(\hat{\Pi}_n)/b_n$ can be highly

nonstandard. ■

We now discuss the result of Proposition 2.3.3 when $r_0 = r$ and its relation to the literature. In this case, $P_{0,2}^\intercal \mathcal{M} Q_{0,2}$ has $k-r$ columns and $\sum_{j=r-r_0+1}^{k-r_0} \sigma_j^2(P_{0,2}^\intercal \mathcal{M} Q_{0,2})$ is equal to the Frobenius norm of $P_{0,2}^\intercal \mathcal{M} Q_{0,2}$. Thus, the asymptotic distribution in (2.32) becomes

$$\|P_{0,2}^\intercal \mathcal{M} Q_{0,2}\|^2 = \mathrm{vec}(P_{0,2}^\intercal \mathcal{M} Q_{0,2})^\intercal \mathrm{vec}(P_{0,2}^\intercal \mathcal{M} Q_{0,2}) \ . \tag{2.35}$$

When $\mathcal{M}$ is a zero mean Gaussian, the limit is a weighted sum of independent $\chi^2(1)$ random variables. Thus, Proposition 2.3.3 includes Robin and Smith (2000) as a special case. If, in addition, the covariance matrix of $\mathrm{vec}(P_{0,2}^\intercal \mathcal{M} Q_{0,2})$ is nonsingular, Kleibergen and Paap (2006) proved that a normalized version of $\tau_n^2 \phi(\hat{\Pi}_n)$ has a $\chi^2((m-r)(k-r))$ asymptotic distribution under $\mathrm{H}_0^{(r)}$. The asymptotic distribution is not a χ^2-type distribution any more if $r_0 < r$. This suggests that the Robin and Smith (2000) test when directly applied to (2.1) may fail to control the asymptotic null rejection rate.

2.3.3 The Bootstrap

Given the nonstandard asymptotic distribution in Proposition 2.3.3, no analytical critical values can be employed for inference. We may resort to the standard bootstrap method (Efron, 1979) to consistently estimate the asymptotic distribution. Unfortunately, the consistency of this method fails due to the degeneracy of ϕ'_{Π_0} under the null (Chen and Fang, 2015). Moreover, the recentered bootstrap does not necessarily correct the inconsistency due to the nondifferentiability of ϕ. As such, we resort to the procedure developed by Chen and Fang (2015) for construction of critical values. See the discussion on m out of n bootstrap and subsampling in Remark 2.3.1.

Recall that the asymptotic distribution is a composition of $\mathcal{M}$ and ϕ''_{Π_0}. Our proposed procedure consists of first estimating $\mathcal{M}$ by bootstrap and then estimating ϕ''_{Π_0}. For the former, let $\hat{\Pi}_n^*$ denote a "bootstrapped version" of $\hat{\Pi}_n$, which is defined as a function of the data $\{X_i\}_{i=1}^n$ and random weights $\{W_i\}_{i=1}^n$ that are independent of $\{X_i\}_{i=1}^n$ into $\mathbf{M}^{m\times k}$.

This general definition allows us to include special cases such as nonparametric, Bayesian, block, score, more generally multiplier and exchangeable bootstrap. To accommodate diverse resampling schemes, we simply impose the following high level condition.

Assumption 2.3.2. *(i)* $\hat{\Pi}_n^* : \{X_i, W_i\}_{i=1}^n \to \mathbf{M}^{m\times k}$ *with* $\{W_i\}_{i=1}^n$ *independent of* $\{X_i\}_{i=1}^n$*; (ii)* $\tau_n(\hat{\Pi}_n^* - \hat{\Pi}_n) \xrightarrow{L^*} \mathcal{M}$ *almost surely, where* $\xrightarrow{L^*}$ *denotes weak convergence with respect to the joint law of* $\{W_i\}_{i=1}^n$ *conditional on* $\{X_i\}_{i=1}^n$.

Assumption 2.3.2(i) defines the bootstrap analog $\hat{\Pi}_n^*$ of $\hat{\Pi}_n$, while Assumption 2.3.2(ii) simply imposes the consistency of the law of $\tau_n(\hat{\Pi}_n^* - \hat{\Pi}_n)$ conditional on the data $\{X_i\}_{i=1}^n$ for the law of $\mathcal{M}$, i.e., the bootstrap works for the estimator $\hat{\Pi}_n$.

Next we examine Assumption 2.3.2 in Examples 2.2.1 and 2.2.3; Examples 2.2.2 and 2.2.4-2.2.7 will be treated in Appendix 2.7.2. In particular, Assumption 2.3.2 is not well satisfied in Example 2.2.2, and we extend the result in Theorem 2.3.1 for it.

Example 2.2.1 (Continued). Let $\{Z_i^*, W_i^*\}_{i=1}^n$ be obtained by nonparametric bootstrapping $\{Z_i, W_i\}_{i=1}^n$ when $\{Z_i, W_i\}_{i=1}^n$ is a sequence of i.i.d. data, and by block bootstrapping $\{Z_i, W_i\}_{i=1}^n$ when $\{Z_i, W_i\}_{i=1}^n$ is a sequence of dependent data. Under certain weak dependence and moment condition, Assumption 2.3.2 is satisfied with

$$\hat{\Pi}_n^* = \frac{1}{n}\sum_{i=1}^n W_i^* Z_i^{*\intercal} \ . \tag{2.36}$$

Multiplier and exchangeable bootstrap may also be employed for i.i.d. data. ■

Example 2.2.3 (Continued). Since $\hat{\Pi}_n$ only depends on $\{\Delta Y_t\}_{t=1}^n$, it suffices to resample $\{\Delta Y_t\}_{t=1}^n$. Note that $\{\Delta Y_t\}_{t=1}^n$ is stationary. Let $\{\Delta Y_t^*\}_{t=1}^n$ be obtained by block bootstrapping $\{\Delta Y_t\}_{t=1}^n$. Under certain weak dependence and moment condition, Assumption 2.3.2 is satisfied with

$$\hat{\Pi}_n^* = \sum_{j=-n+1}^{n-1} k(\frac{j}{b_n})\hat{\Gamma}_n^*(j) \ , \tag{2.37}$$

where $\hat{\Gamma}_n^*(j) \equiv \frac{1}{n}\sum_{t=1}^{n-j} \Delta Y_t^* \Delta Y_{t+j}^{*\intercal}$ for $j \geq 0$, $\hat{\Gamma}_n^*(j) = \hat{\Gamma}_n^*(-j)^\intercal$ for $j < 0$, $k(\cdot)$ and b_n are the same kernel function and bandwidth parameter. See Politis and Romano (1992, 1993) and Politis et al. (1992) for other bootstrap procedures. ■

There are two main methods for estimating ϕ''_{Π_0}: the structure-exploiting approach and the numerical differentiation approach. For the former, we describe how to estimate ϕ''_{Π_0} according to (2.30). Let $\hat{\Pi}_n = \hat{P}_n \hat{\Sigma}_n \hat{Q}_n^\intercal$ be a singular value decomposition of $\hat{\Pi}_n$, where $\hat{P}_n \in \mathbb{S}^{m\times m}$ and $\hat{Q}_n \in \mathbb{S}^{k\times k}$, and $\hat{\Sigma}_n \in \mathbf{M}^{m\times k}$ is diagonal with diagonal entries in descending order. Let $\hat{r}_n \equiv \min\{r, \#\{1 \leq j \leq k : \sigma_j(\hat{\Pi}_n) \geq \kappa_n\}\}$, where $\kappa_n \downarrow 0$ is a tuning parameter that is required to satisfy certain conditions below.[5] Write $\hat{P}_n = [\hat{P}_{1,n}, \hat{P}_{2,n}]$ and $\hat{Q}_n = [\hat{Q}_{1,n}, \hat{Q}_{2,n}]$ for $\hat{P}_{1,n} \in \mathbf{M}^{m\times \hat{r}_n}$ and $\hat{Q}_{1,n} \in \mathbf{M}^{k\times \hat{r}_n}$, respectively. By (2.30), we may estimate ϕ''_{Π_0} by

$$\hat{\phi}''_n(M) = \sum_{j=r-\hat{r}_n+1}^{k-\hat{r}_n} \sigma_j^2(\hat{P}_{2,n}^\intercal M \hat{Q}_{2,n}) \ . \tag{2.38}$$

Note that $\hat{P}_{2,n}$ and $\hat{Q}_{2,n}$ can be chosen up to postmultiplication by $(m-\hat{r}_n)\times(m-\hat{r}_n)$ and $(k-\hat{r}_n)\times(k-\hat{r}_n)$ orthonormal matrices, respectively, but the term on the right hand side of (2.38) is invariant to the choice of $\hat{P}_{2,n}$ and $\hat{Q}_{2,n}$. For the latter, we estimate ϕ''_{Π_0} by

$$\hat{\phi}''_n(M) = \frac{\phi(\hat{\Pi}_n + \kappa_n M) - \phi(\hat{\Pi}_n)}{\kappa_n^2} \ . \tag{2.39}$$

Remark 2.3.1. In effect, m out of n bootstrap and subsampling amounts to estimating $\mathcal{M}$ based on subsamples (with and without replacement, respectively) and ϕ''_{Π_0} via the numerical differentiation approach, in which case the tuning parameters for choosing subsamples and estimation of the derivative coincide. Thus, our bootstrap procedure can be more efficient in two ways. First, our bootstrap procedure makes efficient use of the data in estimating $\mathcal{M}$, since it is based on full samples. Second, our bootstrap procedure also provides alternative method of estimating ϕ''_{Π_0} by exploiting more structural information of

[5] We use $\#A$ to denote the cardinality of a set A. One can theoretically ignore r in the expression of $\hat{r}_n$. However, taking minimum in the expression of $\hat{r}_n$ is a way of imposing the information under the null to ensure that the estimator in (2.38) is well defined and improve power.

the data .

Given a suitable condition on $\kappa_n \downarrow 0$, we are then able to prove that the law of the weak limit in (2.32) is consistently estimated by the law of $\hat{\phi}_n''(\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\})$ conditional on the data. It in turn suggests employing the $1-\alpha$ quantile $\hat{c}_{1-\alpha}$ of $\hat{\phi}_n''(\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\})$ conditional on the data:[6]

$$\hat{c}_{1-\alpha} \equiv \inf\{c \in \mathbf{R} : P_W(\hat{\phi}_n''(\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\}) \leq c) \geq 1-\alpha\} \ . \tag{2.40}$$

Note that $\hat{c}_{1-\alpha}$ is generally infeasible in that it is constructed based on the "exact" distribution of $\hat{\phi}_n''(\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\})$ conditional on the data. Nonetheless, it can be estimated by Monte Carlo simulation and the estimation error can be made arbitrarily small by choosing the number of bootstrap replications (Efron, 1979; Hall, 1992; Horowitz, 2001).

For each realization of $\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\}$, the computation of $\hat{\phi}_n''(\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\})$ requires no more than calculating singular value decompositions with $\hat{\phi}_n''$ in (2.38) and (2.39). When $\hat{\phi}_n''$ is given in (2.38), it is only necessary to calculate the singular value decomposition of $\hat{P}_{2,n}^{\intercal}\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\}\hat{Q}_{2,n}$. When $\hat{\phi}_n''$ is given in (2.39), it is only necessary to calculate the singular value decomposition of $\hat{\Pi}_n + \kappa_n\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\}$. Thus, the computation of simulated critical values is as simple as the computation of the test statistic. Comparisons between the estimators in (2.38) and (2.39) will be investigated in Monte Carlo studies.

The following theorem establishes that the test of rejecting H_0 when $\tau_n^2\phi(\hat{\Pi}_n) > \hat{c}_{1-\alpha}$ controls the asymptotic null rejection rate and is consistent.

Theorem 2.3.1. *Suppose Assumptions 2.3.1 and 2.3.2 hold. Let $\kappa_n \downarrow 0$ and $\tau_n\kappa_n \to \infty$. Let $\hat{c}_{1-\alpha}$ be given in* (2.40) *with $\hat{\phi}_n''$ in* (2.38) *or* (2.39)*. If the cdf of the limit distribution in* (2.32) *is continuous and strictly increasing at its $1-\alpha$ quantile for $\alpha \in (0,1)$, then under* H_0*,*

$$\lim_{n\to\infty} P(\tau_n^2\phi(\hat{\Pi}_n) > \hat{c}_{1-\alpha}) = \alpha \ .$$

[6] P_W denotes the probability with respect to the joint law of the random weights $\{W_i\}_{i=1}^n$.

Furthermore, under H_1,

$$\lim_{n\to\infty} P(\tau_n^2 \phi(\hat{\Pi}_n) > \hat{c}_{1-\alpha}) = 1 \ .$$

Theorem 2.3.1 implies that our tests have the asymptotic null rejection rate that is exactly equal to the nominal level, regardless of whether $r_0 = r$ or $r_0 < r$. This stems from the design of our bootstrap that estimates the asymptotic distribution pointwise in Π_0. In contrast to existing rank tests and the multiple testing method that may have the asymptotic null rejection rate strictly below the nominal level when $r_0 < r$, this distinct feature shall make our tests more powerful. In particular, when Π_0 is close to a matrix with rank strictly less than r, our tests shall be more powerful in detecting H_1 than existing rank tests and the multiple testing method. In addition, in contrast to existing rank tests that may fail to control the asymptotic null rejection rate when $r_0 < r$, our tests control the asymptotic null rejection rate regardless of whether $r_0 = r$ or $r_0 < r$. Theorem 2.3.1 also implies that our tests are consistent.

Several simple, new and powerful tests are immediate from Theorem 2.3.1. First, applying Theorem 2.3.1 to Examples 2.2.1 yields tests for identification in linear IV models. Second, applying Theorem 2.3.1 to Examples 2.2.2 and 2.2.3 yields tests for the existence of stochastic trend and/or cointegration with or without VAR specification, respectively. Third, applying Theorem 2.3.1 to Examples 2.2.4 yields tests for the existence of common features.

We now discuss the quantile requirement on the limit distribution in (2.32) imposed in Theorem 2.3.1. A necessary condition for that requirement to hold is $P_{0,2}^{\intercal}\mathcal{M}Q_{0,2} \neq 0$ with positive probability, that is,

$$P(\mathcal{R}(\mathcal{M}) \cap \mathcal{N}(\Pi_0^{\intercal}) \neq \emptyset) > 0 \text{ and } P(\mathcal{R}(\mathcal{M}^{\intercal}) \cap \mathcal{N}(\Pi_0) \neq \emptyset) > 0 \ ,$$

where $\mathcal{R}(A)$ denotes the range of a matrix A and $\mathcal{N}(A)$ denotes the null space of a matrix A. When $\mathcal{M}$ is zero mean Gaussian and $r_0 = r$, the limit in (2.32) is a weighted sum of independent $\chi^2(1)$ random variables as shown in (2.35). This implies the limit distribution

is continuous, unless the covariance matrix of $\text{vec}(P_{0,2}^{\intercal}\mathcal{M}Q_{0,2})$ is zero. Thus, in this special case, the sufficient and necessary condition for the requirement to hold is nonzero of the covariance matrix of $\text{vec}(P_{0,2}^{\intercal}\mathcal{M}Q_{0,2})$. In contrast, Kleibergen and Paap (2006) requires nonsingularity of the covariance matrix of $\text{vec}(P_{0,2}^{\intercal}\mathcal{M}Q_{0,2})$. In view of this, our tests rely on much weaker conditions than Kleibergen and Paap (2006).

Remark 2.3.2. The requirement on the limit distribution in (2.32) imposed in Theorem 2.3.1 may not be satisfied in testing for perfect multicollinearity in Example 2.2.1. When $\hat{\Pi}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^{\intercal}$, then the limit in (2.32) is degenerate at zero, which can be best seen from (2.32) since $\mathcal{M}Q_{0,2} = 0$. Heuristically, if the smallest singular value of Π_0 is zero, then $\lambda^{\intercal}Z_i$ is constantly zero for some constant $\lambda \in \mathbb{S}^k$ and the smallest singular value of $\hat{\Pi}_n$ is constantly zero. Nevertheless, one can easily prove that the properties of size control and consistency continue to hold. ■

2.4 Simulations and Applications

In this section, we first conduct Monte Carlo studies to examine the finite sample performance of our tests, and show how existing rank tests when directly applied to (2.1) and the multiple testing method may be conservative. We then apply our tests to study identification in stochastic discount factor models (Jagannathan and Wang, 1996). Lastly, we demonstrate how our tests can improve the accuracy of the sequential testing procedure for rank determination.

2.4.1 Simulation Studies

We start with the performance of our tests for the problem in Section 2.2.2. To implement our tests, we use the same estimator $\hat{\Pi}_n$ as in Section 2.2.2 and the same nominal level 5%. The rejection rates, which are based on $10,000$ simulation replications with 500 nonparametric i.i.d. bootstrap replications for each Monte Carlo, are plotted in Figure 2.3. Clearly, Assumptions 2.3.1 and 2.3.2 are satisfied. The result is based on the derivative

estimator in (2.38) with $\kappa_n = n^{-1/4}$, while the result for the derivative estimator in (2.38) with $\kappa_n = n^{-1/3}$ is similar and available upon request. For ease of comparison, we combine Figures 2.1 and 2.3 to yield Figure 2.4, where CF denotes our tests and KP-M is defined in Section 2.2.1. In contrast to KP-M, the null rejection rates of CF are close to the 5% nominal level for all d as shown in Figure 2.3. As expected from Theorem 2.3.1, CF are more powerful than KP-M uniformly over $d \neq 1$ and all $\delta > 0$ as shown in Figure 2.4. In particular, in contrast to KP-M, all power curves of CF lie above the 5% nominal level line. Note the power curves do not coincide since the data generating process (DGP) is varied for different d. Figure 2.4 also shows that the greater the value of d is, the greater the power improvement is. In addition, when $d = 1$, CF are as powerful as KP-M. Thus, these findings confirm that KP-M are too conservative, and CF provide power improvement over KP-M. Given Figure 2.2, the comparison between CF and KP-D is the same.

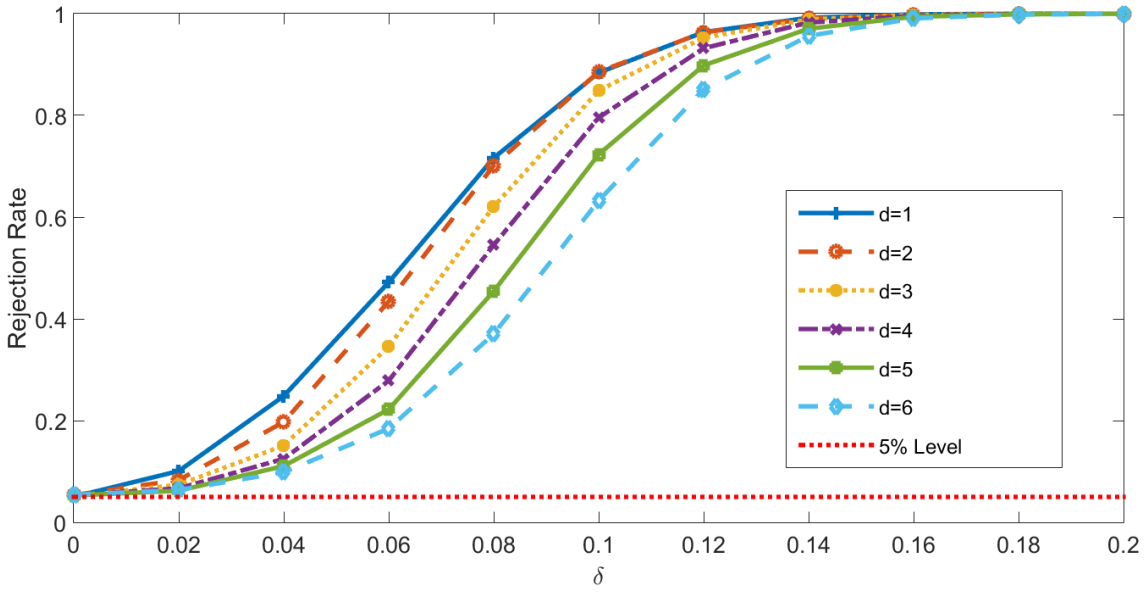


Figure 2.3: The rejection rate of our tests with 5% nominal level

We next investigate the finite sample performance of our tests, the Kleibergen and Paap (2006) test when directly applied, and the multiple testing method in more complicated DGPs with heteroskedasticity, dependence and different sample sizes. We consider

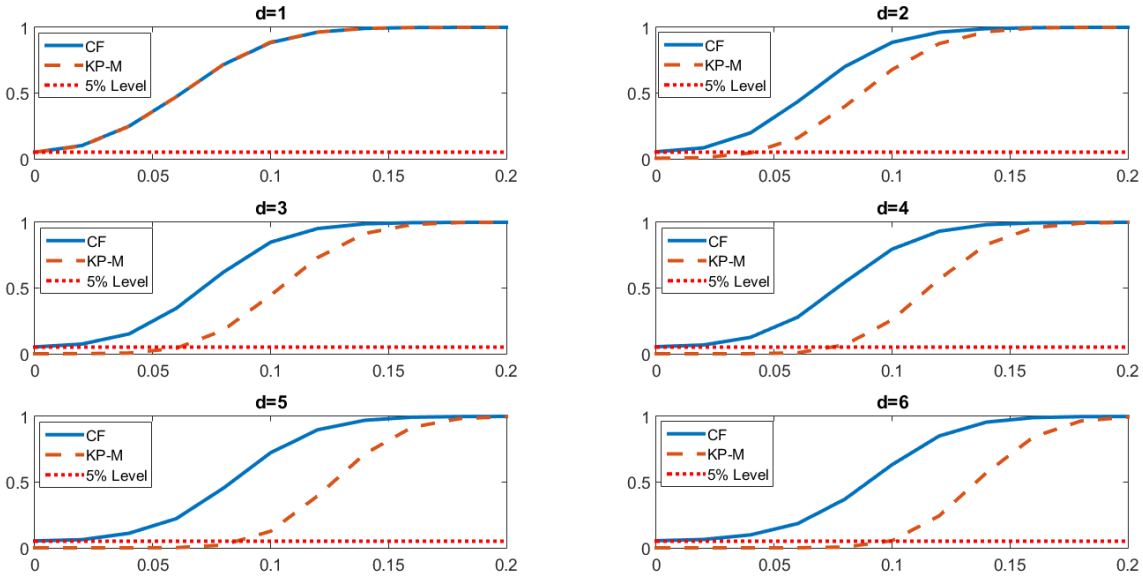


Figure 2.4: Comparison between our tests and the multiple testing method based on the Kleibergen and Paap (2006) test with 5% nominal level

two types of DGPs. For the first DGP (DGP1), we assume

$$Z_t^\intercal = W_t^\intercal \Pi_0 + W_{1,t} u_t^\intercal \text{ with } u_t = v_t - \frac{1}{4}\mathbf{1}_4\mathbf{1}_4^\intercal v_{t-1}, t = 1, \ldots, T,$$

where $v_t \overset{i.i.d.}{\sim} N(0, I_4)$, $W_t \overset{i.i.d.}{\sim} N(0, I_4)$ and $W_{1,t}$ is the first element of W_t. Note the errors now are heteroskedastic and autocorrelated. Let

$$\Pi_0 = \text{diag}(\mathbf{1}_2, \mathbf{0}_2) + \rho I_4 \text{ for } \rho \geq 0 \ .$$

For the second DGP (DGP2), following Kleibergen and Paap (2006) we assume

$$R_t = \Pi_0 F_t + \varepsilon_t \text{ with } \varepsilon_t = v_t + \Gamma v_{t-1}, t = 1, \ldots, T,$$

where $v_t \overset{i.i.d.}{\sim} N(0, \Sigma_v)$ and $F_t \overset{i.i.d.}{\sim} N(0, \Sigma_F)$ with $\Gamma \in \mathbf{M}^{10\times 10}$, $\Sigma_v \in \mathbf{M}^{10\times 10}$ and $\Sigma_F \in \mathbf{M}^{4\times 4}$ given in Appendix 2.7.4. Let

$$\Pi_0 = \beta\alpha^\intercal + \rho\Pi_1 \text{ for } \rho \geq 0 \ ,$$

where $\alpha \in \mathbf{R}^4$, $\beta \in \mathbf{R}^{10}$ and $\Pi_1 \in \mathbf{M}^{10\times 4}$ are given in Appendix 2.7.4. These values are estimates based on the real data used in Section 2.4.2. In view of this, we use DGP2 to mimic possible scenarios in Section 2.4.2 as in Kleibergen and Paap (2006).

We examine the hypotheses (2.1) with $r = 2$ and $r = 3$ for DGP1, and the hypotheses (2.1) with $r = 3$ for DGP2. The design of Π_0 implies that H_0 is true if and only if $\delta = 0$ for both cases. In particular, for DGP1 $r_0 = 2$ under H_0, and for DGP2 $r_0 = 1$ under H_0. So $r = 3$ for both DGPs represents the case when Π_0 is close to a matrix with rank strictly less than r, while $r = 2$ for DGP1 represents the regular case. Given the findings in Figure 2.4, for the hypotheses with $r = 3$ for both DGPs, it shall be expected that our tests are more powerful than the Kleibergen and Paap (2006) test when directly applied and the multiple testing method.

To implement all tests, we estimate Π_0 by $\hat{\Pi}_T = \frac{1}{T}\sum_{t=1}^{T} W_t Z_t^\intercal$ for DGP1 and by $\hat{\Pi}_T = \sum_{t=1}^{T} R_t F_t^\intercal \, (\sum_{t=1}^{T} F_t F_t^\intercal)^{-1}$ for DGP2. It is clear that the asymptotic distribution $\mathcal{M}$ of $\hat{\Pi}_T$ is zero mean Gaussian with convergence rate $\sqrt{T}$, so Assumption 2.3.1 is satisfied. As the data exhibits first order autocorrelation, we adopt the simple block bootstrap (Lahiri, 2003) to resample the data with block size $b = 2$ for implementing our tests. For derivative estimation in (2.38) and (2.39), we set the tuning parameter $\kappa_T = T^{-1/4}$ and $T^{-1/3}$. It is also clear that all assumptions in Kleibergen and Paap (2006) are satisfied. We use HACC matrix estimator with one lag (West, 1997) for the long run covariance matrix estimator. See Appendix 2.7.3 for a review on the Kleibergen and Paap (2006) test.

We let $\rho = 0, 0.1, \cdots, 0.5$ for DGP1 and $\rho = 0, 0.01, \cdots, 0.1$ for DGP2, where ρ represents how strong H_1 deviates away from H_0. We consider $T = 50, 100, 300, 1000$ for DGP1 and $T = 330$ for DGP2. The rejection rates, which are based on $5,000$ simulation replications with 500 bootstrap replications, are reported in Tables 2.1-2.3. We use CF1 and CF2 to denote our tests using derivative estimator in (2.38) and (2.39), respectively, and KP-D and KP-M to denote the Kleibergen and Paap (2006) test when directly applied and the multiple testing method, respectively. The nominal level is 5% throughout.

The main findings are summarized as follows. First, CF1 exhibits good finite sample

performance for all cases, even when $T = 50$. Interestingly, as Tables 2.1 and 2.3 show, the rejection rates of CF1 for $r = 2$ under DGP1 and $r = 3$ under DGP2 are invariant to the choice of κ_T in most of cases. The rejection rates of CF1 for $r = 3$ under DGP1 are not quite sensitive to the choice of κ_T. Second, the performance of CF2 is not as satisfactory as that of CF1 in small samples. In particular, CF2 is over rejected for all cases with $\rho = 0$ when $T = 50$ or 100. This indicates that good performance of CF2 may require a larger T than CF1 does. This may be explained by the fact that the structural method (CF1) exploits more information of the derivative. For large T, CF2 seems to be more powerful than CF1 under DGP1 when $T = 300$ or $1,000$, while CF1 seems to be more powerful than CF2 under DGP2. We leave a thorough comparison between these two methods of derivative estimation for future study. Third, the performance of KP-M and KP-D is less satisfactory than our tests. As Table 2.1 shows, KP-M and KP-D over-reject the null for $r = 2$ under DGP1 with $\rho = 0$ when $T = 50$ or 100. This indicates that good performance of KP-M and KP-D may require a large T. On other other hand, as Tables 2.2 and 2.3 show, KP-M and KP-D under-reject the null for $r = 3$ under DGP1 and DGP2 with $\rho = 0$. This is consistent with the finding in Figure 2.1. Moreover, as expected, CF1 and CF2 are uniformly more powerful than KP-M and KP-D as shown in Tables 2.2 and 2.3.[7] In addition, in our designed simulation, the rejection rates of KP-M and KP-D are similar with insignificant difference, although the latter is slightly more powerful.

2.4.2 Testing for Identification in SDF Models

Following Jagannathan and Wang (1996), the stochastic discount factor (SDF) model based on the conditional capital asset pricing model is specified as

$$E[R_{t+1}F_{t+1}^{\intercal}\gamma_0|\mathcal{I}_t] = \mathbf{1}_m \ , \tag{2.41}$$

[7] In Table 2.2, the rejection rates of KP-M and KP-D under the alternatives are size adjusted ones.

Table 2.1: Rejection rates for $r = 2$ under DGP1

	$\rho = 0$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.0380	0.0374	0.2764	0.2078	0.1276	0.1658
$T = 100$	0.0402	0.0402	0.1958	0.1232	0.0930	0.0952
$T = 300$	0.0450	0.0450	0.1218	0.0512	0.0606	0.0606
$T = 1000$	0.0472	0.0472	0.0752	0.0368	0.0526	0.0526
	$\rho = 0.1$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.0812	0.0812	0.3520	0.2676	0.0720	0.0828
$T = 100$	0.1210	0.1210	0.3356	0.2314	0.1262	0.1316
$T = 300$	0.3458	0.3458	0.5144	0.3600	0.3961	0.3962
$T = 1000$	0.8976	0.8976	0.9238	0.8784	0.9784	0.9152
	$\rho = 0.2$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.2248	0.2248	0.5714	0.4880	0.1904	0.2078
$T = 100$	0.4254	0.4254	0.6950	0.5750	0.4410	0.4520
$T = 300$	0.9350	0.9350	0.9694	0.9366	0.9526	0.9526
$T = 1000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	$\rho = 0.3$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.4776	0.4776	0.7852	0.7208	0.3682	0.4142
$T = 100$	0.8044	0.8044	0.9348	0.8906	0.7964	0.8102
$T = 300$	0.9992	0.9992	0.9980	1.0000	1.0000	1.0000
$T = 1000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	$\rho = 0.4$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.7220	0.7220	0.9236	0.8896	0.5380	0.6212
$T = 100$	0.9618	0.9618	0.9954	0.9832	0.9456	0.9586
$T = 300$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$T = 1000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	$\rho = 0.5$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.8872	0.8872	0.9786	0.9658	0.6696	0.7846
$T = 100$	0.9960	0.9960	0.9994	0.9992	0.9840	0.9946
$T = 300$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$T = 1000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 2.2: Rejection rates for $r = 3$ under DGP1

	$\rho = 0$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.0594	0.0388	0.1410	0.1248	0.0072	0.0156
$T = 100$	0.0556	0.0328	0.1156	0.0944	0.0066	0.0110
$T = 300$	0.0486	0.0324	0.0766	0.0564	0.0050	0.0062
$T = 1000$	0.0550	0.0484	0.0656	0.0544	0.0044	0.0056
	$\rho = 0.1$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.0874	0.0534	0.1770	0.1600	0.0168	0.0270
$T = 100$	0.1114	0.0626	0.1936	0.1624	0.0270	0.0344
$T = 300$	0.2926	0.1562	0.3628	0.3068	0.0926	0.0994
$T = 1000$	0.8070	0.5948	0.8226	0.7730	0.5396	0.5428
	$\rho = 0.2$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.1804	0.1182	0.3334	0.3030	0.0698	0.0910
$T = 100$	0.3162	0.2060	0.4882	0.4342	0.1692	0.1806
$T = 300$	0.7774	0.6724	0.8872	0.8426	0.6644	0.6666
$T = 1000$	0.9988	0.9986	0.9960	0.9994	0.9982	0.9982
	$\rho = 0.3$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.3254	0.2538	0.5566	0.5166	0.1906	0.2432
$T = 100$	0.5678	0.4986	0.7886	0.7414	0.4856	0.4962
$T = 300$	0.9602	0.9576	0.9940	0.9874	0.9602	0.9602
$T = 1000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	$\rho = 0.4$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.4916	0.4460	0.7488	0.7110	0.3544	0.4434
$T = 100$	0.7758	0.7626	0.9422	0.9120	0.7552	0.7656
$T = 300$	0.9972	0.9972	0.9998	0.9996	0.9974	0.9974
$T = 1000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	$\rho = 0.5$					
	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$T = 50$	0.6432	0.6288	0.8798	0.8442	0.5182	0.6290
$T = 100$	0.9146	0.9138	0.9880	0.9766	0.9016	0.9116
$T = 300$	0.9998	0.9998	1.0000	1.0000	0.9998	0.9998
$T = 1000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 2.3: Rejection rates for $r = 3$ under DGP2 when $T = 330$

	CF1		CF2		KP	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	KP-M	KP-D
$\rho = 0.00$	0.0514	0.0514	0.0468	0.0406	0.0006	0.0008
$\rho = 0.01$	0.2834	0.2834	0.1770	0.1104	0.0460	0.0482
$\rho = 0.02$	0.4228	0.4228	0.1648	0.0864	0.0956	0.1018
$\rho = 0.03$	0.5850	0.5850	0.2192	0.1242	0.2044	0.2166
$\rho = 0.04$	0.7526	0.7526	0.3268	0.2388	0.3562	0.3768
$\rho = 0.05$	0.8706	0.8706	0.4944	0.4010	0.5314	0.5598
$\rho = 0.06$	0.9500	0.9500	0.6622	0.5796	0.6898	0.7294
$\rho = 0.07$	0.9822	0.9606	0.8064	0.7388	0.7994	0.8464
$\rho = 0.08$	0.9932	0.9852	0.9032	0.8628	0.8748	0.9276
$\rho = 0.09$	0.9982	0.9936	0.9582	0.9368	0.9144	0.9670
$\rho = 0.10$	0.9998	0.9984	0.9842	0.9754	0.9306	0.9852

where $R_t \in \mathbf{R}^m$ is a vector of returns on m assets at time t, $F_t \in \mathbf{R}^k$ is a vector of common factors at time t, $\mathcal{I}_t$ is the information set at time t, and $\gamma_0 \in \mathbf{R}^k$ is a vector of risk premia. The risk premia γ_0 can be estimated by the generalized method of moments (Hansen, 1982), see, for example, Jagannathan et al. (2002). The GMM estimator of γ_0 is consistent if

$$E[R_{t+1}F_{t+1}^{\intercal}|\mathcal{I}_t] \tag{2.42}$$

is of full rank at time t, see, for example, Hansen (1982) and Newey and McFadden (1994). Therefore, it is of importance to test for the full rank of (2.42) to indicate whether γ_0 is identifiable.

When the conditional expectation of $R_{t+1}F_{t+1}^{\intercal}$ does not depend on $\mathcal{I}_t$ and R_t satisfies a linear factor model

$$R_t = \Pi_0 F_t + \varepsilon_t \tag{2.43}$$

with $E[F_t\varepsilon_t] = 0$ and $E[F_tF_t^{\intercal}]$ being nonsingular, then testing for the full rank of (2.42) is equivalent to testing for the full rank of Π_0. Following Kleibergen and Paap (2006), instead

of testing for the full rank of (2.42), we opt to test whether Π_0 is of full rank. Thus, this amounts to examining the hypotheses (2.1) with $r = k - 1$. We cannot restrict ourself to examine the hypotheses (2.2) since it is unrealistic to assume $r_0 \geq k - 1$ unless $k = 1$.

We use the same set of data as in Kleibergen and Paap (2006). There are returns R_t on 10 portfolios and 4 factors in F_t with observations from July 1963 to December 1990, so $m = 10$, $k = 4$ and $T = 330$. The factors in F_t consist of constant, the return on a value-weighted portfolio, a corporate bond yield spread and a measure of per capita labor income growth. We estimate Π_0 by

$$\hat{\Pi}_T = \sum_{t=1}^{T} R_t F_t^{\intercal} (\sum_{t=1}^{T} F_t F_t^{\intercal})^{-1} \ . \tag{2.44}$$

As demonstrated in Kleibergen and Paap (2006), the data on returns R_t exhibits first order autocorrelation. To compute the test statistics of Kleibergen and Paap (2006) test, we use HACC matrix estimator with one lag (West, 1997) for the long run covariance matrix estimator. To implement our tests, we adopt the simple block bootstrap (Lahiri, 2003) to resample the data with block size $b = 1, 2, 3, 4$. For derivative estimation in (2.39) and (2.38), we set the tuning parameter $\kappa_T = T^{-1/4}$ and $T^{-1/3}$.

The results, which are based on $1,000$ bootstrap replications, are reported in Table 2.4. We use CF1 and CF2 to denote our tests using derivative estimator in (2.38) and (2.39), respectively, and KP-D and KP-M to denote the Kleibergen and Paap (2006) test when directly applied and the multiple testing method based on it, respectively. As Panel A of Table 2.4 indicates, all our tests fail to reject the non-full rank of Π_0 with 5% nominal level, which is consistent with the finding in Kleibergen and Paap (2006). However, the p values of our tests are uniformly smaller than 15% with some smaller than 10%, while the p values of the two conventional tests are larger than 90%. This implies that our tests reject the non-full rank of Π_0 in some cases at the 10% level, while the conventional tests never reject the non-full rank of Π_0 at any conventional significance level. In this sense, the evidence for non-identification of γ_0 from our tests is very weak, while the evidence from

the conventional tests is very strong. Given the drawback of the conventional tests, the conclusion from our tests is more reliable.

Table 2.4: p values for different tests

Panel A: our tests				
	CF1		CF2	
	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$	$\kappa_T = T^{-1/4}$	$\kappa_T = T^{-1/3}$
$b = 1$	0.079	0.079	0.118	0.121
$b = 2$	0.094	0.094	0.113	0.119
$b = 3$	0.103	0.103	0.128	0.140
$b = 4$	0.082	0.082	0.137	0.138
Panel B: conventional tests				
	KP-M		KP-D	
	0.9063		0.9063	

The p value for KP-M is given by the smallest significance level such that the null hypothesis is rejected, which is equal to the maximum p value of all Kleibergen and Paap (2006)'s tests implemented by the multiple testing method.

2.4.3 Rank Determination

Testing for the hypotheses (2.1) only tells whether r_0 satisfies the inequality or not. In many cases, however, we still want to know what r_0 is. In addition to employing the multiple testing method to test for inequality of cointegration rank, Johansen (1995, Chapter 12) also proposed a sequential testing procedure to determine the rank of cointegration in VAR models (see, for instance, Example 2.2.2). More examples that concern the true rank of a matrix can be found in Examples 2.2.5-2.2.7. In this section, we demonstrate how our tests can improve the accuracy of the sequential testing procedure for rank determination.

We first characterize the sequential testing procedure for rank determination in our general framework following Johansen (1995, Chapter 12). For $\alpha \in (0, 1)$, let $\psi_n^{(r)}$ be a test for the hypotheses (2.1) or (2.2) such that $\lim_{n\to\infty} P(\psi_n^{(r)} = 1) = \alpha$ when $r_0 = r$, and $\lim_{n\to\infty} P(\psi_n^{(r)} = 1) = 1$ when $r_0 > r$. For example, it can be any one of existing rank tests or our tests. The sequential testing procedure starts with $q = 0$ and carries out $\psi_n^{(q)}$ with progressively larger q. The rank estimator $\hat{r}_n^*$ is defined as the threshold value q^* when $\psi_n^{(q^*)}$

does not reject the null hypothesis for the first time, and $\hat{r}_n^* = k$ if such q^* does not exist. Formally, $\hat{r}_n^* = k$ if $\psi_n^{(q)} = 1$ for all $0 \leq q \leq k-1$ and otherwise

$$\hat{r}_n^* = \min\{0 \leq q \leq k-1 : \psi_n^{(q)} = 0\} \ . \tag{2.45}$$

Remark 2.4.1. Clearly, $\hat{r}_n^* > r$ is equivalent to $\psi_n^{(q)} = 1$ for all $0 \leq q \leq r$. Thus, for given existing rank tests $\{\psi_n^{(q)}\}_{q=1}^r$, rejecting H_0 by the multiple testing method based on $\{\psi_n^{(q)}\}_{q=1}^r$ is equivalent to $\hat{r}_n^* > r$ where $\hat{r}_n^*$ is based on $\{\psi_n^{(q)}\}_{q=1}^r$. In fact, Kleibergen and Paap (2006) relied on this relation for $r = k-1$ to test for identification of the risk premia parameters in stochastic discount factor models. ■

The following theorem establishes that $\hat{r}_n^*$ is a good estimator for r_0.

Theorem 2.4.1. *For $\alpha \in (0,1)$, let $\psi_n^{(r)}$ be a test for the hypotheses* (2.1) *or* (2.2) *such that $\lim_{n\to\infty} P(\psi_n^{(r)} = 1) = \alpha$ when $r_0 = r$, and $\lim_{n\to\infty} P(\psi_n^{(r)} = 1) = 1$ when $r_0 > r$. Then $\lim_{n\to\infty} P(\hat{r}_n^* < r_0) = 0$,*

$$\lim_{n\to\infty} P(\hat{r}_n^* = r_0) = 1-\alpha \text{ if } r_0 < k \text{ and } 1 \text{ if } r_0 = k \ ,$$

and

$$\lim_{n\to\infty} P(\hat{r}_n^* > r_0) = \alpha \text{ if } r_0 < k \text{ and } 0 \text{ if } r_0 = k \ .$$

Theorem 2.4.1 implies that the true rank is correctly chosen with probability no smaller than $1-\alpha$ asymptotically, a smaller rank is chosen with probability going to zero, and a larger rank is chosen with probability no larger than α asymptotically. In short, $\{\hat{r}_n^*\}$ provides a confidence set for r_0 with asymptotic coverage probability no smaller than $1-\alpha$. Interestingly, Theorem 2.4.1 does not rely on the behavior of $\psi_n^{(q)}$ when $q > r_0$, since the sequential testing procedure carries out $\psi_n^{(q)}$ progressively from $q = 0$ to larger q and terminates before $q = r_0$ with probability no smaller than $1-\alpha$ asymptotically. That is, efficient rank determination does not require the ability of detecting whether $\mathrm{rank}(\Pi_0)$ is strictly less than a hypothesized value. This explains why the hypotheses (2.2) has become

prevalent.

However, the procedure crucially depends on the behavior of $\psi_n^{(q)}$ when $q < r_0$, that is, the power of detecting whether rank(Π_0) is strictly greater than hypothesized values. In particular, the probability of ensuring a no smaller rank crucially depends on the probability of accepting $r_0 > q$ for $q = 0, \ldots, r_0 - 1$, which is the power of $\psi_n^{(q)}$ for $q = 0, \ldots, r_0 - 1$. This suggests that our tests may be leveraged for accuracy improvement in the sequential testing procedure for rank determination, provided the improved power property of our tests over existing rank tests as shown in Sections 2.2.2 and 2.4.1.

To show how the sequential testing procedure based on our tests can be more accurate than that based on existing rank tests, we focus on the case of the Kleibergen and Paap (2006) test and present some simulation evidence. We use the same DGP given in (2.21) and (2.22) with $\delta = 0.1$ and 0.12. The design of Π_0 implies that $r_0 = 6$ for both δ's and all $d = 1, \ldots, 6$. The Kleibergen and Paap (2006) test and our tests are implemented as in Section 2.2.1 and 2.4.1. The probability distributions of $\hat{r}_n^*$, which are based on $5,000$ simulation replications are reported in Figures 2.5 and 2.6. We use CF to denote the sequential testing procedure based on our tests and KP to denote the one based on the Kleibergen and Paap (2006) test. The result is based on $\kappa_n = n^{-1/4}$ and the derivative estimator in (2.38). The result for $\kappa_n = n^{-1/3}$ is similar and is available upon request. As shown in both figures, CF yields more accurate rank estimators than KP uniformly over $d = 1, \ldots, 6$ for both δ's. In particular, KP tends to underestimate the true rank when d increases. The coverage probability of the resulting rank estimator is 5.46% when $d = 6$ and $\delta = 0.1$, and 25.3% when $d = 6$ and $\delta = 0.12$. The coverage probabilities of CF's rank estimator are greater than those of KP's rank estimator.

Remark 2.4.2. To obtain a consistent estimator for r_0, Cragg and Donald (1997) and Robin and Smith (2000) make an adjustment dependent on n to the nominal level α. The consistency of $\hat{r}_n^*$ can be obtained when the adjusted nominal level $\alpha_n \to 0$ as $n \to \infty$ and satisfies certain rate requirement. In fact, the estimator $\hat{r}_n$ used in (2.38) provides a simple

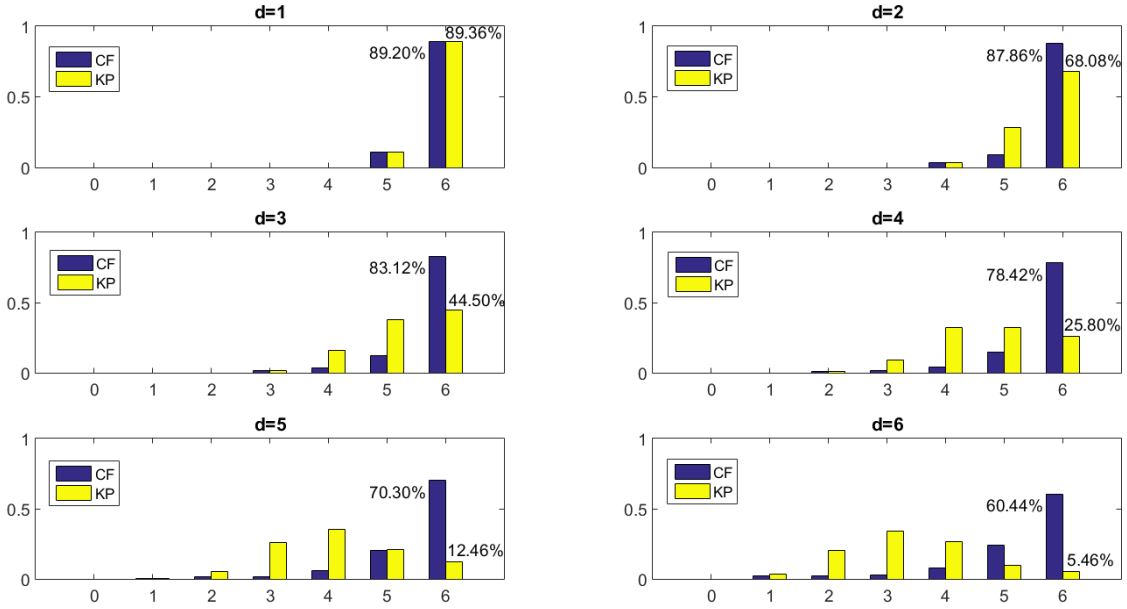


Figure 2.5: Comparison between the sequential testing procedures based on our tests and the Kleibergen and Paap (2006) test with $\delta = 0.1$

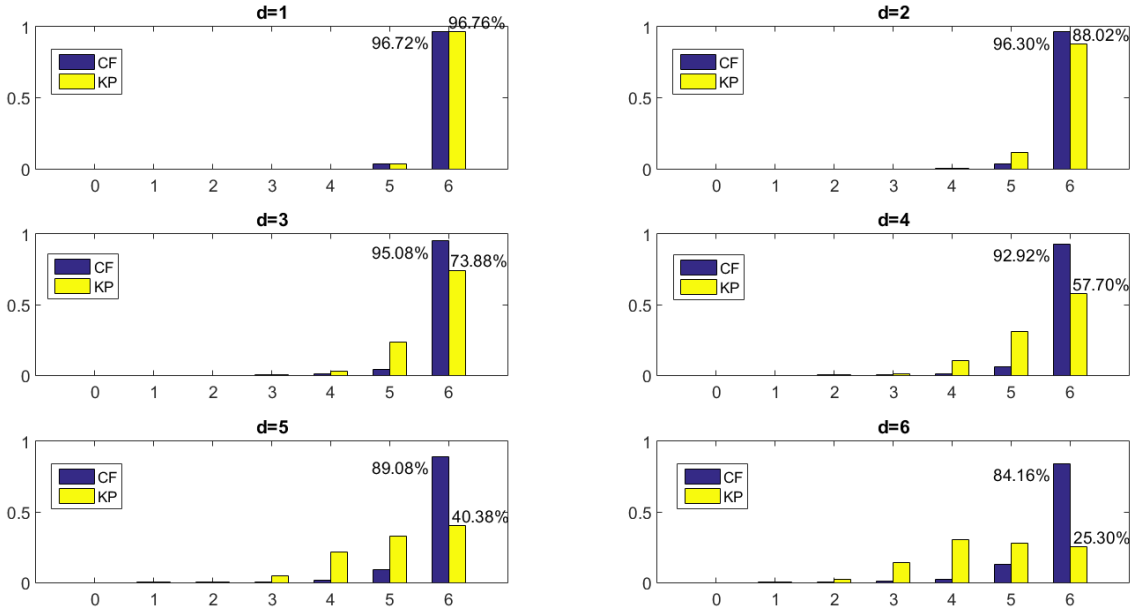


Figure 2.6: Comparison between the sequential testing procedures based on our tests and the Kleibergen and Paap (2006) test with $\delta = 0.12$

and consistent estimator for r_0, see Lemma 2.7.6. ■

2.5 Conclusion

In this paper, we developed a more powerful method for examining a "no greater than" inequality of the rank of a matrix and a more accurate procedure for rank determination in a general setup. We proved that our tests have the asymptotic null rejection rate that is exactly equal to the nominal level regardless of whether the rank is less than or equal to the hypothesized value. Our simulation results showed that our tests are often more powerful than the multiple testing method, and improve the accuracy of the sequential testing procedure for rank determination. We illustrated our methods in several examples, including testing for identification and testing for the existence of stochastic trend and/or cointegration, to show the wide applicability of our methods.

2.6 Acknowledgement

Chapter 2 "Improved Inference on the Rank of a Matrix with Applications to IV and Cointegration Models," in part, is currently being prepared for submission for publication of the material. Chen, Qihui; Fang, Zheng. The dissertation author was the primary investigator and author of this material.

2.7 Appendix

2.7.1 Proofs of Main Results

Proof of Lemma 2.3.1: The proof is based on a simple application of the representation of extremal partial trace. Recall that $\sigma_1^2(\Pi), \ldots, \sigma_k^2(\Pi)$ are eigenvalues of $\Pi^\intercal\Pi$ in descending

order. Let $d \equiv k - r$. It follows by Proposition 1.3.4 in Tao (2012) that

$$\phi(\Pi) = \sum_{j=r+1}^{k} \sigma_j^2(\Pi) = \inf_{u_1,\ldots,u_d} \sum_{j=1}^{d} u_j^\intercal \Pi^\intercal \Pi u_j \ , \tag{2.46}$$

where the infimum is taken over all $u_1, \ldots, u_d \in \mathbf{R}^k$ that are orthonormal. Let $U \equiv [u_1, \ldots, u_d]$. Clearly, $U \in \mathbb{S}^{k\times d}$. By (2.46) and the definition of Frobenius norm, we further have

$$\phi(\Pi) = \inf_{U\in\mathbb{S}^{k\times d}} \mathrm{tr}(U^\intercal \Pi^\intercal \Pi U) = \inf_{U\in\mathbb{S}^{k\times d}} \|\Pi U\|^2 \ . \tag{2.47}$$

The infimum in (2.47) is in fact achieved on $\mathbb{S}^{k\times d}$ because $U \mapsto \|\Pi U\|^2$ is clearly continuous, and $\mathbb{S}^{k\times d}$ is compact since it is closed and bounded. This completes the proof of the lemma.

■

Proof of Proposition 2.3.1: Recall that $d = k - r$. Define $\phi_1 : \mathbf{M}^{m\times k} \to C(\mathbb{S}^{k\times d})$ by $\phi_1(\Pi)(U) = \|\Pi U\|^2$, and $\phi_2 : C(\mathbb{S}^{k\times d}) \to \mathbf{R}$ by $\phi_2(f) = \min\{f(U) : U \in \mathbb{S}^{k\times d}\}$, thus $\phi = \phi_2 \circ \phi_1$ by Lemma 2.3.1. For part (i), we proceed by verifying first order Hadamard directional differentiability of ϕ_1 and ϕ_2, and then conclude by the chain rule.

Let $\{M_n\} \subset \mathbf{M}^{m\times k}$ be such that $M_n \to M \in \mathbf{M}^{m\times k}$ and $t_n \downarrow 0$ as $n \to \infty$. For each $n \in \mathbf{N}$, define $g_n : \mathbb{S}^{k\times d} \to \mathbf{R}$ by

$$g_n(U) = \frac{\|(\Pi + t_n M_n)U\|^2 - \|\Pi U\|^2}{t_n} = \frac{\|\Pi U + t_n M_n U\|^2 - \|\Pi U\|^2}{t_n} \ ,$$

and $g : \mathbb{S}^{k\times d} \to \mathbf{R}$ by $g(U) = 2\mathrm{tr}((\Pi U)^\intercal M U)$. Then by simple algebra we have

$$\begin{aligned} \sup_{U\in\mathbb{S}^{k\times d}} |g_n(U) - g(U)| &= \sup_{U\in\mathbb{S}^{k\times d}} |2\mathrm{tr}((\Pi U)^\intercal (M_n - M)U) + t_n \|M_n U\|^2| \\ &\leq \sup_{U\in\mathbb{S}^{k\times d}} \{2\|\Pi U\|\|(M_n - M)U\| + t_n \|M_n U\|^2\} \ , \end{aligned} \tag{2.48}$$

where the inequality follows by the triangle inequality and the Cauchy-Schwarz inequality

for the trace operator. For the right hand side of (2.48), we have

$$\begin{aligned} &\sup_{U\in\mathbb{S}^{k\times d}}\{2\|\Pi U\|\|(M_n-M)U\|+t_n\|M_nU\|^2\} \\ \leq &\sup_{U\in\mathbb{S}^{k\times d}}\{2\|\Pi\|\|U\|\|M_n-M\|\|U\|+t_n\|M_n\|^2\|U\|^2\}=o(1)\ , \end{aligned} \tag{2.49}$$

where we exploited the sub-multiplicativity of Frobenius norm and the fact that $\|U\|=\sqrt{d}$ and that $M_n\to M$ as well as $t_n\downarrow 0$ as $n\to\infty$. We thus conclude from (2.48) and (2.49) that $g_n\to g$ uniformly in $C(\mathbb{S}^{k\times d})$, or equivalently ϕ_1 is first order Hadamard directionally differentiable at Π with derivative $\phi'_{1,\Pi}:\mathbf{M}^{m\times k}\to C(\mathbb{S}^{k\times d})$ given by

$$\phi'_{1,\Pi}(M)(U)=2\mathrm{tr}((\Pi U)^\intercal MU)\ . \tag{2.50}$$

On the other hand, Theorem 3.1 in Shapiro (1991) implies that $\phi_2:C(\mathbb{S}^{k\times d})\to\mathbf{R}$ is first order Hadamard directionally differentiable at any $f\in C(\mathbb{S}^{k\times d})$ with derivative $\phi'_{2,f}:C(\mathbb{S}^{k\times d})\to\mathbf{R}$ given by

$$\phi'_{2,f}(h)=\min_{U\in\Psi(f)}h(U)\ , \tag{2.51}$$

where, by abuse of notation, $\Psi(f)\equiv\arg\min_{U\in\mathbb{S}^{k\times d}}f(U)$. Combining (2.50), (2.51) and the chain rule (Shapiro, 1990, Proposition 3.6), we may now conclude that $\phi:\mathbf{M}^{m\times k}\to\mathbf{R}$ is first order Hadamard directionally differentiable at any $\Pi\in\mathbf{M}^{m\times k}$ with the derivative $\phi'_\Pi:\mathbf{M}^{m\times k}\to\mathbf{R}$ given by

$$\phi'_\Pi(M)=\phi'_{2,\phi_1(\Pi)}\circ\phi'_{1,\Pi}(M)=\min_{U\in\Psi(\Pi)}2\mathrm{tr}((\Pi U)^\intercal MU)\ .$$

This completes the proof of part (i) of the proposition.

For part (ii), note that $\phi(\Pi)=0$ implies that $\Pi U=0$ for all $U\in\Psi(\Pi)$ and hence $\phi'_\Pi(M)=0$ for all $M\in\mathbf{M}^{m\times k}$. Recall that $\{M_n\}\subset\mathbf{M}^{m\times k}$ with $M_n\to M\in\mathbf{M}^{m\times k}$ and

$t_n \downarrow 0$ as $n \to \infty$. By Lemma 2.3.1 we have

$$|\phi(\Pi + t_n M_n) - \phi(\Pi + t_n M)| \leq |\min_{U \in \mathbb{S}^{k\times d}} \|(\Pi + t_n M_n)U\| - \min_{U \in \mathbb{S}^{k\times d}} \|(\Pi + t_n M)U\|| \times (\min_{U \in \mathbb{S}^{k\times d}} \|(\Pi + t_n M_n)U\| + \min_{U \in \mathbb{S}^{k\times d}} \|(\Pi + t_n M)U\|) \ , \tag{2.52}$$

where the inequality follows by the formula $a^2 - b^2 = (a + b)(a - b)$. For the first term on the right hand side of (2.52), we have

$$|\min_{U \in \mathbb{S}^{k\times d}} \|(\Pi + t_n M_n)U\| - \min_{U \in \mathbb{S}^{k\times d}} \|(\Pi + t_n M)U\|| \leq t_n \sqrt{d}\|M_n - M\| = o(t_n) \ , \tag{2.53}$$

where the inequality follows by the Lipschitz continuity of the infimum operator, the triangle inequality, the sub-multiplicativity of Frobenius norm and $\|U\| = \sqrt{d}$ for $U \in \mathbb{S}^{k\times d}$. For the second term on the right hand side of (2.52), we have

$$\min_{U \in \mathbb{S}^{k\times d}} \|(\Pi + t_n M_n)U\| + \min_{U \in \mathbb{S}^{k\times d}} \|(\Pi + t_n M)U\| \leq \|(\Pi + t_n M_n)U^*\| + \|(\Pi + t_n M)U^*\| \leq t_n\|M_n\|\|U^*\| + t_n\|M\|\|U^*\| = O(t_n) \ , \tag{2.54}$$

where the first inequality follows by letting U^* be an element from $\Psi(\Pi)$, and the second inequality follows by $\Pi U^* = 0$, the sub-multiplicativity of Frobenius norm and the fact that $\|U^*\| = \sqrt{d}$ and that $M_n \to M$ as $n \to \infty$. Combining (2.52)-(2.54), we thus obtain

$$|\phi(\Pi + t_n M_n) - \phi(\Pi + t_n M)| = o(t_n^2) \ . \tag{2.55}$$

Next, for $\epsilon > 0$, let $\Psi(\Pi)^\epsilon \equiv \{U \in \mathbb{S}^{k\times d} : \min_{U' \in \Psi(\Pi)} \|U' - U\| \leq \epsilon\}$ and $\Psi(\Pi)_1^\epsilon \equiv \{U \in \mathbb{S}^{k\times d} : \min_{U' \in \Psi(\Pi)} \|U' - U\| \geq \epsilon\}$. In what follows we consider the nontrivial case $\Pi \neq 0$ and $M \neq 0$. In this case, $\Psi(\Pi) \subsetneqq \mathbb{S}^{k\times d}$ in view of Proposition 1.3.4 in Tao (2012) and hence $\Psi(\Pi)_1^\epsilon \neq \emptyset$ for ϵ sufficiently small. Let $\sigma_{\min}^+(\Pi)$ denote the smallest positive singular value of Π which exists since $\Pi \neq 0$, and $\Delta \equiv 3\sqrt{2}[\sigma_{\min}^+(\Pi)]^{-1} \max_{U \in \mathbb{S}^{k\times d}} \|MU\| > 0$ since

$M \neq 0$. Then it follows that for all n sufficiently large

$$\begin{aligned}\min_{U\in\Psi(\Pi)_1^{t_n\Delta}} \|(\Pi + t_n M)U\| &\geq \min_{U\in\Psi(\Pi)_1^{t_n\Delta}} \|\Pi U\| - t_n \max_{U\in\mathbb{S}^{k\times d}} \|MU\| \\ &\geq \frac{\sqrt{2}}{2} t_n \sigma^+_{\min}(\Pi)\Delta - t_n \max_{U\in\mathbb{S}^{k\times d}} \|MU\| > t_n \max_{U\in\mathbb{S}^{k\times d}} \|MU\| \\ &\geq \min_{U\in\Psi(\Pi)} \|(\Pi + t_n M)U\| \geq \sqrt{\phi(\Pi + t_n M)}\ , \end{aligned} \tag{2.56}$$

where the first inequality follows by the Lipschitz continuity of the infimum operator, the triangle inequality and the fact that $\Psi(\Pi)_1^{t_n\Delta} \subset S^{k\times d}$, the second inequality follows by Lemma 2.7.1, the third inequality follows by the definition of Δ, and the fourth inequality holds by the fact that $\Pi U = 0$ for $U \in \Psi(\Pi)$. By (2.56), we thus obtain that for all n sufficiently large

$$\phi(\Pi + t_n M) = \min_{U\in\Psi(\Pi)^{t_n\Delta}} \|(\Pi + t_n M)U\|^2\ . \tag{2.57}$$

Now, for fixed $U \in \Psi(\Pi)$, $\Delta > 0$ and $t \in \mathbf{R}$, let $\Gamma^\Delta \equiv \{V \in \mathbf{M}^{k\times d} : \|V\| \leq \Delta\}$ and $\Gamma_U^\Delta(t) \equiv \{V \in \Gamma^\Delta : U + tV \in \mathbb{S}^{k\times d}\} = \{V \in \Gamma^\Delta : V^\intercal U + U^\intercal V = -tV^\intercal V\}$. Define a correspondence $\varphi : \mathbf{R} \twoheadrightarrow \mathbb{S}^{k\times d} \times \Gamma^\Delta$ by $\varphi(t) = \{(U, V) : U \in \Psi(\Pi), V \in \Gamma_U^\Delta(t)\}$. Then the right hand side of (2.57) can be written as

$$\begin{aligned}\min_{U\in\Psi(\Pi)^{t_n\Delta}} \|(\Pi + t_n M)U\|^2 &= \min_{(U,V)\in\varphi(t_n)} \|(\Pi + t_n M)(U + t_n V)\|^2 \\ &= t_n^2 \min_{(U,V)\in\varphi(t_n)} \|\Pi V + MU\|^2 + o(t_n^2)\ , \end{aligned} \tag{2.58}$$

where the second equality follows by the fact that $\Pi U = 0$ for all $U \in \Psi(\Pi)$ and $\|MV\| \leq \|M\|\Delta$ for all $V \in \Gamma^\Delta$. By Lemma 2.7.2, $\varphi(t)$ is continuous at $t = 0$. Moreover, φ is obviously compact-valued. We may then obtain by Theorem 17.31 in Aliprantis and Border

(2006) that

$$\min_{(U,V)\in\varphi(t_n)} \|\Pi V + MU\|^2 = \min_{(U,V)\in\varphi(0)} \|\Pi V + MU\|^2 + o(1)$$
$$= \min_{U\in\Psi(\Pi)} \min_{V\in\mathbf{M}^{k\times d}} \|\Pi V + MU\|^2 + o(1) \ , \tag{2.59}$$

where the second equality holds by letting Δ sufficiently large in view of Lemma 2.7.3. Combining (2.55), (2.57), (2.58) and (2.59) then yields part (ii) of the proposition. ■

Proof of Proposition 2.3.2 Recall that $d = k - r$ and let $d^* \equiv k - r^*$. Noting that the column vectors in Q_2 form a orthonormal basis for the null space of Π_0, we may rewrite $\Psi(\Pi)$ as $\Psi(\Pi) = \{Q_2 V : V \in \mathbb{S}^{d^*\times d}\}$. This together with the projection theorem implies

$$\phi''_{\Pi}(M) = \min_{V\in\mathbb{S}^{d^*\times d}} \|(I - \Pi(\Pi^\intercal\Pi)^-\Pi^\intercal)MQ_2V\|^2 \ , \tag{2.60}$$

where A^- denotes the Moore-Penrose inverse of a generic matrix A. By the singular value decomposition of Π, we have

$$(I - \Pi(\Pi^\intercal\Pi)^-\Pi^\intercal)P = P - P\Sigma Q^\intercal(Q\Sigma^\intercal P^\intercal P\Sigma Q^\intercal)^- Q\Sigma^\intercal P^\intercal P$$
$$= P - P\Sigma Q^\intercal Q(\Sigma^\intercal P^\intercal P\Sigma)^- Q^\intercal Q\Sigma^\intercal P^\intercal P = P - P\Sigma(\Sigma^\intercal\Sigma)^-\Sigma^\intercal = [0, P_2] \ , \tag{2.61}$$

where the second equality exploited Theorem 20.5.6 in Harville (2008), the third equality follows from P and Q being orthonormal, and the fourth equality is obtained by carrying out the Moore-Penrose inverse by Exercise 2.7.4 in Magnus and Neudecker (2007) and noting that Σ is diagonal. In view of (2.61), we have

$$\min_{V\in\mathbb{S}^{d^*\times d}} \|(I - \Pi(\Pi^\intercal\Pi)^-\Pi^\intercal)MQ_2V\|^2 = \min_{V\in\mathbb{S}^{d^*\times d}} \|[0, P_2]P^\intercal MQ_2V\|^2$$
$$= \min_{V\in\mathbb{S}^{d^*\times d}} \|P_2P_2^\intercal MQ_2V\|^2 = \min_{V\in\mathbb{S}^{d^*\times d}} \|P_2^\intercal MQ_2V\|^2 = \sum_{j=r-r^*+1}^{k-r^*} \sigma_j^2(P_2^\intercal MQ_2) \ , \tag{2.62}$$

where the third equality follows from $P_2^\intercal P_2 = I_{m-r^*}$ and the final equality follows from Lemma 2.3.1. Combining (2.60) and (2.62) concludes the proof of the proposition. ■

PROOF OF PROPOSITION 2.3.3: The first and second results are straightforward application of Theorem 2.1 in Fang and Santos (2015) and Chen and Fang (2015) by noting that $\phi'_{\Pi_0} = 0$ under H_0, respectively. In particular, Assumptions 2.1(i)-(ii) are satisfied in view of Proposition 2.3.1 and Assumption 2.2 is satisfied by Assumption 2.3.1. ■

PROOF OF THEOREM 2.3.1: By Lemma 2.7.5 and the maintained assumptions, each of the two derivative estimators are consistent for ϕ''_{Π_0} in the sense that they satisfy Assumption 3.4 in Chen and Fang (2015). This, together with Lemma A.2 in Chen and Fang (2015), Assumption 2.3.2, Proposition 2.3.3, and the cdf of the limit distribution being strictly increasing at its $1-\alpha$ quantile $c_{1-\alpha}$, implies that $\hat{c}_{1-\alpha} \xrightarrow{p} c_{1-\alpha}$, following exactly the same proof of Corollary 3.2 in Fang and Santos (2015). Then under H_0, the first claim of the theorem follows from combining Proposition 2.3.3, Slutsky's lemma, $c_{1-\alpha}$ being a continuity point of the limit distribution and the portmanteau theorem.

For the second part of the theorem, let $\mathbb{G}_n^* \equiv \tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\}$. By the definition of $\hat{c}_{1-\alpha}$, if $P_W(\hat{\phi}''_n(\mathbb{G}_n^*) \leq \tau_n^2\phi(\hat{\Pi}_n)) \geq 1-\alpha$, then we must have $\hat{c}_{1-\alpha} \leq \tau_n^2\phi(\hat{\Pi}_n)$ and hence

$$P(\tau_n^2\phi(\hat{\Pi}_n) \geq \hat{c}_{1-\alpha}) \geq P_X(P_W(\hat{\phi}''_n(\mathbb{G}_n^*) \leq \tau_n^2\phi(\hat{\Pi}_n)) \geq 1-\alpha) \ . \tag{2.63}$$

We shall show that the right side of (2.63) tends to one as $n \to \infty$ for each of the two derivative estimators. First, consider the numerical estimator (2.39). Note that

$$\begin{aligned} P_W(\hat{\phi}''_n(\mathbb{G}_n^*) \leq \tau_n^2\phi(\hat{\Pi}_n)) &= P_W(\frac{\phi(\hat{\Pi}_n + \kappa_n\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\}) - \phi(\hat{\Pi}_n)}{\kappa_n^2} \leq \tau_n^2\phi(\hat{\Pi}_n)) \\ &\geq P_W(\frac{\phi(\hat{\Pi}_n + \kappa_n\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\})}{\kappa_n^2} \leq \tau_n^2\phi(\hat{\Pi}_n)) \\ &= P_W(\phi(\hat{\Pi}_n + \kappa_n\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\}) \leq (\kappa_n\tau_n)^2\phi(\hat{\Pi}_n)) \ . \end{aligned} \tag{2.64}$$

Since $\hat{\Pi}_n \xrightarrow{p} \Pi_0$ and ϕ is continuous at Π_0, the continuous mapping theorem implies that: under H_1,

$$\phi(\hat{\Pi}_n) \xrightarrow{p} \phi(\Pi_0) > 0 \ . \tag{2.65}$$

By Assumptions 2.3.1 and 2.3.2, together with the assumption that $\kappa_n = o(1)$ as $n \to \infty$ and continuity of ϕ, we have $\phi(\hat{\Pi}_n + \kappa_n\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\}) = O_{P_W}(1)$ with probability approaching one. Consequently, by $\kappa_n\tau_n \to \infty$, with probability approaching one,

$$P_W(\phi(\hat{\Pi}_n + \kappa_n\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\}) \leq (\kappa_n\tau_n)^2\phi(\hat{\Pi}_n)) \to 1 > 1 - \alpha \ . \tag{2.66}$$

By the dominated convergence theorem, we may conclude from results (2.64), (2.65) and (2.66) that

$$P_X(P_W(\hat{\phi}_n''(\mathbb{G}_n^*) \leq \tau_n^2\phi(\hat{\Pi}_n)) \geq 1 - \alpha) \to 1 \ . \tag{2.67}$$

This implies the second claim of the theorem holds when $\hat{\phi}_n''$ is the numerical derivative estimator. Second, consider the derivative estimator (2.38). Recall that $\hat{d}_n = k - \hat{r}_n$ and $d = k - r$. By Lemma 2.3.1, we have

$$\begin{aligned} P_W(\hat{\phi}_n''(\mathbb{G}_n^*)) \leq \tau_n^2\phi(\hat{\Pi}_n) &= P_W(\min_{U \in \mathbb{S}^{\hat{d}_n \times d}} \|\hat{P}_{2,n}^\intercal \tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\}\hat{Q}_{2,n}U\|^2 \leq \tau_n^2\phi(\hat{\Pi}_n)) \\ &\geq P_W(\|\tau_n\{\hat{\Pi}_n^* - \hat{\Pi}_n\}\|^2 mkd \leq \tau_n^2\phi(\hat{\Pi}_n)) \ , \end{aligned}$$

where the second inequality exploited $\|\hat{P}_{2,n}^\intercal\|^2\|\hat{Q}_{2,n}\|^2 \leq mk$ and $\|U\|^2 = d$. The second claim of the theorem then follows by analogous arguments as above. ■

Proof of Theorem 2.4.1: We prove the results for three different cases: when $r_0 = k$, when $1 \leq r_0 \leq k-1$ and when $r_0 = 0$. It suffices to show the first two results. First, we show the second result. When $r_0 = k$, we have

$$\lim_{n\to\infty} P(\hat{r}_n^* = r_0) = \lim_{n\to\infty} P(\psi_n^{(0)} = 1, \cdots, \psi_n^{(k-1)} = 1) \geq 1 - \sum_{q=0}^{k-1}(1 - \lim_{n\to\infty} P(\psi_n^{(q)} = 1)) = 1 \ ,$$

where the inequality follows by the Boole's inequality. When $1 \leq r_0 \leq k-1$, we have

$$\begin{aligned}\lim_{n\to\infty} P(\hat{r}_n^* = r_0) &= \lim_{n\to\infty} P(\psi_n^{(0)} = 1, \cdots, \psi_n^{(r_0-1)} = 1, \psi_n^{(r_0)} = 0) \\ &\leq 1 - \lim_{n\to\infty} P(\psi_n^{(r_0)} = 1) = 1 - \alpha \ , \end{aligned} \tag{2.68}$$

where the inequality follows by the fact that $P(A) \leq P(B)$ for $A \subset B$, and

$$\begin{aligned}\lim_{n\to\infty} P(\hat{r}_n^* = r_0) &= \lim_{n\to\infty} P(\psi_n^{(0)} = 1, \cdots, \psi_n^{(r_0-1)} = 1, \psi_n^{(r_0)} = 0) \\ &\geq 1 - \sum_{q=0}^{r_0-1} (1 - \lim_{n\to\infty} P(\psi_n^{(q)} = 1)) - \lim_{n\to\infty} P(\psi_n^{(r_0)} = 1) = 1 - \alpha \ , \end{aligned} \tag{2.69}$$

where the inequality follows by the Boole's inequality. Combining results (2.68) and (2.69) gives the result when $1 \leq r_0 \leq k-1$. When $r_0 = 0$, we have

$$\lim_{n\to\infty} P(\hat{r}_n^* = r_0) = \lim_{n\to\infty} P(\psi_n^{(0)} = 0) = 1 - \lim_{n\to\infty} P(\psi_n^{(0)} = 1) = 1 - \alpha \ .$$

Next, we show the first result. When $r_0 = k$, we have

$$\lim_{n\to\infty} P(\hat{r}_n^* < r_0) \leq \sum_{q=0}^{k-1} (1 - \lim_{n\to\infty} P(\psi_n^{(q)} = 1)) = 0 \ ,$$

where the inequality holds by the Boole's inequality. When $1 \leq r_0 \leq k-1$, we have

$$\lim_{n\to\infty} P(\hat{r}_n^* < r_0) \leq \sum_{q=0}^{r_0-1} (1 - \lim_{n\to\infty} P(\psi_n^{(q)} = 1)) = 0 \ ,$$

where the inequality holds by the Boole's inequality. When $r_0 = 0$, obviously $P(\hat{r}_n^* < r_0) = 0$. This completes the proof of the theorem. ■

Lemma 2.7.1. *Suppose* $\Pi \in \mathbf{M}^{m\times k}$ *with* $\Pi \neq 0$ *and rank*$(\Pi) \leq r$. *For* $\epsilon > 0$, *let* $\Psi(\Pi)_1^\epsilon$ *be given in the proof of Proposition 2.3.1. Let* $\sigma_{\min}^+(\Pi)$ *be the smallest positive singular value*

of Π. *Then for all sufficiently small* $\epsilon > 0$, *we have*

$$\min_{U \in \Psi(\Pi)_1^\epsilon} \|\Pi U\| \geq \frac{\sqrt{2}}{2} \sigma_{\min}^+(\Pi)\epsilon \ .$$

PROOF: Let $\Pi = P\Sigma Q^\intercal$ be a singular value decomposition of Π, where $P \in \mathbb{S}^{m\times m}$ and $Q \in \mathbb{S}^{k\times k}$ are orthonormal, and $\Sigma \in \mathbf{M}^{m\times k}$ is diagonal with diagonal entries in descending order. Recall that $d = k - r$ and let $r^* \equiv \text{rank}(\Pi)$. For $U \in \mathbb{S}^{k\times d}$, let $U_Q \equiv Q^\intercal U$ and write $U_Q^\intercal = [U_Q^{(1)\intercal}, U_Q^{(2)\intercal}]$ such that $U_Q^{(1)} \in \mathbf{M}^{r^*\times d}$. Then we have that for $U \in \mathbb{S}^{k\times d}$,

$$\|\Pi U\| = \|P\Sigma Q^\intercal U\| = \|\Sigma U_Q\| \geq \sigma_{\min}^+(\Pi)\|U_Q^{(1)}\| \ , \tag{2.70}$$

where the second equality follows by $P^\intercal P = I_m$, and the inequality follows by the fact that Σ is diagonal with diagonal entries in descending order and $\sigma_{\min}^+(\Pi) = \sigma_{r^*}(\Pi)$ is the smallest positive entry. Let $U_Q^{(2)} = P_U^{(2)}\Sigma_U^{(2)}Q_U^{(2)\intercal}$ be a singular value decomposition of $U_Q^{(2)}$ where $Q_U^{(2)} \in \mathbb{S}^{d\times d}$, $P_U^{(2)} \in \mathbb{S}^{(k-r^*)\times(k-r^*)}$ and $\Sigma_U^{(2)} \in \mathbf{M}^{(k-r^*)\times d}$ a diagonal matrix with diagonal entries in descending order. Note that $k - r^* \geq d$ since $r^* \leq r$. It follows that for $U \in \mathbb{S}^{k\times d}$,

$$\|U_Q^{(2)}\|^2 = \sum_{j=1}^{d} \sigma_j^2(U_Q^{(2)}) \leq \sum_{j=1}^{d} \sigma_j(U_Q^{(2)}) = \text{tr}([I_d, \mathbf{0}_{r-r^*}]\Sigma_U^{(2)}) \ , \tag{2.71}$$

where the inequality follows by the fact that $\sigma_j(U_Q^{(2)}) \in [0, 1]$ as singular values of $U_Q^{(2)}$ due to $U_Q^{(2)\intercal}U_Q^{(2)} + U_Q^{(1)\intercal}U_Q^{(1)} = I_d$, and the second equality follows by noting that the diagonal entries of $\Sigma_U^{(2)}$ are singular values of $U_Q^{(2)}$. Since $\|U_Q^{(1)}\|^2 + \|U_Q^{(2)}\|^2 = \|U_Q\|^2 = d$, thus combining (2.70) and (2.71) yields that for $U \in \mathbb{S}^{k\times d}$,

$$\|\Pi U\| \geq \sigma_{\min}^+(\Pi)\sqrt{d - \text{tr}([I_d, \mathbf{0}_{r-r^*}]\Sigma_U^{(2)})} \ . \tag{2.72}$$

Since $\|U_Q^{(1)}\|^2 + \|\Sigma_U^{(2)}\|^2 = \|U_Q^{(1)}\|^2 + \|U_Q^{(2)}\|^2 = d$ and $\|[I_d, \mathbf{0}_{r-r^*}]^\intercal\|^2 = d$, then simple algebra yields that for $U \in \mathbb{S}^{k\times d}$,

$$2(d - \text{tr}([I_d, \mathbf{0}_{d-r^*}]\Sigma_U^{(2)})) = \|U_Q^{(1)}\|^2 + \|\Sigma_U^{(2)} - [I_d, \mathbf{0}_{r-r^*}]^\intercal\|^2 \ . \tag{2.73}$$

Write $Q = [Q_1, Q_2]$ such that $Q_1 \in \mathbf{M}^{k \times r^*}$. Since $Q_1^\intercal Q_1 = I_{r^*}$, $Q_2^\intercal Q_2 = I_{k-r^*}$ and $Q_1^\intercal Q_2 = 0$ as well as $P_U^{(2)}$ and $Q_U^{(2)}$ are orthonormal, we then have that for $U \in \mathbb{S}^{k \times d}$,

$$\|U_Q^{(1)}\|^2 + \|\Sigma_U^{(2)} - [I_d, \mathbf{0}_{r-r^*}]^\intercal\|^2 = \|Q_1 U_Q^{(1)} + Q_2 P_U^{(2)}(\Sigma_U^{(2)} - [I_d, \mathbf{0}_{r-r^*}]^\intercal) Q_U^{(2)\intercal}\|^2 \ . \tag{2.74}$$

Since $U_Q^{(1)} = Q_1^\intercal U$ and $U_Q^{(2)} = Q_2^\intercal U$ by construction and $Q_1 Q_1^\intercal U + Q_2 Q_2^\intercal U = U$ by $QQ^\intercal = I_k$, we then have that for $U \in \mathbb{S}^{k \times d}$,

$$|Q_1 U_Q^{(1)} + Q_2 P_U^{(2)}(\Sigma_2 - [I_d, \mathbf{0}_{r-r^*}]^\intercal) Q_U^{(2)\intercal}\|^2 = \|U - Q_2 P_U^{(2)} [I_d, \mathbf{0}_{r-r^*}]^\intercal Q_U^{(2)\intercal}\|^2 \ . \tag{2.75}$$

Clearly, $Q_2 P_U^{(2)} [I_d, \mathbf{0}_{r-r^*}]^\intercal Q_U^{(2)\intercal} \in \Psi(\Pi)$, so combining (2.73)- (2.75) yields that for $U \in \mathbb{S}^{k \times d}$,

$$2(d - \mathrm{tr}([I_d, \mathbf{0}_{r-r^*}] \Sigma_U^{(2)})) \geq \min_{U' \in \Psi(\Pi)} \|U - U'\|^2 \ . \tag{2.76}$$

Since $\Pi \neq 0$, then $\Psi(\Pi)_1^\epsilon \neq \emptyset$ for all sufficiently small $\epsilon > 0$ by Proposition 1.3.4 in Tao (2012). Fix such an ϵ. By the definition of $\Psi(\Pi)_1^\epsilon$, combining (2.72) and (2.76) yields that for all $U \in \Psi(\Pi)_1^\epsilon$,

$$\|\Pi U\| \geq \frac{\sqrt{2}}{2} \sigma_{\min}^+(\Pi) \min_{U' \in \Psi(\Pi)} \|U - U'\| \geq \frac{\sqrt{2}}{2} \sigma_{\min}^+(\Pi) \epsilon \ . \tag{2.77}$$

Then the lemma follows by applying minimum over $\Psi(\Pi)_1^\epsilon$ to both sides of (2.77) and noting that the result continues to hold for all sufficiently small $\epsilon > 0$. ■

Lemma 2.7.2. *Let the correspondence φ be as in the proof of Proposition 2.3.1. Then $\varphi(t)$ is continuous at $t = 0$.*

PROOF: Fix $U_0 \in \Psi(\Pi)$, and define the correspondence $\bar{\varphi} : \mathbf{R} \twoheadrightarrow \Gamma^\Delta$ given by $\bar{\varphi}(t) = \Gamma_{U_0}^\Delta(t)$, where $\Psi(\Pi)$, Γ^Δ and $\Gamma_{U_0}^\Delta(t)$ are given in the proof of Proposition 2.3.1. Recall that $d = k - r$. For each $\{t_n\}$ satisfying $t_n \downarrow 0$ and each $V_0 \in \bar{\varphi}(0)$, consider the function $f : \Gamma^\Delta \to \mathbf{M}^{k \times d}$

given by

$$f(V) = V_0 - \frac{t_n}{2} U_0 V^\intercal V \ .$$

Since f is continuous and Γ^Δ is compact, f is a compact map in the sense of Granas and Dugundji (2003). It follows from Theorem 0.2.3 in Granas and Dugundji (2003) that one of the following two cases must happen: i) f has a fixed point $V_{1n} \in \Gamma^\Delta$, and ii) there exists some $V_{2n} \in \Gamma^\Delta$ such that $\|V_{2n}\| = \Delta$ and $V_{2n} = \lambda_n f(V_{2n})$ with $\lambda_n \equiv \frac{\Delta}{\|f(V_{2n})\|} \in (0,1)$. In case i), since $U_0 \in \Psi(\Pi)$, $V_0 \in \bar{\varphi}(0)$ and $f(V_{1n}) = V_{1n}$, then by simple algebra we have

$$V_{1n}^\intercal U_0 + U_0^\intercal V_{1n} = (V_0 - \frac{t_n}{2} U_0 V_{1n}^\intercal V_{1n})^\intercal U_0 + U_0^\intercal (V_0 - \frac{t_n}{2} U_0 V_{1n}^\intercal V_{1n}) = -t_n V_{1n}^\intercal V_{1n} \ . \tag{2.78}$$

This together with $V_{1n} \in \Gamma^\Delta$ implies that $V_{1n} \in \bar{\varphi}(t_n)$. Moreover, since $f(V_{1n}) = V_{1n}$, $\|U_0\| = \sqrt{d}$ and $V_{1n} \in \Gamma^\Delta$, then by the sub-multiplicativity of Frobenius norm we have

$$\|V_{1n} - V_0\| = \|\frac{t_n}{2} U_0 V_{1n}^\intercal V_{1n}\| \leq \frac{t_n}{2} \sqrt{d} \Delta^2 \ . \tag{2.79}$$

In case ii), since $U_0 \in \Psi(\Pi)$, $\lambda_n^2 V_0 \in \bar{\varphi}(0)$ and $\lambda_n V_{2n} = \lambda_n^2 f(V_{2n})$, then by analogous calculations as in (2.78), we have

$$(\lambda_n V_{2n})^\intercal U_0 + U_0^\intercal (\lambda_n V_{2n}) = -t_n (\lambda_n V_{2n})^\intercal (\lambda_n V_{2n}) \ .$$

This together with $\lambda_n V_{2n} \in \Gamma^\Delta$ due to $\lambda_n \in (0,1)$ and $V_{2n} \in \Gamma^\Delta$ implies that $\lambda_n V_{2n} \in \bar{\varphi}(t_n)$. Moreover, since $\lambda_n V_{2n} = \lambda_n^2 f(V_{2n})$, then by analogous calculations as in (2.79), we have

$$\|\lambda_n V_{2n} - V_0\| \leq \|\lambda_n^2 f(V_{2n}) - \lambda_n^2 V_0\| + |\lambda_n^2 - 1| \|V_0\| \leq \frac{t_n}{2} \sqrt{d} \Delta^2 + |\lambda_n^2 - 1| \Delta \ , \tag{2.80}$$

where the first inequality follows the triangle inequality and the second inequality follows since $\lambda_n \in (0,1)$. Now, for each $n \in \mathbf{N}$, define V_n^* to be V_{1n} if case (i) happens and $\lambda_n V_{2n}$ otherwise. Let $\delta_n \equiv 1$ if case (i) happens and $\delta_n \equiv \lambda_n$ otherwise. Then $V_n^* \in \Gamma_{U_0}^\Delta(t_n)$ for all

$n \in \mathbf{N}$, and combination of (2.79) and (2.80) yields

$$\|V_n^* - V_0\| \leq \frac{t_n}{2}\sqrt{d}\Delta^2 + |\delta_n^2 - 1|\Delta \to 0 \ ,$$

where we exploited the fact that if V_{2n} exists infinitely often, $\delta_n = \lambda_n = \frac{\Delta}{\|f(V_{2n})\|} \to 1$ due to $f(V_{2n}) \to V_0$ as $n \to \infty$ and $\|V_0\| \leq \Delta$, and $t_n \to 0$ as $n \to \infty$. It follows that $\bar{\varphi}(t)$ is lower hemicontinuous at $t = 0$ by Theorem 17.21 in Aliprantis and Border (2006).

The lower hemicontinuity of $\varphi(t)$ at $t = 0$ follows easily from that of $\bar{\varphi}(t)$ again by Theorem 17.21 in Aliprantis and Border (2006). To see this, let $t_n \to 0$ and $(U_0, V_0) \in \varphi(0)$. Define (U_n^*, V_n^*) to be $U_n^* = U_0$ and V_n^* be as in previous construction for all $n \in \mathbf{N}$. Clearly, $(U_n^*, V_n^*) \to (U_0, V_0)$, implying that $\varphi(t)$ is lower hemicontinuous at $t = 0$. Since $\varphi(t)$ is contained in the compact set $\mathbb{S}^{k\times d} \times \Gamma^\Delta$ for all t, $\varphi(t)$ is upper hemicontinuous at $t = 0$ by Theorem 17.20 in Aliprantis and Border (2006). We have therefore showed that $\varphi(t)$ is continuous at $t = 0$. ■

Lemma 2.7.3. *Suppose $\Pi \in \mathbf{M}^{m\times k}$ with $\Pi \neq 0$ and rank$(\Pi) \leq r$, and $M \in \mathbf{M}^{m\times k}$ with $M \neq 0$. Let $\Psi(\Pi)$ given in the proof of Proposition 2.3.1. For $U \in \Psi(\Pi)$ and $\Delta > 0$, let $\Gamma_U^\Delta(0)$ be as in the proof of Proposition 2.3.1. Recall that $d = k - r$. When Δ is sufficiently large, then for all $U \in \Psi(\Pi)$,*

$$\min_{V\in\Gamma_U^\Delta(0)} \|\Pi V + MU\|^2 = \min_{V\in\mathbf{M}^{k\times d}} \|\Pi V + MU\|^2 \ .$$

Proof: Recall that $\Pi = P\Sigma Q^\intercal$ is a singular value decomposition of Π, where $P \in \mathbb{S}^{m\times m}$ and $Q \in \mathbb{S}^{k\times k}$ are orthonormal, and $\Sigma \in \mathbf{M}^{m\times k}$ is diagonal with diagonal entries in descending order. Recall that $r^* = \text{rank}(\Pi) < r$. We may write $\Sigma = [\Sigma_1, 0]$ such that $\Sigma_1 \in \mathbf{M}^{m\times r^*}$ is of full rank with $r^* < r$. It follows that

$$\min_{V\in\mathbf{M}^{k\times d}} \|\Pi V + MU\|^2 = \min_{V\in\mathbf{M}^{r^*\times d}} \|[P\Sigma_1 V + MU\|^2 \ . \tag{2.81}$$

By the projection theorem, the minimum on the right hand side of (2.81) is attained at

some point, say $V_1^* \in \mathbf{M}^{r^* \times d}$. Moreover, V_1^* is uniformly bounded over $U \in \Psi(\Pi)$. Let $V^* \equiv Q[V_1^{*\intercal}, 0]^\intercal \in \mathbf{M}^{k \times d}$, then the minimum on the left hand side of (2.81) is attained at V^*. Recall that $Q = [Q_1, Q_2]$, where $Q_1 \in \mathbf{M}^{k \times r^*}$. Then $V^* = Q_1 V_1^* \in \Gamma_U^\Delta(0)$ for all $U \in \Psi(\Pi)$, when Δ is sufficiently large. It implies that the minimum on the right hand side of (2.81) is attained within $\Gamma_U^\Delta(0)$ as well for all $U \in \Psi(\Pi)$, when Δ is sufficiently large. This implies that when Δ is sufficiently large,

$$\min_{V \in \Gamma_U^\Delta(0)} \|\Pi V + MU\|^2 \leq \min_{V \in \mathbf{M}^{k \times d}} \|\Pi V + MU\|^2$$

for all $U \in \Psi(\Pi)$. The reverse inequality is simply true since $\Gamma_U^\Delta(0) \subset \mathbf{M}^{k \times d}$ all $U \in \Psi(\Pi)$ and all $\Delta > 0$. This completes the proof of the lemma. ■

Lemma 2.7.4. *Suppose rank*$(\Pi) \leq r$ *and let* $\phi_\Pi'' : \mathbf{M}^{m \times k} \to \mathbf{R}$ *be given in Proposition 2.3.1. If rank*$(\Pi) = r$*, there exists a bilinear map* $\Phi_\Pi'' : \mathbf{M}^{m \times k} \times \mathbf{M}^{m \times k} \to \mathbf{R}$ *such that* $\phi_\Pi''(M) = \Phi_\Pi''(M, M)$ *for all* $M \in \mathbf{M}^{m \times k}$*; if rank*$(\Pi) < r$*, such a* Φ_Π'' *does not exist.*

Proof: Recall that $\Pi = P\Sigma Q^\intercal$ is a singular value decomposition of Π, where $P \in \mathbb{S}^{m \times m}$ and $Q \in \mathbb{S}^{k \times k}$ are orthonormal, and $\Sigma \in \mathbf{M}^{m \times k}$ is diagonal with diagonal entries in descending order. Recall that $d = k - r$. If $\text{rank}(\Pi) = r$, then Proposition 2.3.2 and Lemma 2.3.1 imply

$$\phi_\Pi''(M) = \min_{V \in \mathbb{S}^{d \times d}} \|P_2^\intercal M Q_2 V\|^2 = \|P_2^\intercal M Q_2\|^2 \ ,$$

for all $M \in \mathbf{M}^{m \times k}$, which is a quadratic form corresponding to the bilinear form $\Phi_\Pi''(M_1, M_2)$ $\equiv \text{tr}(Q_2^\intercal M_1^\intercal P_2 P_2^\intercal M_2 Q_2)$ for $M_1 \in \mathbf{M}^{m \times k}$ and $M_2 \in \mathbf{M}^{m \times k}$.

Next, suppose that $\text{rank}(\Pi) < r_0$ and assume that there exists a bilinear map Φ_Π'' corresponding to ϕ_Π''. In turn, bilinearity of Φ_Π'' implies that

$$\phi_\Pi''(M_1) + \phi_\Pi''(M_2) = \frac{\phi_\Pi''(M_1 + M_2) + \phi_\Pi''(M_1 - M_2)}{2} \tag{2.82}$$

for all $M_1 \in \mathbf{M}^{m \times k}$ and $M_2 \in \mathbf{M}^{m \times k}$. Recall that $r^* = \text{rank}(\Pi)$. If $M = P_2 H Q_2^\intercal$ for some

$H \in \mathbf{M}^{(m-r^*)\times(k-r^*)}$, then Proposition 2.3.2 and Lemma 2.3.1 imply

$$\phi''_{\Pi}(M) = \sigma^2_{r-r^*+1}(H) + \cdots + \sigma^2_{k-r^*}(H) \ . \tag{2.83}$$

Now, let $H_1 \in \mathbf{M}^{(m-r^*)\times(k-r^*)}$ be diagonal with the (j,j)th entry equal to 1 for $j = 1, \ldots, k-r^*$ and $H_2 \in \mathbf{M}^{(m-r^*)\times(k-r^*)}$ be diagonal with the (j,j)th entry equal to -1 for $j = 1$ and 1 for $j = 2, \ldots, k-r^*$. Set $M_i = P_2 H_i Q_2^\intercal$ for $i = 1, 2$, the result in (2.83) implies $\phi''_{\Pi}(M_1) = \phi''_{\Pi}(M_2) = k-r$, $\phi''_{\Pi}(M_1+M_2) = 4(k-r)-4$ and $\phi''_{\Pi}(M_1-M_2) = 0$. It follows that

$$2(k-r) = \phi''_{\Pi}(M_1) + \phi''_{\Pi}(M_2) \neq \frac{\phi''_{\Pi}(M_1+M_2) + \phi''_{\Pi}(M_1-M_2)}{2} = 2(k-r) - 2 \ ,$$

which contradicts the result (2.82). Thus, the second result of the lemma follows. ■

Lemma 2.7.5. *Suppose Assumption 2.3.1 holds, $\kappa_n \downarrow 0$ and $\tau_n \kappa_n \to \infty$. Let $\hat{\phi}''_n$ be constructed as in* (2.39) *or* (2.38)*. Then we have under* H_0*,*

$$\hat{\phi}''_n(M_n) \xrightarrow{p} \phi''_{\Pi_0}(M)$$

whenever $M_n \to M$ as $n \to \infty$ for $\{M_n\} \subset \mathbf{M}^{m\times k}$ and $M \in \mathbf{M}^{m\times k}$.

Proof: When $\hat{\phi}''_n$ is constructed as in (2.39), the result of the lemma follows by Proposition 3.1 of Chen and Fang (2015). Next we consider the derivative estimator (2.38). Recall that $d = k - r$ and let $\hat{d}_n \equiv k - \hat{r}_n$. By Lemma 2.3.1, we have

$$\begin{aligned} |\hat{\phi}''_n(M_n) - \hat{\phi}''_n(M)| \leq & \, | \min_{U\in\mathbb{S}^{\hat{d}_n\times d}} \|\hat{P}_{2,n}^\intercal M_n \hat{Q}_{2,n} U\| - \min_{U\in\mathbb{S}^{\hat{d}_n\times d}} \|\hat{P}_{2,n}^\intercal M \hat{Q}_{2,n} U\| | \\ & \times \big(\min_{U\in\mathbb{S}^{\hat{d}_n\times d}} \|\hat{P}_{2,n}^\intercal M_n \hat{Q}_{2,n} U\| + \min_{U\in\mathbb{S}^{\hat{d}_n\times d}} \|\hat{P}_{2,n}^\intercal M \hat{Q}_{2,n} U\| \big) \ , \end{aligned} \tag{2.84}$$

where the inequality follows by the formula $(a^2 - b^2) = (a+b)(a-b)$. For the first term on

the right hand side of (2.84), we have

$$|\min_{U\in\mathbb{S}^{\hat{d}_n\times d}}\|\hat{P}_{2,n}^{\intercal}M_n\hat{Q}_{2,n}U\| - \min_{U\in\mathbb{S}^{\hat{d}_n\times d}}\|\hat{P}_{2,n}^{\intercal}M\hat{Q}_{2,n}U\|| \le \sqrt{kmd}\|M_n - M\| = o_p(1)\ , \tag{2.85}$$

where the inequality follows by the Lipschitz continuity of the infimum operator, the triangle inequality and $\|\hat{P}_{2,n}\| \le \sqrt{m}$, $\|\hat{Q}_{2,n}\| \le \sqrt{k}$ and $\|U\| = \sqrt{r}$ for all $U \in \mathbb{S}^{\hat{d}_n\times d}$, and the equality follows since $M_n \to M$. For the second term on the right hand side of (2.84), we have

$$\min_{U\in\mathbb{S}^{\hat{d}_n\times d}}\|\hat{P}_{2,n}^{\intercal}M_n\hat{Q}_{2,n}U\| + \min_{U\in\mathbb{S}^{\hat{d}_n\times d}}\|\hat{P}_{2,n}^{\intercal}M\hat{Q}_{2,n}U\| \le \sqrt{kmd}\|M_n\| + \sqrt{kmd}\|M\|\ , \tag{2.86}$$

where the inequality follows by the sub-multiplicability of the Frobenius norm, $\|\hat{P}_{2,n}\| \le \sqrt{m}$, $\|\hat{Q}_{2,n}\| \le \sqrt{k}$ and $\|U\| = \sqrt{r}$ for all $U \in \mathbb{S}^{\hat{d}_n\times d}$. Combining (2.84)-(2.86), then we obtain

$$|\hat{\phi}_n''(M_n) - \hat{\phi}_n''(M)| = o_p(1)\ . \tag{2.87}$$

Recall that $\phi''_{\Pi_0}(M) = \sum_{j=r-r_0+1}^{k-r_0}\sigma_j^2(P_{0,2}^{\intercal}MQ_{0,2})$. By (2.87), Lemma 2.3.1 and 2.7.6, it suffices to show that given $\hat{r}_n = r_0$,

$$|\sum_{j=r-\hat{r}_n+1}^{k-\hat{r}_n}\sigma_j^2(\hat{P}_{2,n}^{\intercal}M\hat{Q}_{2,n}) - \sum_{j=r-r_0+1}^{k-r_0}\sigma_j^2(P_{0,2}^{\intercal}MQ_{0,2})| = o_p(1)\ . \tag{2.88}$$

Let $\hat{r}_n = r_0$. Let $\hat{q}_j$ be the jth column of $\hat{Q}_{2,n}$. Since $Q_0 \in \mathbb{S}^{k\times k}$, we may write $\hat{q}_j = Q_0\hat{u}_j$ for some (random) $\hat{u}_j \in \mathbb{S}^{k\times 1}$. Noting that $\hat{q}_j$ is the eigenvector of $\hat{\Pi}_n^{\intercal}\hat{\Pi}_n$ associated with the eigenvalue $\sigma_{r_0+j}^2(\hat{\Pi}_n)$ due to $\hat{r}_n = r_0$, we then have

$$\begin{aligned}&[\hat{\Pi}_n^{\intercal}\hat{\Pi}_n - \Pi_0^{\intercal}\Pi_0 - (\sigma_{r_0+j}^2(\hat{\Pi}_n) - \sigma_{r_0+j}^2(\Pi_0))I_k + \Pi_0^{\intercal}\Pi_0 - \sigma_{r_0+j}^2(\Pi_0)I_k]Q_0\hat{u}_j \\ &\qquad = [\hat{\Pi}_n^{\intercal}\hat{\Pi}_n - \sigma_{r_0+j}^2(\hat{\Pi}_n)I_k]\hat{q}_j = 0\ .\end{aligned} \tag{2.89}$$

Noting that $\|\hat{\Pi}_n^{\intercal}\hat{\Pi}_n - \Pi_0^{\intercal}\Pi_0\| = o_p(1)$ and $|\sigma_{r_0+j}^2(\hat{\Pi}_n) - \sigma_{r_0+j}^2(\Pi_0)| = o_p(1)$ by the continuous

mapping theorem, the Weyl inequality (Tao, 2012, Exercise 1.3.22(iv)) and Assumption 2.3.1, we then conclude from (2.89) that

$$o_p(1) = [\Pi_0^\intercal \Pi_0 - \sigma^2_{r_0+j}(\Pi_0) I_k] Q_0 \hat{u}_j = Q_0 \Sigma_0^\intercal \Sigma_0 \hat{u}_j \ , \tag{2.90}$$

where we exploited the singular value decomposition $\Pi_0 = P_0 \Sigma_0 Q_0^\intercal$, and the fact that $\sigma^2_{r_0+j}(\Pi_0) = 0$. Since the first r_0 diagonal elements of the diagonal matrix $\Sigma_0^\intercal \Sigma_0$ are positive and Q_0 being nonsingular, we may conclude from result (2.90) that the first r_0 elements of $\hat{u}_j$ are $o_p(1)$ and moreover by the definition of $\hat{q}_j$ that for some random $U_2 \in \mathbb{S}^{(k-r_0)\times(k-r_0)}$,

$$\hat{Q}_{2,n} = Q_{0,2} U_2 + o_p(1) \ , \tag{2.91}$$

By an analogous argument, we have that for some random $V_2 \in \mathbb{S}^{(m-r_0)\times(m-r_0)}$,

$$\hat{P}_{2,n} = P_{0,2} V_2 + o_p(1) \ . \tag{2.92}$$

Combining results (2.91) and (2.92) and the continuous mapping theorem yields that given $\hat{r}_n = r_0$,

$$\| \hat{P}_{2,n}^\intercal M \hat{Q}_{2,n} - V_2^\intercal P_{0,2}^\intercal M Q_{0,2} U_2 \| = o_p(1) \ . \tag{2.93}$$

Thus, (2.88) is obtained by (2.93), the continuous mapping theorem and the fact that the singular values of $V_2^\intercal P_{0,2}^\intercal M Q_{0,2} U_2$ are equal to those of $P_{0,2}^\intercal M Q_{0,2}$. This completes the proof of the lemma. ■

Lemma 2.7.6. *Suppose Assumption 2.3.1 holds,* $\kappa_n \downarrow 0$ *and* $\tau_n \kappa_n \to \infty$. *Let* $\hat{r}_n = \min\{r, \#\{1 \le j \le k : \sigma_j(\hat{\Pi}_n) \ge \kappa_n\}\}$. *Then we have under* H_0,

$$\lim_{n\to\infty} P(\hat{r}_n = r_0) = 1 \ .$$

PROOF: Noting that $\hat{r}_n > r_0$ implies $\sigma_{r_0+1}(\hat{\Pi}_n) \geq \kappa_n$ and that $\sigma_{r_0+1}(\Pi_0) = 0$, we then have

$$\begin{aligned}\limsup_{n\to\infty} P(\hat{r}_n > r_0) &\leq \limsup_{n\to\infty} P(|\sigma_{r_0+1}(\hat{\Pi}_n) - \sigma_{r_0+1}(\Pi_0)| \geq \kappa_n) \\ &\leq \limsup_{n\to\infty} P(\|\tau_n(\hat{\Pi}_n - \Pi_0)\| \geq \tau_n\kappa_n) = 0 \ , \end{aligned} \tag{2.94}$$

where the first inequality follows by $P(A) \leq P(B)$ for $A \subset B$, the second inequality follows by the Weyl inequality (Tao, 2012, Exercise 1.3.22(iv)), and the equality follows by Assumption 2.3.1 and $\tau_n\kappa_n \to \infty$. Noting that $\hat{r}_n < r_0$ implies $\sigma_{r_0}(\hat{\Pi}_n) < \kappa_n$, we then have

$$\begin{aligned}\limsup_{n\to\infty} P(\hat{r}_n < r_0) &\leq \limsup_{n\to\infty} P(|\sigma_{r_0}(\hat{\Pi}_n) - \sigma_{r_0}(\Pi_0)| > -\kappa_n + \sigma_{r_0}(\Pi_0)) \\ &\leq \limsup_{n\to\infty} P(\|\tau_n(\hat{\Pi}_n - \Pi_0)\| \geq \tau_n\sigma_{r_0}(\Pi_0)(1 - \kappa_n/\sigma_{r_0}(\Pi_0)) = 0 \ , \end{aligned} \tag{2.95}$$

where the first inequality follows by $P(A) \leq P(B)$ for $A \subset B$, the second inequality follows by the Weyl inequality (Tao, 2012, Exercise 1.3.22(iv)), and the equality follows by Assumption 2.3.1, $\sigma_{r_0}(\Pi_0) > 0$, $\tau_n \to \infty$ and $\kappa_n \downarrow 0$. Combining (2.94) and (2.95) yields

$$\limsup_{n\to\infty} P(\hat{r}_n \neq r_0) \leq \limsup_{n\to\infty} P(\hat{r}_n < r_0) + \limsup_{n\to\infty} P(\hat{r}_n > r_0) = 0 \ .$$

This completes the proof of the lemma by noting that $\lim_{n\to\infty} P(\hat{r}_n = r_0) = 1 - \lim_{n\to\infty} P(\hat{r}_n \neq r_0) = 1$. ■

2.7.2 Results for Examples 2.2.1-2.2.7

Example 2.2.2 (Continued). Suppose $\{Y_t\}_{t=1}^n$ is a sequence of data from Example 2.2.2. Let $\hat{\Pi}_n$ be the least squares estimator

$$\hat{\Pi}_n = \frac{1}{n}\sum_{t=2}^{n} \Delta Y_t Y_{t-1}^{\intercal} (\frac{1}{n}\sum_{t=2}^{n} Y_{t-1}Y_{t-1}^{\intercal})^{-1} \ . \tag{2.96}$$

Let $D_n \equiv \text{diag}(\sqrt{n}\mathbf{1}_{r_0}, n\mathbf{1}_{k-r_0})$ and $B_0 \equiv [Q_{0,1}, P_{0,2}]^{\intercal}$, where r_0, $Q_{0,1}$ and $P_{0,2}$ are given in Proposition 2.3.3. By Lemma A.2 of Liao and Phillips (2015), if Φ_0 has eigenvalues on or

inside the unit circle, then

$$(\hat{\Pi}_n - \Pi_0)B_0^{-1}D_nB_0 \xrightarrow{L} \mathcal{M} = \mathcal{M}_1 + \mathcal{M}_2 \ , \tag{2.97}$$

where $\mathcal{M}_1 \in \mathbf{M}^{k\times k}$ with $\text{vec}(\mathcal{M}_1) \sim N(0, \Sigma \otimes (Q_{0,1}\Sigma_1^{-1}Q_{0,1}^\intercal))$ and $\Sigma_1 \equiv \text{Var}(Q_{0,1}^\intercal Y_t)$, and $\mathcal{M}_2 \in \mathbf{M}^{k\times k}$ with

$$\mathcal{M}_2 \sim \Sigma^{1/2}\int_0^1 dB_k(t)B_k(t)^\intercal\Sigma^{1/2}P_{0,2}(P_{0,2}^\intercal\Sigma^{1/2}\int_0^1 B_k(t)B_k(t)^\intercal dt\Sigma^{1/2}P_{0,2})^{-1}P_{0,2}^\intercal$$

and $B_k(t)$ is a $k\times 1$ Brownian motion defined on the unit interval with identity covariance matrix at time t. Given that Assumption 2.3.1 is not satisfied since the rates in D_n are not homogenous unless $r_0 = 0$ or $r_0 = k$, we extend Proposition 2.3.3 to accommodate this case. Next we focus on the nontrivial case of testing for the existence of stochastic trend. By Proposition 2.7.2, the asymptotic distribution of $n^2\phi(\hat{\Pi}_n)$ under H_0 is given by

$$\sum_{j=r-r_0+1}^{k-r_0}\sigma_j^2(\Sigma_{r_0}^{1/2}\int_0^1 dB_{k-r_0}(t)B_{k-r_0}(t)^\intercal(\int_0^1 B_{k-r_0}(t)B_{k-r_0}(t)^\intercal dt)^{-1}\Sigma_{r_0}^{-1/2}P_{0,2}^\intercal Q_{0,2}) \ , \tag{2.98}$$

where $\Sigma_{r_0} = P_{0,2}^\intercal\Sigma P_{0,2}$ and $Q_{0,2}$ is given in Proposition 2.3.3. When $r_0 < k-1$, the asymptotic distribution can be highly nonstandard. Note that $P_{0,2}$ and $Q_{0,2}$ are identified up to postmultiplication by $(k-r_0)\times(k-r_0)$ orthonormal matrices, so the weak limits in (2.97) and (2.98) are invariant to the choice of $P_{2,0}$ and $Q_{2,0}$.

Another distinct feature of this example is that $\mathcal{M}$ depends on Π_0, in particular, on r_0. This presents a challenge for estimating $\mathcal{M}$ by bootstrap. We propose a residual based bootstrap following Swensen (2006) and Cavaliere et al. (2012). To this end, we need a consistent estimator for r_0, that can be obtained by various methods, for example, the estimator $\hat{r}_n$ used in (2.38). We propose the following bootstrap algorithm.

1. Given the consistent estimator $\hat{r}_n$ of r_0, calculate the reduced rank estimate $\hat{\Pi}_{r,n}$ and the corresponding residuals $\hat{u}_{r,t}$, for example, following Johansen (1991). Let $\hat{u}_{r,t}^c \equiv \hat{u}_{r,t} - n^{-1}\sum_{t=1}^n \hat{u}_{r,t}$, i.e., $\hat{u}_{r,t}^c$ are recentered residuals of $\hat{u}_{r,t}$.

2. Check that $\det|(1-z)I_k - \hat{\Pi}_{r,n}z|$ has $k - \hat{r}_n$ roots equal to one and all other roots outside the unit circle. If so, proceed to the next step.

3. Construct the bootstrap sample $\{Y_t^*\}_{t=1}^n$ recursively from (2.7) with the initial value Y_0, $\Pi_0 = \hat{\Pi}_{r,n}$, and u_t^* being generated from $\{\hat{u}_{r,t}^c\}_{t=1}^n$ by the nonparametric bootstrap. Calculate the least squares estimator

$$\hat{\Pi}_n^* = \frac{1}{n}\sum_{t=2}^{n} \Delta Y_t^* Y_{t-1}^{*\intercal} (\frac{1}{n}\sum_{t=2}^{n} Y_{t-1}^* Y_{t-1}^{*\intercal})^{-1} \ . \tag{2.99}$$

Let $\hat{B}_n$ is the analog of B_0 and $\hat{D}_n$ is the analog of D_n by letting $\Pi_0 = \hat{\Pi}_{r,n}$. It then can be proved that

$$(\hat{\Pi}_n^* - \hat{\Pi}_{r,n})\hat{B}_n^{-1}\hat{D}_n\hat{B}_n \xrightarrow{L^*} \mathcal{M} \tag{2.100}$$

almost surely, where $\xrightarrow{L^*}$ denotes the weak convergence conditional on the data. That is, the law of the weak limit $\mathcal{M}$ is consistently estimated by the proposed bootstrap. Note that Assumption 2.3.2 is not satisfied.

Given that Assumptions 2.3.1 and 2.3.2 are not satisfied, we extend Theorem 2.3.1 to accommodate this case. Let $\kappa_n \downarrow 0$, $n\kappa_n \to \infty$, and $\hat{\phi}_n''$ be given in (2.38). We note that the same argument in the proof of Theorem 3.2 of Fang and Santos (2015) and Theorem 3.3 of Chen and Fang (2015) can be applied to prove that the law of the weak limit in (2.98) is consistently estimated by the law of

$$\hat{\phi}_n''((\hat{\Pi}_n^* - \hat{\Pi}_{r,n})\hat{B}_n^{-1}\hat{D}_n\hat{B}_n) \tag{2.101}$$

conditional on the data. Let $\hat{c}_{1-\alpha}$ be the $1-\alpha$ quantile of (2.101) conditional on the data. Then the same argument in the proof of Theorem 2.3.1 can be applied to prove that the test of rejecting H_0 when $n^2\phi(\hat{\Pi}_n) > \hat{c}_{1-\alpha}$ controls the asymptotic null rejection rate and is consistent. ■

Example 2.2.4-2.2.7 (Continued). The analysis here is similar to Example 2.2.1. Sup-

pose the data is generated in Examples 2.2.4-2.2.7. In Example 2.2.4, let $\hat{\Pi}_n$ be the least squares estimator of Γ_0 from regressing Y_t on Z_t and W_t based on (2.11). In Examples 2.2.5-2.2.7, let $\hat{\Pi}_n$ be the method of moment estimators based on (2.14), (2.16) and (2.18), respectively. Then, under certain weak dependence and moment condition, Assumption 2.3.1 is satisfied by all of four examples with $\tau_n = \sqrt{n}$ and $\mathcal{M}$ bing a zero mean Gaussian. Specifically, in Example 2.2.4 the Gaussian limit follows by the standard result of linear regression, and Examples 2.2.5-2.2.7 the Gaussian limit follows by the central limit theorem.

Let the resampled data be generated by the nonparametric bootstrap when the original data is a sequnce of i.i.d. data, and by a block bootstrap when the original data is a sequence of dependent data. Then, under certain weak dependence and moment condition, in Example 2.2.4 Assumption 2.3.2 is satisfied with $\hat{\Pi}^*_n$ being be the least squares estimator of Γ_0 from regressing Y^*_t on Z^*_t and W^*_t based on (2.11), and in Examples 2.2.5-2.2.7 Assumption 2.3.2 is satisfied with $\hat{\Pi}^*_n$ being the method of moment estimators based on (2.14), (2.16) and (2.18), respectively. ■

Proposition 2.7.1. *Let $\phi : \mathbf{M}^{k\times k} \to \mathbf{R}$ be defined as in* (2.24)*. For $\Pi \in \mathbf{M}^{k\times k}$ satisfying $\phi(\Pi) = 0$, let r^*, P_2, Q_1 and Q_2 be given in Proposition 2.3.2. Let $B^* \equiv [Q_1, P_2]^\intercal$. Then for $\Pi \in \mathbf{M}^{k\times k}$ satisfying $\phi(\Pi) = 0$, we have*

$$\lim_{n\to\infty} \frac{\phi(\Pi + M_n T^*_n B^*)}{t^4_n} = \sum_{j=r-r^*+1}^{k-r^*} \sigma^2_j(P_2^\intercal M Q_2) \quad \textit{with } T^*_n \equiv \mathrm{diag}(t_n \mathbf{1}_{r^*}, t^2_n \mathbf{1}_{k-r^*}) \ ,$$

for all sequences $\{M_n\} \subset \mathbf{M}^{k\times k}$ and $\{t_n\} \subset \mathbf{R}^+$ such that $t_n \downarrow 0$, $M_n B^* \to M \in \mathbf{M}^{m\times k}$ as $n \to \infty$.

Proof: Let $\{M_n\} \subset \mathbf{M}^{k\times k}$ be such that $M_n B^* \to M \in \mathbf{M}^{k\times k}$ and $t_n \downarrow 0$ as $n \to \infty$. Write $M_n = [M_{n,1}, M_{n,2}]$ such that $M_{n,1} \in \mathbf{M}^{k\times r^*}$, and $M = M_1 + M_2$ with $M_{n,1}Q_1^\intercal \to M_1$ and $M_{n,2}P_2^\intercal \to M_2$. Clearly, $M_1 U = 0$ for all $U \in \Psi(\Pi)$. Recall that $d = k - r$. For $\epsilon > 0$, let $\Psi(\Pi)^\epsilon$ and $\Psi(\Pi)^\epsilon_1$ be given in the proof of Proposition 2.3.1. In what follows we consider the nontrivial case with $\Pi \neq 0$ and $M_2 \neq 0$. In this case, $\Psi(\Pi) \subsetneqq \mathbb{S}^{k\times d}$ in

view of Proposition 1.3.4 in Tao (2012) and hence $\Psi(\Pi)_1^{\epsilon} \neq \emptyset$ for ϵ sufficiently small. Let $\sigma_{\min}^{+}(\Pi)$ be the smallest positive singular value of Π, which exists since $\Pi \neq 0$. Let $\Delta \equiv 5\sqrt{2}[\sigma_{\min}^{+}(\Pi)]^{-1}(\max_{U\in\mathbb{S}^{k\times d}}\|M_2U\| + \max_{U\in\mathbb{S}^{k\times d}}\|M_1U\|) > 0$, which holds since $M_2 \neq 0$. Then it follows that for all n sufficiently large

$$\begin{aligned}
&\min_{U\in\Psi(\Pi)_1^{t_n\Delta}} \|(\Pi + M_nT_n^*B^*)U\| \geq \min_{U\in\Psi(\Pi)_1^{t_n\Delta}} \|\Pi U\| - \max_{U\in\mathbb{S}^{k\times d}} \|M_nT_n^*B^*U\| \\
&\geq \frac{\sqrt{2}}{2} t_n \sigma_{\min}^{+}(\Pi)\Delta - t_n \max_{U\in\mathbb{S}^{k\times d}} \|M_{n,1}Q_1^{\intercal}U\| - t_n^2 \max_{U\in\mathbb{S}^{k\times d}} \|M_{n,2}P_2^{\intercal}U\| \\
&> t_n^2 \max_{U\in\mathbb{S}^{k\times d}} \|M_{n,2}P_2^{\intercal}U\| \geq \min_{U\in\Psi(\Pi)} \|(\Pi + M_nT_n^*B^*)U\| \geq \sqrt{\phi(\Pi + M_nT_n^*B^*)}\ , \qquad (2.102)
\end{aligned}$$

where the first inequality follows by the Lipschitz continuity of the infimum operator, the triangle inequality and the fact that $\Psi(\Pi)_1^{t_n\Delta} \subset \mathbb{S}^{k\times d}$, the second inequality follows by Lemma 2.7.1 and the triangle inequality, the third inequality follows by the definition of Δ, $t_n \downarrow 0$, $M_{n,1}Q_1^{\intercal} \to M_1$ and $M_{n,2}P_2^{\intercal} \to M_2$ as $n \to \infty$, the fourth inequality holds by the fact that $\Pi U = 0$ and $Q_1^{\intercal}U = 0$ for $U \in \Psi(\Pi)$, and the last inequality follows by Lemma 2.3.1. Let Γ^{Δ} and the correspondence $\varphi : \mathbf{R} \twoheadrightarrow \mathbb{S}^{k\times d} \times \Gamma^{\Delta}$ be given in the proof of Proposition 2.3.1. Then it follows that

$$\begin{aligned}
\max_{U\in\Psi(\Pi)^{t_n\Delta}} \|M_nT_n^*B^*U\| &\leq t_n \max_{(U,V)\in\varphi(t_n)} \|(M_{n,1}Q_1^{\intercal})(U + t_nV)\| + t_n^2 \max_{U\in\mathbb{S}^{k\times d}} \|M_{n,2}P_2^{\intercal}U\| \\
&\leq t_n^2 \max_{V\in\Gamma^{\Delta}} \|M_{n,1}Q_1^{\intercal}V\| + t_n^2 \max_{U\in\mathbb{S}^{k\times d}} \|M_{n,2}P_2^{\intercal}U\|\ , \qquad (2.103)
\end{aligned}$$

where the first inequality follows by the triangle inequality and the fact that $\Psi(\Pi)^{t_n\Delta} \subset \mathbb{S}^{k\times d}$, and the second inequality follows by the fact that $Q_1^{\intercal}U = 0$ for $U \in \Psi(\Pi)$ and $\varphi(t_n) \subset \Psi(\Pi) \times \Gamma^{\Delta}$. By analogous arguments in (2.102), we have for all n sufficiently large

$$\begin{aligned}
&\min_{U\in\Psi(\Pi)_1^{t_n^{3/2}\Delta}\cap\Psi(\Pi)^{t_n\Delta}} \|(\Pi + M_nT_n^*B^*)U\| \geq \min_{U\in\Psi(\Pi)_1^{t_n^{3/2}\Delta}} \|\Pi U\| - \max_{U\in\Psi(\Pi)^{t_n\Delta}} \|M_nT_n^*B^*U\| \\
&\geq \frac{\sqrt{2}}{2} t_n^{3/2} \sigma_{\min}^{+}(\Pi)\Delta - t_n^2 \max_{V\in\Gamma^{\Delta}} \|M_{n,1}Q_1^{\intercal}V\| - t_n^2 \max_{U\in\mathbb{S}^{k\times d}} \|M_{n,2}P_2^{\intercal}U\| \\
&> t_n^2 \max_{U\in\mathbb{S}^{k\times d}} \|M_{n,2}P_2^{\intercal}U\| \geq \min_{U\in\Psi(\Pi)} \|(\Pi + M_nT_n^*B^*)U\| \geq \sqrt{\phi(\Pi + M_nT_n^*B^*)}\ , \qquad (2.104)
\end{aligned}$$

where the first inequality follows by the Lipschitz continuity of the infimum operator, the triangle inequality and the fact that $\Psi(\Pi)_1^{t_n^{3/2}\Delta} \cap \Psi(\Pi)^{t_n\Delta} \subset \Psi(\Pi)_1^{t_n^{3/2}\Delta}$ and $\Psi(\Pi)_1^{t_n^{3/2}\Delta} \cap \Psi(\Pi)^{t_n\Delta} \subset \Psi(\Pi)^{t_n\Delta}$, the second inequality follows by (2.103) and Lemma 2.7.1, the third inequality follows by the definition of Δ and Γ^{Δ}, $t_n \downarrow 0$, $M_{n,1}Q_1^{\intercal} \to M_1$ and $M_{n,2}P_2^{\intercal} \to M_2$ as $n \to \infty$, the fourth inequality holds by the fact that $\Pi U = 0$ and $Q_1^{\intercal}U = 0$ for $U \in \Psi(\Pi)$, and the last inequality follows by Lemma 2.3.1. By analogous arguments in (2.104), we have for all n sufficiently large

$$\min_{U \in \Psi(\Pi)_1^{t_n^2\Delta} \cap \Psi(\Pi)^{t_n^{3/2}\Delta}} \|(\Pi + M_n T_n^* B^*)U\| > \sqrt{\phi(\Pi + M_n T_n^* B^*)} \; . \tag{2.105}$$

Combining (2.102), (2.104), (2.105) and Lemma 2.3.1, we thus obtain that for all n sufficiently large

$$\phi(\Pi + M_n T_n^* B^*) = \min_{U \in \Psi(\Pi)^{t_n^2\Delta}} \|(\Pi + M_n T_n^* B^*)U\|^2 \; . \tag{2.106}$$

Now, for the right hand side of (2.106), we have

$$\begin{aligned} &| \min_{U \in \Psi(\Pi)^{t_n^2\Delta}} \|(\Pi + M_n T_n^* B^*)U\|^2 - \min_{U \in \Psi(\Pi)^{t_n^2\Delta}} \|(\Pi + t_n M_1 + t_n^2 M_2)U\|^2| \\ &\quad \leq (O(t_n^2) + O(t_n^2)) \max_{U \in \Psi(\Pi)^{t_n^2\Delta}} \|(t_n(M_{1,n}Q_1^{\intercal} - M_1) + t_n^2(M_{2,n}P_2^{\intercal} - M_2))U\| \; , \end{aligned} \tag{2.107}$$

where the inequality follows by the formula $a^2 - b^2 = (a+b)(a-b)$, the Lipschitz inequality of the infimum operator, the triangle inequality, and the fact that $\min_{U \in \Psi(\Pi)^{t_n^2\Delta}} \|(\Pi + M_n T_n^* B^*)U\| = O(t_n^2)$ and $\min_{U \in \Psi(\Pi)^{t_n^2\Delta}} \|(\Pi + M T_n^* B^*)U\| = O(t_n^2)$. For the second term

on the right hand side of (2.107), we have

$$\begin{aligned}
&\max_{U\in\Psi(\Pi)^{t_n^2\Delta}} \|(t_n(M_{1,n}Q_1^\intercal - M_1) + t_n^2(M_{2,n}P_2^\intercal - M_2))U\| \\
&\leq t_n \max_{(U,V)\in\varphi(t_n^2)} \|(M_{n,1}Q_1^\intercal - M_1)(U + t_n^2 V)\| + t_n^2 \max_{U\in\Psi(\Pi)^{t_n^2\Delta}} \|(M_{n,2}P_2^\intercal - M_2)U\| \\
&\leq \max_{V\in\Gamma^\Delta} t_n^3 \|(M_{n,1}Q_1^\intercal - M_1)V\| + t_n^2 \max_{U\in\Psi(\Pi)^{t_n^2\Delta}} \|(M_{n,2}P_2^\intercal - M_2)U\| = o(t_n^2)\ , \qquad (2.108)
\end{aligned}$$

where the first inequality follows by the triangle inequality and the definition of $\varphi(t_n^2)$, the second inequality follows by the fact that $Q_1^\intercal U = 0$ and $M_1 U = 0$ for $U \in \Psi(\Pi)$ and $\varphi(t_n^2) \subset \Psi(\Pi) \times \Gamma^\Delta$, and the equality follows by applying the sub-multiplicativity of Frobenius norm and the fact that $M_{n,1}Q_1^\intercal \to M_1$ and $M_{n,2}P_2^\intercal \to M_2$ as $n \to \infty$. Combining (2.106), (2.107) and (2.108), we then obtain

$$\phi(\Pi + M_n T_n^* B^*) = \min_{U\in\Psi(\Pi)^{t_n^2\Delta}} \|(\Pi + t_n M_1 + t_n^2 M_2)U\|^2 + o(t_n^4)\ . \qquad (2.109)$$

Next, the first term on the right hand side of (2.109) can be written as

$$\begin{aligned}
\min_{U\in\Psi(\Pi)^{t_n^2\Delta}} \|(\Pi + t_n M_1 + t_n^2 M_2)U\|^2 &= \min_{(U,V)\in\varphi(t_n^2)} \|(\Pi + t_n M_1 + t_n^2 M_2)(U + t_n^2 V)\|^2 \\
&= t_n^4 \min_{(U,V)\in\varphi(t_n^2)} \|\Pi V + MU\|^2 + o(t_n^4)\ , \qquad (2.110)
\end{aligned}$$

where the second equality follows by the fact that $\Pi U = 0$ and $M_1 U = 0$ for $U \in \Psi(\Pi)$ and $\|V\| \leq \Delta$ for all $V \in \Gamma^\Delta$. By analogous arguments in (2.59), we have

$$\min_{(U,V)\in\varphi(t_n^2)} \|\Pi V + MU\|^2 = \min_{U\in\Psi(\Pi)} \min_{V\in\mathbf{M}^{k\times d}} \|\Pi V + MU\|^2 + o(1)\ . \qquad (2.111)$$

Combining (2.109), (2.110) and (2.111), we may conclude that

$$\lim_{n\to\infty} \frac{\phi(\Pi + M_n T_n^* B^*)}{t_n^4} = \min_{U\in\Psi(\Pi)} \min_{V\in\mathbf{M}^{k\times d}} \|\Pi V + MU\| = \sum_{j=r-r^*+1}^{k-r^*} \sigma_j^2(P_2^\intercal M Q_2)\ , \qquad (2.112)$$

where the second equality follows by Proposition 2.3.2. This completes the proof of the

lemma. ■

Proposition 2.7.2. *Suppose* $\Pi_0 \in \mathbf{M}^{k\times k}$*, and let* r_0*,* $Q_{0,1}$ *and* $P_{0,2}$ *be given in Proposition 2.3.3. Suppose there are* $\hat{\Pi}_n : \{X_i\}_{i=1}^n \to \mathbf{M}^{k\times k}$ *such that* $(\hat{\Pi}_n - \Pi_0)B_0^{-1}D_nB_0 \xrightarrow{L} \mathcal{M}$ *for some* $\tau_n \uparrow \infty$ *and random matrix* $\mathcal{M} \in \mathbf{M}^{k\times k}$*, where* $D_n \equiv diag(\tau_n \mathbf{1}_{r_0}, \tau_n^2 \mathbf{1}_{k-r_0})$ *and* $B_0 \equiv [Q_{0,1}, P_{0,2}]^\intercal$*. Then we have under* H_0*,*

$$\tau_n^4 \phi(\hat{\Pi}_n) \xrightarrow{L} \sum_{j=r-r_0+1}^{k-r_0} \sigma_j^2(P_{0,2}^\intercal \mathcal{M} Q_{0,2}) \ .$$

Proof: For each $n \in \mathbf{N}$, define $g_n : \mathbf{M}^{k\times k} \to \mathbf{R}$ by

$$g_n(M) \equiv \tau_n^4 \phi(\Pi_0 + MD_n^{-1}B_0) \ . \tag{2.113}$$

By Proposition 2.7.1, $g_n(M_n) \to \sum_{j=r-r^*+1}^{k-r^*} \sigma_j^2(P_2^\intercal M Q_2)$ whenever $M_n B^* \to M$. Note that $\tau_n^4\phi(\hat{\Pi}_n) = g_n((\hat{\Pi}_n - \Pi_0)B_0^{-1}D_n)$, then the result of the proposition follows by Theorem 1.11.1(i) in van der Vaart and Wellner (1996a). ■

2.7.3 Kleibergen and Paap (2006)'s Test

For ease of reference, we review the rank test by Kleibergen and Paap (2006). Let $\hat{\Pi}_n \in \mathbf{M}^{m\times k}$ be an estimator for $\Pi_0 \in \mathbf{M}^{m\times k}$ that satisfies Assumption 2.3.1 with $\tau_n = \sqrt{n}$ and $\mathrm{vec}(\mathcal{M}) \sim N(0, \Omega)$ for some positive semidefinite matrix Ω. Let $\hat{\Omega}_n$ be a consistent estimator of Ω. Let $\hat{\Pi}_n = \hat{P}_n \hat{\Sigma}_n \hat{Q}_n^\intercal$ be a singular value decomposition of $\hat{\Pi}_n$, where $\hat{P}_n \in \mathbb{S}^{m\times m}$ and $\hat{Q}_n \in \mathbb{S}^{k\times k}$, and $\hat{\Sigma}_n \in \mathbf{M}^{m\times k}$ is diagonal with diagonal entries in descending order. Write $\hat{P}_n = [\hat{A}_n, \hat{B}_n]$ and $\hat{Q}_n = [\hat{C}_n, \hat{D}_n]$ for $\hat{A}_n \in \mathbf{M}^{m\times r}$ and $\hat{C}_n \in \mathbf{M}^{k\times r}$, and let $\hat{S}_n$ be the right bottom $(m-r)\times(k-r)$ block submatrix of $\hat{\Sigma}_n$. Then the test statistic for the hypotheses (2.2) is given by

$$\mathrm{rk}(r) = n\mathrm{vec}(\hat{S}_n)^\intercal [(\hat{D}_n \otimes \hat{B}_n)^\intercal \hat{\Omega}_n (\hat{D}_n \otimes \hat{B}_n)]^{-1} \mathrm{vec}(\hat{S}_n) \ , \tag{2.114}$$

where $\otimes$ denotes the kronecker product. Thus, the rank test with the nominal level $\alpha \in (0,1)$ rejects the null $\mathrm{H}_0^{(r)}$ in the hypotheses (2.2) whenever $\mathrm{rk}(r) > \chi^2((m-r)(k-r), 1-\alpha)$. Note that $\hat{B}_n$ and $\hat{D}_n$ can be chosen up to postmultiplication by $(m-r)\times(m-r)$ and $(k-r)\times(k-r)$ orthonormal matrices, respectively, but $\mathrm{rk}(r)$ is invariant to the choice of $\hat{B}_n$ and $\hat{D}_n$.

In order to examine the asymptotic behavior of the rank test when $\mathrm{rank}(\Pi_0) < r$, we consider the case with $\Pi_0 = \mathbf{0}_{2\times 2}$, Ω is positive definite and $r = 1$. Let $\mathcal{M} = \mathcal{P}\mathcal{W}\mathcal{Q}$ be a singular value decomposition of $\mathcal{M}$, where $\mathcal{P} \in \mathbb{S}^{2\times 2}$ and $\mathcal{Q} \in \mathbb{S}^{2\times 2}$, and $\mathcal{W} \in \mathbf{M}^{2\times 2}$ is diagonal with diagonal entries in descending order. Write $\mathcal{P} = [\mathcal{P}_1, \mathcal{P}_2]$ and $\mathcal{Q} = [\mathcal{Q}_1, \mathcal{Q}_2]$ for $\mathcal{P}_1 \in \mathbf{M}^{2\times 1}$ and $\mathcal{Q}_1 \in \mathbf{M}^{2\times 1}$, and let $\mathcal{S}$ be (2,2)th entry of $\mathcal{W}$. Then by Lemma 2.7.7, the asymptotic distribution of rk(1) is given by

$$\mathrm{rk}(1) \xrightarrow{L} \frac{\mathcal{S}^2}{(\mathcal{Q}_2 \otimes \mathcal{P}_2)^\intercal \Omega (\mathcal{Q}_2 \otimes \mathcal{P}_2)} \,. \tag{2.115}$$

Note that $\mathcal{P}_2$ and $\mathcal{Q}_2$ can be chosen up to a sign, respectively, but the asymptotic distribution is invariant to the choice of $\mathcal{P}_2$ and $\mathcal{Q}_2$.

We now plot the distribution function of the weak limit in (2.115) by simulation. We consider two values of Ω:

$$\Omega_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \Omega_2 = \begin{bmatrix} 1 & 0 & 0 & -0.9\sqrt{5} \\ 0 & 1 & 0.9\sqrt{5} & 0 \\ 0 & 0.9\sqrt{5} & 5 & 0 \\ -0.9\sqrt{5} & 0 & 0 & 5 \end{bmatrix} .$$

The distribution functions based on 100,000 simulation replications are plotted in Figure 2.7. The weak limit when $\Omega = \Omega_1$ is first order dominated by the $\chi^2(1)$ random variable, and the weak limit when $\Omega = \Omega_2$ first order dominates the $\chi^2(1)$ random variable. This implies that directly applying the test to (2.1) will under-reject the null when $\Omega = \Omega_1$, and

will over-reject the null when $\Omega = \Omega_2$.

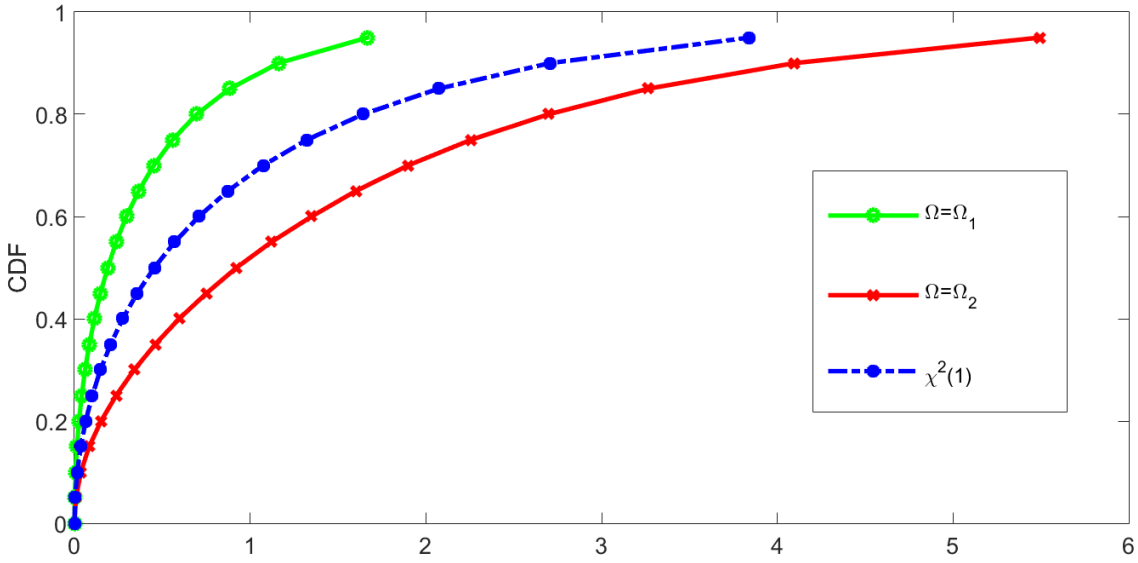


Figure 2.7: The distribution function of the weak limit of rk(1) when $\Pi_0 = \mathbf{0}_{2\times 2}$

Lemma 2.7.7. *Let* rk(r) *be given in* (2.114). *Suppose* $\Pi_0 = \mathbf{0}_{2\times 2}$ *and* Ω *is positive definite. Then the asymptotic distribution of* rk(1) *is given in* (2.115).

PROOF: For $x \in \mathbf{R}$, let $\text{sgn}(x) \equiv 1\{x \geq 0\} - 1\{x < 0\}$. Note that $\hat{D}_n$ and $\mathcal{Q}_2$ are the eigenvalue of $n\hat{\Pi}_n^\intercal \hat{\Pi}_n$ and $\mathcal{M}^\intercal \mathcal{M}$ associated with the smallest eigenvalue. By analogous arguments in Lemma 4.3 of Bosq (2000), we have

$$\|\text{sgn}(\hat{D}_n^\intercal \mathcal{Q}_2)\hat{D}_n - \mathcal{Q}_2\| \leq \frac{2\sqrt{2}}{\sigma_1^2(\mathcal{M}) - \sigma_2^2(\mathcal{M})} \|n\hat{\Pi}_n^\intercal \hat{\Pi}_n - \mathcal{M}^\intercal \mathcal{M}\| \ . \tag{2.116}$$

Similarly, we have

$$\|\text{sgn}(\hat{B}_n^\intercal \mathcal{P}_2)\hat{B}_n - \mathcal{P}_2\| \leq \frac{2\sqrt{2}}{\sigma_1^2(\mathcal{M}) - \sigma_2^2(\mathcal{M})} \|n\hat{\Pi}_n \hat{\Pi}_n^\intercal - \mathcal{M}\mathcal{M}^\intercal\| \ . \tag{2.117}$$

Note that $\sqrt{n}\hat{S}_n = \sigma_2(\sqrt{n}\hat{\Pi}_n)$ and $\mathcal{S} = \sigma_2(\mathcal{M})$. By the fact that singular values are continuous, (2.116), (2.117) and the continuous mapping theorem, we thus obtain that

$$(\sqrt{n}\hat{S}_n, \text{sgn}(\hat{B}_n^\intercal \mathcal{P}_2)\hat{B}_n^\intercal, \text{sgn}(\hat{D}_n^\intercal \mathcal{Q}_2)\hat{D}_n^\intercal) \xrightarrow{L} (\mathcal{S}, \mathcal{P}_2^\intercal, \mathcal{Q}_2^\intercal) \ . \tag{2.118}$$

Note that rk(1) does not change by replacing $\hat{B}_n$ and $\hat{D}_n$ with $\text{sgn}(\hat{B}_n^\intercal \mathcal{P}_2)\hat{B}_n$ and $\text{sgn}(\hat{D}_n^\intercal \mathcal{Q}_2)\hat{D}_n$, respectively, so the result of the lemma follows by (2.118) together with the continuous mapping theorem. ■

2.7.4 Parameters in Section 2.4.1

The values of parameters for DGP2 in the simulation studies in Section 2.4.1 are as follows:

- The value of Σ_F is specified as the sample correlation matrix of $\{F_t\}_{t=1}^T$, where $\{F_t\}_{t=1}^T$ is the real data in Section 2.4.2;
- The values of α and β are specified as $\alpha = (0.0813, -0.0271, -0.6203, -0.0460)^\intercal$ and $\beta = (-0.3411, -0.1277, -0.3838, -0.5312, -0.2728, -0.3527, -0.2188, -0.2934, -0.2035, -0.3427)^\intercal$;
- The value of Π_1 is specified as $\Pi_1 = \bar{\Pi}_T - \beta\alpha^\intercal$, where $\bar{\Pi}_T = \sum_{t=1}^T R_t F_t^\intercal (\sum_{t=1}^T F_t F_t^\intercal)^{-1}$ with $\{F_t, R_t\}_{t=1}^T$ being the real data in Section 2.4.2;

- The value of Γ is specified as

$$\Gamma = \begin{bmatrix} 0.0312 & 0.0255 & -0.0185 & 0.0591 & 0.0389 & 0.0953 & -0.1515 & 0.2286 & -0.0806 & -0.1659 \\ 0.0346 & -0.0166 & -0.0608 & 0.0743 & 0.0794 & -0.0043 & -0.2194 & 0.2959 & -0.0043 & 0.0016 \\ -0.0304 & 0.0624 & -0.1347 & 0.1054 & -0.0369 & -0.0187 & -0.0989 & 0.3571 & 0.0133 & -0.1731 \\ -0.0414 & 0.0951 & 0.0029 & -0.0497 & -0.0586 & 0.0910 & -0.0903 & 0.1850 & 0.0616 & -0.0865 \\ -0.0570 & -0.0845 & 0.0606 & -0.0143 & -0.1971 & 0.0528 & 0.0403 & 0.1935 & -0.0114 & 0.1141 \\ -0.0649 & -0.0738 & 0.0030 & 0.0335 & 0.0346 & -0.0432 & -0.0787 & 0.2199 & -0.0266 & -0.0013 \\ -0.0334 & -0.1163 & -0.0139 & -0.0218 & -0.0390 & 0.0128 & -0.0645 & 0.1299 & 0.1105 & 0.0097 \\ -0.1029 & 0.0368 & 0.0737 & -0.0005 & -0.1686 & 0.0254 & 0.0184 & 0.0966 & -0.0176 & 0.0596 \\ -0.1153 & 0.0008 & 0.0373 & 0.0185 & -0.0927 & 0.1029 & 0.0546 & 0.0529 & -0.1792 & 0.0798 \\ -0.0737 & -0.0669 & 0.0500 & 0.1466 & -0.1359 & 0.0617 & 0.1090 & 0.0402 & -0.0659 & -0.0440 \end{bmatrix};$$

- The value of Σ_v is specified as

$$\Sigma_v = \frac{1}{100}\begin{bmatrix} 0.19 & 0.09 & 0.07 & 0.05 & 0.04 & 0.03 & 0.02 & -0.01 & 0.00 & -0.01 \\ 0.09 & 0.11 & 0.06 & 0.05 & 0.04 & 0.04 & 0.03 & 0.01 & 0.02 & 0.01 \\ 0.07 & 0.06 & 0.10 & 0.05 & 0.04 & 0.04 & 0.03 & 0.03 & 0.02 & 0.01 \\ 0.05 & 0.05 & 0.05 & 0.08 & 0.04 & 0.04 & 0.04 & 0.03 & 0.02 & 0.01 \\ 0.04 & 0.04 & 0.04 & 0.04 & 0.08 & 0.05 & 0.05 & 0.05 & 0.04 & 0.03 \\ 0.03 & 0.04 & 0.04 & 0.04 & 0.05 & 0.08 & 0.06 & 0.05 & 0.05 & 0.03 \\ 0.02 & 0.03 & 0.03 & 0.04 & 0.05 & 0.06 & 0.08 & 0.06 & 0.05 & 0.03 \\ -0.01 & 0.01 & 0.03 & 0.03 & 0.05 & 0.05 & 0.06 & 0.10 & 0.07 & 0.05 \\ 0.00 & 0.02 & 0.02 & 0.02 & 0.04 & 0.05 & 0.05 & 0.07 & 0.09 & 0.04 \\ -0.01 & 0.01 & 0.01 & 0.01 & 0.03 & 0.03 & 0.03 & 0.05 & 0.04 & 0.07 \end{bmatrix}.$$

Chapter 3

Robust and Optimal Estimation for Partially Linear Instrumental Variables Models with Partial Identification

Abstract

This chapter studies robust and optimal estimation of the slope coefficients in a partially linear instrumental variables model with nonparametric partial identification. We establish the root-n asymptotic normality of a penalized sieve minimum distance estimator of the slope coefficients. We show that the asymptotic normality holds regardless of whether the nonparametric function is point identified or only partially identified. However, in the presence of nonparametric partial identification, the model is not regular in the sense of Bickel et al. (1993) and the asymptotic variance matrix may depend on the penalty, so classical efficiency analysis does not apply. We then develop an optimally penalized estimator which minimizes the asymptotic variance of a linear functional of the slope coefficients es-

timator through employing an optimal penalty, and propose a feasible two-step procedure. To conduct inference, a consistent variance matrix estimator is provided. Monte Carlos simulations examine finite sample performance of our penalized estimators.

3.1 Introduction

Recently nonparametric identification failure in the nonparametric instrumental variables (NPIV) model has attracted much attention in the literature. As originally discussed in Newey and Powell (2003), nonparametric identification requires the so-called completeness condition that is much stronger than the usual covariance restrictions needed for parametric identification. Santos (2012) discussed that even with restrictions on the parameter space, the completeness condition is still a strong requirement and may fail to hold for a rich class of models. Moreover, recent work by Canay et al. (2013) showed that the completeness condition is not directly testable. In light of these, identification, estimation and inference for the NPIV model allowing for nonparametric partial identification have been extensively studied. Without nonparametric identification, Severini and Tripathi (2012) derived necessary and sufficient conditions for the identification of linear functional of the nonparametric function, a necessary condition for its $\sqrt{n}$ estimability, and the associated efficiency bound. Based on the necessary condition, Santos (2011) developed a feasible $\sqrt{n}$ asymptotically normal estimator for the identifiable linear functional. For inference, Santos (2012) developed methods for hypothesis testing for linear restrictions on the nonparametric function, which are robust to a lack of nonparametric identification. In addition, Liao and Jiang (2011) adopted a Bayesian approach to estimate the identified set of the nonparametric function.

Nonparametric identification failure may occur in semiparametric conditional moment restriction models (Ai and Chen, 2003, 2007; Chen and Pouzo, 2009) as well. In particular, Florens et al. (2012) demonstrated that the completeness condition is necessary to identify the nonparametric function in the partially linear instrumental variables (PLIV) model while it is not needed for the identification of the slope coefficients. This motivates us to consider robust and optimal estimation of the parametric components in semiparametric conditional moment restriction models without nonparametric identification. Our focus is on robust and optimal estimation while the current literature focuses on robust inference,

see Chen et al. (2011a), Hong (2012), Tao (2014) and Chernozhukov et al. (2015) for conditional moment restriction models, and Chen et al. (2011b) for the maximum likelihood setting. It is well known that in a nonstandard setting such as what is considered here, optimal estimation and inference have to be considered separately. So our paper complements the existing literature. An attractive feature of the PLIV model is that nonparametric identification failure does not affect the identifiability of the parametric component, which may not be true in general nonlinear models (Chen et al., 2014a). As such, in this paper we focus on the PLIV model and consider robust and optimal estimation of the parametric component allowing for nonparametric partial identification. To the best of our knowledge, this is the first paper studying optimal estimation for semiparametric conditional moment restriction models without nonparametric identification.

Existing estimation methods for the PLIV model rely on the identification requirement of full parameters. Florens et al. (2012) studied the kernel method for estimating the slope coefficients. Ai and Chen (2003) studied the sieve minimum distance (SMD) estimation of smooth semiparametric conditional moment restriction models with a compactness assumption, which include the PLIV model as a special case, while Chen and Pouzo (2009) studied the penalized sieve minimum distance (PSMD) estimation of nonsmooth semiparametric conditional moment restriction models without a compactness assumption, allowing for both well-posed and ill-posed problems. To establish the $\sqrt{n}$ asymptotic normality of the slope coefficients estimators, all existing methods require a strong-norm consistency and a weak-norm convergence rate faster than $n^{-1/4}$ of the nonparametric function estimator. Without nonparametric identification, strong-norm consistency generally fails while the sufficiently fast weak-norm convergence rate can still be guaranteed. In other words, the lack of identification of the nuisance nonparametric function does not affect $\sqrt{n}$ estimability of the slope coefficients, but presents important technical challenges in deriving the asymptotic distribution. As a result, these slope coefficients estimators do not necessarily exhibit asymptotic normality, which creates substantial challenges for inferential purposes. For parametric models, Phillips (1989) and Choi and Phillips (1992) showed that the asymp-

totic distribution of instrumental variables (IV) estimator for the identified coefficients in partially identified structural equations is a variance matrix mixture of normals.

To fix the problem, we use penalization to select and consistently estimate a unique element from the identified set of the nuisance nonparametric function, which fortunately is enough to obtain the asymptotic normality. Specifically, we design a penalty function that has a unique minimizer over the identified set and add it to the model-based criterion. When the model-based criterion fails to identify the true nuisance nonparametric function, the penalty takes effect to select a unique element from the identified set. Given an appropriate penalty tuning parameter, the desired consistency and convergence rate of the estimator for the selected nuisance nonparametric function follows and then the asymptotic normality of the slope coefficients estimator is assured. If nonparametric identification is assumed, the results are consistent with Ai and Chen (2003) and Chen and Pouzo (2009). When the nonparametric function is only partially identified, the slope coefficients estimator still enjoys the usual property of being asymptotically normal. For nonparametric models, the method of achieving identification by penalization has been studied in Florens et al. (2011) and Chen and Pouzo (2012a) to obtain a consistent estimator of the parameter of interest. Here our ultimate goal is to obtain an asymptotically normal estimator of the slope coefficients, rather than a consistent estimator of the nuisance nonparametric function. In contrast to Chen and Pouzo (2009) that used penalization to deal with the ill-posed problem arising from discontinuity, we use penalization to deal with the ill-posed problem arising from noninjectivity (Kress, 2013).

For our penalized estimator of the slope coefficients, the asymptotic variance matrix may depend on the penalty in the presence of nonparametric partial identification. Heuristically, different penalty functions pin down a different function in the identified set, which then give arise to a different identification noise defined as the discrepancy between the true nonparametric function and the function selected by the given penalty. The identification noise appears as an important part as the error term in the asymptotic variance matrix. The dependence can be explained by the perhaps expected but unfortunate finding that the

model with nonparametric partial identification is not regular in the sense of Bickel et al. (1993), in particular, the slope coefficients are not continuous in the underlying distribution. This in turn implies that classical efficiency analysis cannot be applied. As such, we study the optimality in the sense of minimizing the asymptotic variance of a linear functional of the slope coefficients estimator through employing an optimal penalty, and develop a two-step feasible optimally penalized estimator. It is worth pointing out that the feasibility requires only consistency of the penalty estimator, which is in contrast to the sufficiently fast convergence rate requirement on the weight estimator in the optimally weighted estimation (Ai and Chen, 2003; Chen and Pouzo, 2009). In fact, only a consistent initial slope coefficients estimator is needed in the two-step procedure and thus is easy to satisfy. With an optimal weight, our optimally penalized estimator is superior over the estimators in Ai and Chen (2003) and Chen and Pouzo (2009). When the nonparametric function is point identified, our estimator gives an efficient estimator. When the nonparametric function is only partially identified, our estimator exhibits asymptotic normality with locally minimized variance. To conduct inference, we provide a consistent variance matrix estimator, which is directly available in the two-step procedure.

The remainder of the chapter is organized as follows. Section 3.2 discusses the motivation arising from the identification concern. Section 3.3 establishes the strong-norm consistency, weak-norm convergence rate and asymptotic normality of the penalized estimator. Section 3.4 develops an optimally penalized estimator. In Section 3.5, a consistent variance matrix estimator is provided, while Monte Carlo simulation studies are presented in Section 3.6. Section 3.7 briefly concludes. All the proofs are collected in the appendices. For a random vector V, we use its calligraphic version $\mathcal{V}$ to denote its support and $L^2(V) \equiv \{g : \mathcal{V} \to \mathbf{R} : \mathrm{E}[g^2(V)] < \infty\}$.

3.2 The Model

We consider the PLIV model

$$Y = X'\beta_0 + \phi_0(Z) + \varepsilon \text{ and } \mathrm{E}[\varepsilon|W] = 0, \tag{3.1}$$

where (3.1) specifies the structural equation for the dependent variable $Y \in \mathbf{R}$, and $W \in \mathbf{R}^{d_w}$ are IVs that are mean independent of the error term $\varepsilon \in \mathbf{R}$. The structural function is partially linear in potentially endogenous variables $X \in \mathbf{R}^{d_x}$ and $Z \in \mathbf{R}^{d_z}$ for some $\beta_0 \in \mathbf{R}^{d_x}$ and $\phi_0 \in L^2(Z)$. This covers the linear IV model and the popular partially linear regression model (Robinson, 1988) as special cases with $\phi_0 = 0$ and X and Z being exogenous, respectively. As in the partially linear regression model, the slope coefficients β_0 are the parameter of interest while the nonparametric function ϕ_0 is the nuisance parameter. Our concern differs from that in semi-nonparametric models (Blundell et al., 2007), where the parameter of interest is the nonparametric function.

Let $T_X : \mathbf{R}^{d_x} \to L^2(W)$ be given by $T_X(\beta) = \mathrm{E}[X'\beta|W = \cdot]$ and $T_Z : L^2(Z) \to L^2(W)$ be given by $T_Z(\phi) = \mathrm{E}[\phi(Z)|W = \cdot]$. Florens et al. (2012) demonstrated that the identification of ϕ_0 in (3.1) requires the injectivity of T_Z, namely the completeness condition on the joint distribution of (Z, W), i.e.,

$$\mathrm{E}[\phi(Z)|W] = 0 \Rightarrow \phi = 0. \tag{3.2}$$

Santos (2012) discussed that (3.2) fails to hold for a rich class of models even if the parameter space for ϕ_0 is restricted. Specifically, to ensure (3.2), the parameter space and the null space of T_Z have to have a zero intersection. This is a strong and undesired requirement. Moreover, Santos (2012) demonstrated that any distribution of (Z, W) under which (3.2) holds is arbitrarily closed to a distribution under which (3.2) does not hold. As such, imposing the identification of ϕ_0 may easily run into the misspecification trouble. On the other hand, Canay et al. (2013) showed that there does not exist a test for (3.2) with

nontrival power.

Fortunately, the identification of β_0 does not hinge on (3.2). As demonstrated in Florens et al. (2012), the sufficient and necessary conditions for the identification of β_0 are the injectivity of T_X and $\mathcal{R}(T_X) \cap \mathcal{R}(T_Z) = \{0\}$, where $\mathcal{R}(T_X)$ denotes the range of T_X. For the former, the existence of a finite number of IVs satisfying certain rank condition is sufficient as in the linear IV model, and there are plenty of well developed rank tests in the literature, see Camba-Mendez and Kapetanios (2009b) for a review. Under the additively separable structure, the latter is similar to the non-multicolinearity condition. This is not as strong as (3.2), and a nontrivial test for it can be developed following Santos (2012), see Remark 3.2.1. See Florens et al. (2012) for more discussions. As such, it is routine to impose the identification of β_0. Motivated by these results, we consider the estimation of β_0 without the completeness condition (3.2).

Remark 3.2.1. One is able to construct a nontrivial test for $\mathcal{R}(T_X) \cap \mathcal{R}(T_Z) = \{0\}$ allowing for a violation of (3.2). The negation is that there exists $\phi^* \in L^2(Z)$ such that

$$\inf_{\beta \in \mathbb{S}^{d_x}} |\operatorname{E}[X'\beta - \phi^*(Z)|W]| = 0,$$

where $\mathbb{S}^{d_x} \equiv \{x \in \mathbf{R}^{d_x} : \|x\| = 1\}$. Let $\mathbf{\Phi}^*$ be the parameter space for ϕ^*. Following Santos (2012), under contain regularity conditions the negation is equivalent to

$$\inf_{\beta \in \mathbb{S}^{d_x}, \phi \in \mathbf{\Phi}^*} \sup_{t \in \mathcal{T}} |\operatorname{E}[(X'\beta - \phi(Z))\omega(t, W)]| = 0$$

for some $\omega : \mathcal{T} \times \mathcal{W} \to \mathbf{R}$, where $\mathcal{T} \subset \mathbf{R}^{d_t}$ is a known compact set. It suggests employing the following test statistic

$$S_n = \inf_{\beta \in \mathbb{S}^{d_x}, \phi \in \mathbf{\Phi}^*_{J_n}} \sup_{t \in \mathcal{T}_n} |\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i'\beta - \phi(Z_i))\omega(t, W_i)|,$$

where $\{X_i, Z_i, W_i\}_{i=1}^n$ is a set of observations, $\{\Phi^*_{J_n}\}_{n=1}^\infty$ and $\{\mathcal{T}_n\}_{n=1}^\infty$ are sieve spaces that grow to be dense in $\mathbf{\Phi}^*$ and $\mathcal{T}$, respectively. The limiting distribution of S_n can be similarly

developed as in Theorem 3.1 of Santos (2012) and the bootstrap critical values with size control and power property can be similarly developed as in Theorem 3.2 and Corollary 3.1 of Santos (2012), respectively.

3.3 Robust Estimation

Without the completeness condition (3.2), the SMD estimator by Ai and Chen (2003) and the PSMD estimator by Chen and Pouzo (2009) do not necessarily provide a $\sqrt{n}$ asymptotically normal estimator of β_0. This is illustrated in a simple linear IV model, see Lemma 3.9.1. In particular, the estimator is still $\sqrt{n}$ consistent, but the asymptotic distribution is highly nonstandard. The failure is due to the inconsistency of the estimator of ϕ_0. To fix the problem, we use penalization to select and consistently estimate a unique element from the identified set of ϕ_0, which fortunately is enough to obtain the asymptotic normality.

To proceed, let $\mathbf{B} \subset \mathbf{R}^{d_x}$ and $\mathbf{\Phi} \subset L^2(Z)$ denote the parameter space for β_0 and ϕ_0, respectively. Without the completeness condition (3.2), (β_0, ϕ_0) is not necessarily unique solution to

$$\min_{\beta \in \mathbf{B}, \phi \in \mathbf{\Phi}} \mathrm{E}[(\mathrm{E}[Y - X'\beta - \phi(Z)|W])^2 \sigma^{-2}(W)] \tag{3.3}$$

where the weight $\sigma^2(\cdot) > 0$ is introduced to address potential heteroscedasticity. This in turn implies that (β_0, ϕ_0) can not be consistently estimated by minimizing the sample analog of a nonparametric version of (3.3) with ϕ restricted to a sieve space for $\mathbf{\Phi}$, which is the method pursued by Ai and Chen (2003). Nevertheless, to obtain a $\sqrt{n}$ asymptotically normal estimator of β_0, it suffices to obtain a consistent estimator of any element from the identified set of ϕ_0, rather than a consistent estimator of ϕ_0. Let $\mathcal{N}(T_Z)$ denote the null space of T_Z, $\mathcal{N}(T_Z)^\perp$ denote the orthogonal complement of $\mathcal{N}(T_Z)$ and $\phi_0^\perp$ denote the

projection of ϕ_0 in $\mathcal{N}(T_Z)^\perp$. It is observed that $(\beta_0, \phi_0^\perp)$ is the unique solution to

$$\min_{\beta\in\mathbf{B},\phi\in\mathbf{\Phi}\cap\mathcal{N}(T_Z)^\perp} \mathrm{E}[(\mathrm{E}[Y - X'\beta - \phi(Z)|W])^2\sigma^{-2}(W)] \tag{3.4}$$

provided $\phi_0^\perp \in \mathbf{\Phi}$. It suggests that $(\beta_0, \phi_0^\perp)$ can be consistently estimated by minimizing the sample analog of a nonparametric version of (3.4) with ϕ restricted to a sieve space for $\mathbf{\Phi} \cap \mathcal{N}(T_Z)^\perp$. However, it is not straightforwardly feasible since $\mathcal{N}(T_Z)$ is unknown and a sieve space for $\mathbf{\Phi} \cap \mathcal{N}(T_Z)^\perp$ is not directly available.

Next we introduce how penalization solves the problem. Let $\mathbf{\Phi}_0$ denote the identified set of ϕ_0,

$$\mathbf{\Phi}_0 \equiv \{\phi_0 + \phi \in \mathbf{\Phi} : \mathrm{E}[\phi(Z)|W] = 0\}. \tag{3.5}$$

Without the completeness condition (3.2), $\mathbf{\Phi}_0$ is not necessarily a singleton. Let $P : \mathbf{\Phi} \to [0, \infty)$ be a penalty function that has a unique minimizer over $\mathbf{\Phi}_0$, which is denoted $\phi_P \equiv \arg\min_{\phi\in\mathbf{\Phi}_0} P(\phi)$. Let $(\beta_{\lambda_n,P}, \phi_{\lambda_n,P})$ denote a solution to

$$\min_{\beta\in\mathbf{B},\phi\in\mathbf{\Phi}} \mathrm{E}[(\mathrm{E}[Y - X'\beta - \phi(Z)|W])^2\sigma^{-2}(W)] + \lambda_n P(\phi) \tag{3.6}$$

where $0 < \lambda_n = o(1)$ is a penalty tuning parameter. Given the identification of β_0, a significant deviation of β from β_0 leads to a significant deviation of the first term from zero, which dominates the second term for all sufficiently large n since $0 < \lambda_n = o(1)$. A significant deviation of ϕ from ϕ_P outside $\mathbf{\Phi}_0$ leads to a significant deviation of the first term from zero while a significant deviation of ϕ from ϕ_P within $\mathbf{\Phi}_0$ leads to a significant deviation of the second term from its minimal over $\mathbf{\Phi}_0$. Therefore, it follows that $(\beta_{\lambda_n,P}, \phi_{\lambda_n,P})$ converges to (β_0, ϕ_P) as n goes to infinity. Let $\{Y_i, X_i, Z_i, W_i\}_{i=1}^n$ be a set of observations satisfying (3.1). Display (3.6) suggests that a feasible consistent estimator $(\hat{\beta}_P, \hat{\phi}_P)$ for

(β_0, ϕ_P) solves [1]

$$\inf_{\beta \in \mathbf{B}, \phi \in \mathbf{\Phi}_{J_n}} \frac{1}{n} \sum_{i=1}^{n} (\hat{\mathrm{E}}[Y - X'\beta - \phi(Z)|W_i])^2 \hat{\sigma}^{-2}(W_i) + \lambda_n \hat{P}(\phi), \tag{3.7}$$

where $\mathbf{\Phi}_{J_n}$ is a sieve space for $\mathbf{\Phi}$, $\hat{\mathrm{E}}[\cdot|W]$ is a series estimator for $\mathrm{E}[\cdot|W]$, $\hat{\sigma}^2(\cdot)$ and $\hat{P}(\cdot)$ are estimators for $\sigma^2(\cdot)$ and $P(\cdot)$, respectively. In particular, if $P(\phi) = \mathrm{E}[\phi^2(Z)]$ and $\phi_0^{\perp} \in \mathbf{\Phi}$, then $\phi_P = \phi_0^{\perp}$ since $\phi_0^{\perp}$ has the smallest norm over $\mathbf{\Phi}_0$. So our penalization method is equivalent to the nonstraightforwardly feasible method arising from (3.4). In addition, a variety of penalties can be employed and thus the penalization method applies more broadly.

For completeness, we review how to construct sieve spaces and series estimators following Chen (2007). Let $\{q_j(\cdot)\}_{j=1}^{\infty}$ denote a sequence of known basis functions (such as power series, splines, Fourier series, etc.), with the property that its linear combination can approximate any function in $\mathbf{\Phi}$ well. Let $q^{J_n}(z) \equiv (q_1(z), \ldots, q_{J_n}(z))'$, then the linear sieve space $\mathbf{\Phi}_{J_n}$ is given by

$$\mathbf{\Phi}_{J_n} \equiv \{\phi \in \mathbf{\Phi} : \phi(z) = q^{J_n}(z)'\beta, \beta \in \mathbf{R}^{J_n}\}. \tag{3.8}$$

Let $\{p_j(\cdot)\}_{j=1}^{\infty}$ denote a sequence of known basis functions (such as power series, splines, Fourier series, etc.), with the property that its linear combination can approximate any square integrable real-valued function of w well. Let $p^{k_n}(w) \equiv (p_1(w), \ldots, p_{k_n}(w))'$ and $P \equiv (p^{k_n}(W_1), \ldots, p^{k_n}(W_n))'$, then the series estimator is given by

$$\hat{\mathrm{E}}[Y - X'\beta - \phi(Z)|W] \equiv p^{k_n}(W)'(P'P)^{-} \sum_{i=1}^{n} p^{k_n}(W_i)(Y_i - X_i'\beta - \phi(Z_i)) \tag{3.9}$$

for any $\beta \in \mathbf{B}$ and $\phi \in \mathbf{\Phi}$.

[1] The estimators are indexed by P to stress the potential dependence of the asymptotic properties of the slope coefficients estimator and of the probability limit of the nonparametric function estimator on $P(\cdot)$. On the other hand, the dependence of the asymptotic properties of the slope coefficients estimator on $\sigma^2(\cdot)$ is suppressed for notational simplicity as it is not our main concern.

3.3.1 Strong-Norm Convergence

To establish the consistency of $(\hat{\beta}_P, \hat{\phi}_P)$ under the strong norm $\|\cdot\| + \|\cdot\|_\infty$,[2] we require $\mathbf{\Phi}$ to be a set of smooth functions, which ensures both consistency and the uniform behavior of the empirical process on the parameter space. In particular, we assume $\mathbf{\Phi}$ is bounded under the Sobolev norm $\|\cdot\|_{\infty,\gamma_z}$. To define $\|\cdot\|_{\infty,\gamma_z}$, for λ a d_z dimensional vector of nonnegative integers, let $|\lambda| \equiv \sum_{i=1}^{d_z} \lambda_i$ and $D^\lambda \phi(z) \equiv \partial^{|\lambda|}\phi(z)/\partial z_1^{\lambda_1} \ldots \partial z_{d_z}^{\lambda_{d_z}}$ For $\gamma_z \in \mathbf{R}$, let $\underline{\gamma_z}$ denote the greatest integer smaller than γ_z. Then the norm $\|\cdot\|_{\infty,\gamma_z}$ is given by

$$\|\phi\|_{\infty,\gamma_z} \equiv \max_{|\lambda| \leq \underline{\gamma_z}} \sup_{z \in \mathcal{Z}} |D^\lambda \phi(z)| + \max_{|\lambda| = \underline{\gamma_z}} \sup_{z \neq z'} \frac{|D^\lambda \phi(z) - D^\lambda \phi(z')|}{\|z - z'\|^{\gamma_z - \underline{\gamma_z}}} \tag{3.10}$$

A function ϕ with $\|\phi\|_{\infty,\gamma_z} < \infty$ has partial derivatives up to order $\underline{\gamma_z}$ uniformly bounded, and partial derivatives of order $\underline{\gamma_z}$ Hölder continuous with the exponent $\gamma_z - \underline{\gamma_z} \in (0, 1]$. Let $C_M^{\gamma_z}(\mathcal{Z})$ be the set of all continuous functions $\phi : \mathcal{Z} \to \mathbf{R}$ with $\|\phi\|_{\infty,\gamma_z} \leq M$, then these properties hold uniformly in $\phi \in C_M^{\gamma_z}(\mathcal{Z})$. Specifically, we assume $\mathbf{\Phi} = C_M^{\gamma_z}(\mathcal{Z})$ for some $\gamma_z > d_z/2$ and $M > 0$, and thus $\mathbf{\Phi}$ is compact under $\|\cdot\|_\infty$. Given this, we are able to establish the consistency of $\hat{\phi}_P$ under $\|\cdot\|_\infty$, though only the consistency under L^2 norm is needed.

Remark 3.3.1. In our setting, there are two possible sources for ill-posedness of T_Z, which are the noninjectivity of T_Z and the discontinuity of its inverse correspondence, see Kress (2013). As the latter is not our main concern, for simplicity we impose compactness of $\mathbf{\Phi}$ to circumvent it as in Newey and Powell (2003), Ai and Chen (2003) and Santos (2012). Alternatively, the second source of ill-posedness can be circumvented by employing a lower semicompact penalty as in Chen and Pouzo (2009, 2012a), which we do not pursue here. Therefore, our penalty concentrates on dealing with the first source of ill-posedness.

We proceed by imposing the following assumptions.

Assumption 3.3.1. *(i)* $\{Y_i, X_i, Z_i, W_i\}_{i=1}^n$ *is a set of independently and identically distributed observations satisfying* (3.1)*; (ii)* $\beta_0 \in \mathbf{B} \subset \mathbf{R}^{d_z}$ *that is compact and* $\phi_0 \in \mathbf{\Phi} =$

[2]For $\beta \in \mathbf{B}$ and $\phi \in \mathbf{\Phi}$, the strong norm of (β, ϕ) is given by $\|\beta\| + \|\phi\|_\infty$.

$C_M^{\gamma_z}(\mathcal{Z})$ *for some* $\gamma_z > d_z/2$ *and* $M > 0$*; (iii)* $\mathcal{Z}$ *is compact and convex with nonempty interior; (iv)* $\sup_{w\in\mathcal{W}} E[\|X\|^2|W=w] < \infty$ *and* $\sup_{w\in\mathcal{W}} E(\varepsilon^2|W=w) < \infty$*; (v)* $\mathcal{R}(T_X) \cap \mathcal{R}(T_Z) = \{0\}$ *and* T_X *is injective.*

Assumption 3.3.2. *(i)* $\{\mathbf{\Phi}_J : J \geq 1\}$ *is a sequence of nonempty closed subsets such that* $\mathbf{\Phi}_J \subset \mathbf{\Phi}$ *for all* J*; (ii) For* $\phi \in \mathbf{\Phi}$*, there is* $\Pi_n\phi \in \mathbf{\Phi}_{J_n}$ *such that* $\|\Pi_n\phi - \phi\|_\infty = O(J_n^{-\alpha_z})$ *with* $\alpha_z > 0$*.*

Assumption 3.3.3. *(i)* $\sup_{w\in\mathcal{W}} |\sigma^2(w) - \hat{\sigma}^2(w)| = o_p(1)$ *for some* $\sigma^2 : \mathcal{W} \to (0,\infty)$ *with* $0 < \inf_{w\in\mathcal{W}} \sigma^2(w) \leq \sup_{w\in\mathcal{W}} \sigma^2(w) < \infty$*; (ii)* $\sup_{\phi\in\mathbf{\Phi}} |\hat{P}(\phi) - P(\phi)| = O_p(\delta_{P,n})$ *for some* $\delta_{P,n} = o(1)$ *and* $P : \mathbf{\Phi} \to [0,\infty)$*; (iii)* $P(\cdot)$ *has a unique minimizer* ϕ_P *over* $\mathbf{\Phi}_0$ *and* $P(\cdot)$ *is continuous over* $\mathbf{\Phi}$ *with respect to* $\|\cdot\|_\infty$*.*

Assumption 3.3.4. *(i)* $\mathcal{W}$ *is compact and connected with Lipschitz continuous boundary; (ii) The density of* W *is bounded and bounded away from zero over* $\mathcal{W}$*; (iii) The eigenvalues of* $\mathrm{E}[p^{k_n}(W)p^{k_n}(W)']$ *are bounded and bounded away from zero for all* k_n*; (iv) Either* $k_n\xi_n^2 = o(n)$ *or* $k_n \log(k_n) = o(n)$ *for* $p^{k_n}(w)$ *a polynomial spline sieve, where* $\xi_n \equiv \sup_{w\in\mathcal{W}} \|p^{k_n}(w)\|$*; (v) There is* $\Pi_x \in \mathbf{R}^{d_x \times k_n}$ *such that* $\sup_{w\in\mathcal{W}} \|\mathrm{E}[X|W=w] - \Pi_x p^{k_n}(w)\| = O(k_n^{-\alpha_w})$ *with* $\alpha_w > 0$*; (vi) For* $\phi \in \mathbf{\Phi}$*, there is* $\pi_\phi \in \mathbf{R}^{k_n}$ *such that* $\sup_{w\in\mathcal{W}} |\mathrm{E}[\phi(Z)|W=w] - p^{k_n}(w)'\pi_\phi| = O(k_n^{-\alpha_w})$ *uniformly over* $\phi \in \mathbf{\Phi}$*.*

Assumption 3.3.1 specifies the data generating process. Specifically, Assumptions 3.3.1(ii) and (iii) impose the compact parameter spaces; Assumption 3.3.1(iv) imposes finite moment conditions and Assumption 3.3.1(v) assumes the identification condition for β_0. Assumption 3.3.2 specifies $\mathbf{\Phi}_{J_n}$ and its approximation error. In particular, the polynomial rate of approximation error is satisfied with $\alpha_z = \gamma_z/d_z$ under Assumptions 3.3.1(ii) and (iii) if $\mathbf{\Phi}_{J_n}$ is a sieve space with tensor-product of power series, splines or Fourier series. Assumption 3.3.3 specifies the weight and penalty and their estimators. Specifically, Assumption 3.3.3(i) requires uniform convergence of the weight estimator to a nondegenerate and bounded weight, which is obviously satisfied by the identity weight; Assumption 3.3.3(ii) requires uniform convergence of the penalty estimator over $\mathbf{\Phi}$, which is obviously

satisfied if $\hat{P}(\cdot) = P(\cdot)$; Assumption 3.3.3(iii) requires uniqueness of the minimizer of the penalty over $\mathbf{\Phi}_0$ and continuity of $P(\cdot)$ over $\mathbf{\Phi}$, which are key to ensure consistency. In particular, $\mathbf{\Phi}_0$ is convex and compact under Assumptions 3.3.1(ii) and (iii). Then Assumption 3.3.3(iii) is satisfied by all L^2-type penalties, i.e., $P(\phi) = \int_{\mathcal{Z}} \phi^2(z) d\mu(z)$, where μ is a finite measure. For $P(\phi) = \mathrm{E}[\phi^2(Z)]$, the natural estimator is given by $\hat{P}(\phi) = \frac{1}{n}\sum_{i=1}^n \phi^2(Z_i)$ and Assumption 3.3.3(ii) is satisfied under Assumptions 3.3.1(i)-(iii), see Lemma 3.9.2. Assumption 3.3.4 is standard in the use of series estimators for conditional mean functions, see Newey (1997), Huang (1998, 2003) and Chen and Pouzo (2009, 2012a). In particular, the polynomial rate of approximation error can be satisfied under certain smoothness as in Assumption 3.3.2.

The consistency of $(\hat{\beta}_P, \hat{\phi}_P)$ under $\|\cdot\| + \|\cdot\|_\infty$ is established in the following theorem.

Theorem 3.3.1. *Suppose Assumptions 3.3.1-3.3.4 hold. Let $(\hat{\beta}_P, \hat{\phi}_P)$ be the estimator in (3.7). If $J_n + d_x \leq k_n$, $0 < \lambda_n = o(1)$ and $\max\{\frac{k_n}{n}, k_n^{-2\alpha_w}, J_n^{-2\alpha_z}\} = o(\lambda_n)$, then $\|\hat{\beta}_P - \beta_0\| = o_p(1)$ and $\|\hat{\phi}_P - \phi_P\|_\infty = o_p(1)$.*

Theorem 3.3.1 shows that $(\hat{\beta}_P, \hat{\phi}_P)$ converges in probability to (β_0, ϕ_P) under $\|\cdot\| + \|\cdot\|_\infty$ provided the penalty tuning parameter λ_n is appropriately chosen. In particular, it is required to dominate the estimation error of the series estimator for conditional mean $(\frac{k_n}{n} + k_n^{-2\alpha_w})$ and the approximation error of the sieve space for the parameter space $(J_n^{-2\alpha_z})$. Heuristically, the penalty is effective in choosing a unique function in $\mathbf{\Phi}_0$ only when the estimation error of the model-based criterion and the approximation error of the sieve space are asymptotically negligible relative to the penalty term. On the other hand, λ_n is required to decay to zero so that the penalty term selects a unique element from $\mathbf{\Phi}_0$ rather than from $\mathbf{\Phi}$. The argument for Theorem 3.3.1 is similar to the proof of Theorem A.1 in Chen and Pouzo (2012a), though we only penalize part of the parameters. Here we provide an explicit lower bound for λ_n, which ensures the possibility of establishing the asymptotic normality of $\hat{\beta}_P$, see Assumption 3.3.6 and Remark 3.3.2.

3.3.2 Weak-Norm Convergence Rate

Following Ai and Chen (2003), we define the weak norm $\|(\beta_1,\phi_1)-(\beta_2,\phi_2)\|_w$ for any $(\beta_1,\phi_1),(\beta_2,\phi_2)\in\mathbf{B}\times\mathbf{\Phi}$ as

$$\|(\beta_1,\phi_1)-(\beta_2,\phi_2)\|_w^2\equiv \mathrm{E}[(\mathrm{E}[X'(\beta_1-\beta_2)|W]+\mathrm{E}[\phi_1(Z)-\phi_2(Z)|W])^2\sigma^{-2}(W)]. \quad (3.11)$$

It implies that the identified set $\{\beta_0\}\times\mathbf{\Phi}_0$ is an equivalence class under $\|\cdot\|_w$ since $\|(\beta_0,\phi_1)-(\beta_0,\phi_2)\|_w^2=0$ for any $\phi_1,\phi_2\in\mathbf{\Phi}_0$. It turns out that establishing the $\|\cdot\|_w$-convergence rate of $(\hat\beta_P,\hat\phi_P)$ does not present much technical challenge.

The theorem below establishes the rate of convergence of $(\hat\beta_P,\hat\phi_P)$ under $\|\cdot\|_w$.

Theorem 3.3.2. *Suppose Assumptions 3.3.1-3.3.4 hold. Let $(\hat\beta_P,\hat\phi_P)$ be the estimator in (3.7). If $J_n+d_x\le k_n$, $0<\lambda_n=o(1)$ and $\max\{\frac{k_n}{n},k_n^{-2\alpha_w},J_n^{-2\alpha_z}\}=o(\lambda_n)$, then $\|(\hat\beta_P,\hat\phi_P)-(\beta_0,\phi_P)\|_w=o_p(\sqrt{\lambda_n})$.*

Theorem 3.3.2 shows that $(\hat\beta_P,\hat\phi_P)$ converges in probability to (β_0,ϕ_P) under $\|\cdot\|_w$ at a rate faster than $\sqrt{\lambda_n}$. A rate faster than $n^{-1/4}$ requires $\lambda_n=O(n^{-1/2})$. In fact, the result still holds by replacing (β_0,ϕ_P) with any element from $\{\beta_0\}\times\mathbf{\Phi}_0$ as it is a equivalent class under $\|\cdot\|_w$ as discussed above. Moreover, without Assumptions 3.3.1(v) and 3.3.3(iii), it can be shown that $\|(\hat\beta_P,\hat\phi_P)-(\beta_0,\phi_0)\|_w=O_p(\sqrt{\lambda_n})$. So despite the loss of strong-norm consistency, the weak-norm convergence rate of $(\hat\beta_P,\hat\phi_P)$ is not lost in the absence of identification. In fact, the argument for Theorem 3.3.2 does not present much difference from the corresponding results in Chen and Pouzo (2009, 2012a), but the result can be improved from $O_p(\sqrt{\lambda_n})$ to $o_p(\sqrt{\lambda_n})$ due to Theorem 3.3.1.

3.3.3 Asymptotic Normality

Given the $\|\cdot\|+\|\cdot\|_\infty$ consistency and $\|\cdot\|_w$ convergence rate of $(\hat\beta_P,\hat\phi_P)$, we are now able to establish the asymptotic normality of $\hat\beta_P$. We first illustrate why the $\|\cdot\|+\|\cdot\|_\infty$ consistency is crucial for establishing the asymptotic normality.

For $j = 1, \ldots, d_x$, let $X^{(j)}$ be the jth component of X and ϕ_j^* be a solution to

$$\min_{\phi \in L^2(Z)} \mathrm{E}[(\mathrm{E}[X^{(j)} - \phi(Z)|W])^2 \sigma^{-2}(W)], \tag{3.12}$$

which solves the first order condition

$$\mathrm{E}[\mathrm{E}[X^{(j)} - \phi_j^*(Z)|W]\,\mathrm{E}[\phi(Z)|W]\sigma^{-2}(W)] = 0 \text{ for all } \phi \in L^2(Z). \tag{3.13}$$

Let $\Phi^* \equiv (\phi_1^*, \ldots, \phi_{d_x}^*)'$. Without the completeness condition (3.2), Φ^* may not be unique, but $\mathrm{E}[X - \Phi^*(Z)|W]$ is unique as the objective function in (3.12) is strictly convex in $\mathrm{E}[X^{(j)} - \phi(Z)|W]$. Therefore $\mathrm{E}[X - \Phi^*(Z)|W]$ and

$$\Gamma \equiv \mathrm{E}[\mathrm{E}[X - \Phi^*(Z)|W](\mathrm{E}[X - \Phi^*(Z)|W])'\sigma^{-2}(W)] \tag{3.14}$$

are independent of the choice of Φ^*, where the latter is positive definite by Assumption 3.3.1(v). By result (3.13) and the law of iterated expectation, we have the linear representation

$$\begin{aligned}\sqrt{n}(\hat{\beta}_P - \beta_0) &= \sqrt{n}\Gamma^{-1}\,\mathrm{E}[\mathrm{E}[X - \Phi^*(Z)|W](\mathrm{E}[X'(\hat{\beta}_P - \beta_0)|W] \\ &\qquad + \mathrm{E}[\hat{\phi}_P(Z) - \phi_P(Z)|W])\sigma^{-2}(W)] \\ &= -\sqrt{n}\Gamma^{-1}\,\mathrm{E}[\mathrm{E}[X - \Phi^*(Z)|W](Y - X'\hat{\beta}_P - \hat{\phi}_P(Z))\sigma^{-2}(W)]. \end{aligned} \tag{3.15}$$

In the appendix, we prove that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{E}[X - \Phi^*(Z)|W_i](Y_i - X_i'\hat{\beta}_P - \hat{\phi}_P(Z_i))\sigma^{-2}(W_i) = o_p(1), \tag{3.16}$$

which can be imagined as the analog of the orthogonality of residuals and regressors in the least squared regression. In the appendix, we also prove that the class $\mathcal{F} \equiv \{f : \mathbf{R} \times \mathbf{R}^{d_x} \times \mathcal{Z} \times \mathcal{W} \to \mathbf{R} : f(y, x, z, w) = \mathrm{E}[X - \Phi^*(Z)|W = w](y - x'\beta - \phi(z))\sigma^{-2}(w), (\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi}\}$ is Donsker. By the stochastic equicontinuity of the stochastic process indexed by $f \in \mathcal{F}$

and the $\|\cdot\| + \|\cdot\|_\infty$ consistency of $(\hat{\beta}_P, \hat{\phi}_P)$ to (β_0, ϕ_P) in Theorem 3.3.1, results (3.15) and (3.16) imply

$$\sqrt{n}(\hat{\beta}_P - \beta_0) = \Gamma^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{E}[X - \Phi^*(Z)|W_i](Y_i - X_i'\beta_0 - \phi_P(Z_i))\sigma^{-2}(W_i) + o_p(1), \quad (3.17)$$

which delivers the asymptotic normality by the Central Limit Theorem and the Slutsky lemma. The usefulness of $\|\cdot\| + \|\cdot\|_\infty$ consistency arises from the application of the stochastic equicontinuity, while $\|\cdot\|_w$ consistency is not enough since the function class $\mathcal{F}$ is not continuous in $\mathbf{B} \times \mathbf{\Phi}$ with respect to $\|\cdot\|_w$. The proof procedure is different from Ai and Chen (2003) and Chen and Pouzo (2009): both rely on nonparametric identification and the latter in addition demands strong-norm convergence rates.

To formalize the arguments above, we first impose the following assumptions.

Assumption 3.3.5. *(i) $\phi_j^* \in \mathbf{\Phi}$ for all $j = 1, \ldots, d_x$; (ii) There exists $\Pi_\sigma^* \in \mathbf{R}^{d_x \times k_n}$ such that $\sup_{w \in \mathcal{W}} \|\mathrm{E}[X - \Phi^*(Z)|W = w]\sigma^{-2}(w) - \Pi_\sigma^* p^{k_n}(w)\| = O(k_n^{-\alpha_w})$.*

Assumption 3.3.6. *(i) $n^{-1}\sum_{i=1}^n [\hat{\sigma}^2(W_i) - \sigma^2(W_i)]^2 = O_p(\delta_{\sigma,n}^2)$ for some $\delta_{\sigma,n} = o(n^{-1/4})$; (ii) There exists $\epsilon_n = o(n^{-1/2})$ such that $\lambda_n \max_{1 \le j \le d_x} \sup_{\phi \in \mathbf{\Phi}} |\hat{P}(\phi \pm \epsilon_n \Pi_n \phi_j^*) - \hat{P}(\phi)| = o_p(\epsilon_n^2)$.*

Assumption 3.3.7. *β_0 is in the interior of $\mathbf{B}$ and ϕ_P is in the interior of $\mathbf{\Phi}$ with respect to $\|\cdot\|_\infty$.*

Assumption 3.3.5 requires ϕ_j^* to lie in $\mathbf{\Phi}$ for all $j = 1, \ldots, d_x$, and the polynomial rate of the series approximation error for $\mathrm{E}[X - \Phi^*(Z)|W = w]\sigma^{-2}(w)$ as in Assumptions 3.3.4(v) and (vi). In particular, Assumption 3.3.5(i) is only required for one Φ^* if it is not unique. Assumption 3.3.6 further restricts the weight and penalty estimators. Specifically, Assumption 3.3.6(i) requires a sufficiently fast convergence rate of the weight estimator while Assumption 3.3.6(ii) requires uniform continuity of the penalty estimator, which is satisfied by all L^2-type penalties whenever $\lambda_n = o(n^{-1/2})$. In particular, a uniform convergence rate of the weight estimator (Ai and Chen, 2003; Chen and Pouzo, 2009) is not

required. Assumption 3.3.7 imposes the standard assumption that the true and the selected parameters are in the interior of the parameter spaces.

Remark 3.3.2. Assumption 3.3.6(ii) ensures that the estimation error of the penalty term is not involved in the asymptotic distribution of $\hat{\beta}_P$. It requires that $\lambda_n = o(n^{-1/2})$. Besides in the study of general conditional moment models (Chen and Pouzo, 2009), similar assumption can be found in the study of linear LASSO regression (Knight and Fu, 2000), GMM LASSO estimation (Caner, 2009) and GMM estimation with moment selection (Liao, 2013; Cheng and Liao, 2015).

The following theorem confirms the asymptotic normality of $\hat{\beta}_P$.

Theorem 3.3.3. *Suppose Assumptions 3.3.1-3.3.7 hold. Let* $(\hat{\beta}_P, \hat{\phi}_P)$ *be the estimator in* (3.7). *If* $J_n + d_x \leq k_n$, $0 < \lambda_n$, $\max\{\frac{k_n}{n}, k_n^{-2\alpha_w}, J_n^{-2\alpha_z}\} = o(\lambda_n)$ *and* $\lambda_n = o(n^{-1/2})$, *then*

$$\sqrt{n}(\hat{\beta}_P - \beta_0) \xrightarrow{L} N(0, V_P),$$

where $V_P \equiv \Gamma^{-1}\Sigma_P\Gamma^{-1}$ *with*

$$\Sigma_P \equiv \mathrm{E}[\mathrm{E}[X - \Phi^*(Z)|W](\mathrm{E}[X - \Phi^*(Z)|W])'\sigma^{-4}(W)\sigma_P^2(W)]$$

and

$$\sigma_P^2(W) \equiv \mathrm{E}[(\varepsilon + \phi_0(Z) - \phi_P(Z))^2|W].$$

Theorem 3.3.3 demonstrates that $\sqrt{n}(\hat{\beta}_P - \beta_0)$ is asymptotically normally distributed with mean zero and variance matrix V_P. In particular, the asymptotic normality of $\hat{\beta}_P$ holds regardless of whether ϕ_0 is point identified or not. When the completeness condition (3.2) holds, $\sqrt{n}(\hat{\beta}_P - \beta_0)$ has exactly the same asymptotic property as the estimators in Ai and Chen (2003) and Chen and Pouzo (2009). Leveraging their results in turn implies that V_P is the efficiency bound in the model with the completeness condition (3.2) if $\sigma^2(W) = \mathrm{E}[\varepsilon^2|W]$. When the completeness condition (3.2) does not hold, $\hat{\beta}_P$ still enjoys the usual property of being asymptotically normal.

In the present of nonparametric partial identification, V_P nevertheless may depend on the penalty $P(\cdot)$ in addition to the weight $\sigma^2(\cdot)$. Specifically, different $P(\cdot)$'s pin down a different ϕ_P in $\mathbf{\Phi}_0$, which gives arise to a different identification noise defined as the discrepancy between ϕ_0 and ϕ_P. In particular, the dependence is through $\sigma_P^2(W)$, which is conditional variance of the error term and the identification noise. Thus, different $P(\cdot)$'s offer different asymptotic normal distributions. This may be expected from (3.17) as different ϕ_P's give different asymptotically linear statistics, or influence functions (Bickel et al., 1993). It turns out that this can be explained by the finding that the model with nonparametric partial identification is not regular in the sense of Bickel et al. (1993), in particular, the slope coefficients are not continuous in the underlying distribution. This is illustrated in a simple linear IV model, see Lemma 3.9.9.

3.4 Optimal Estimation

Given the irregularity in the presence of nonparametric partial identification, classical efficiency analysis cannot be applied. However, minimizing V_P with respect to the weight $\sigma^2(\cdot)$ and the penalty $P(\cdot)$ can be pursued as follows. In particular, an optimal $\sigma^2(\cdot)$ exists to minimize V_P for a given $P(\cdot)$ as in the efficiency analysis for the model with the completeness condition (3.2) (Ai and Chen, 2003; Chen and Pouzo, 2009). Specifically, an optimal $\sigma^2(\cdot)$ is given by $\sigma_P^2(\cdot)$, see Lemma 3.9.10. So a feasible optimally weighted estimator can be developed as in Ai and Chen (2003) and Chen and Pouzo (2009), and gives an efficient estimator in the model with the completeness condition (3.2). Unfortunately, in general an optimal $P(\cdot)$ does not exist to minimize V_P for a given $\sigma^2(\cdot)$. In turn, we focus on the study of optimal $P(\cdot)$ in the sense of minimizing the asymptotic variance of a linear functional of the slope coefficients estimator for a given $\sigma^2(\cdot)$.

3.4.1 Optimal Penalty

For $0 \neq \lambda \in \mathbf{R}^{d_x}$, the asymptotic variance of $\sqrt{n}\lambda'(\hat{\beta}_P - \beta_0)$ is given by $\lambda' V_P \lambda$. Since $\lambda' V_P \lambda$ is strictly convex in ϕ_P, the natural penalty for minimizing $\lambda' V_P \lambda$ is given by $P_\lambda(\cdot)$ with

$$P_\lambda(\phi) = \lambda' \Gamma^{-1} \Sigma(\phi) \Gamma^{-1} \lambda, \tag{3.18}$$

where $\Sigma(\phi) \equiv \mathrm{E}[\mathrm{E}[X - \Phi^*(Z)|W](\mathrm{E}[X - \Phi^*(Z)|W])' \sigma^{-4}(W)(Y - X'\beta_0 - \phi(Z))^2]$. Note that $P_\lambda(\cdot)$ is strictly convex and continuous in $\mathbf{\Phi}$ with respect to $\|\cdot\|_\infty$, then there exists a unique minimizer of $P_\lambda(\cdot)$ over the convex and compact set $\mathbf{\Phi}_0$, see Lemma 3.4.1. Obviously, $P_\lambda(\cdot)$ offers a solution to minimizing $\lambda' V_P \lambda$. However, in the absence of nonparametric identification, the minimizer ϕ_{P_λ} is not necessarily equal to ϕ_0 and may depend on λ, so in general there does not exist a $P^*(\cdot)$ such that $V_P - V_{P^*}$ is positive semidefinite for all $P(\cdot)$ except some special cases. If $\sigma_P^2(W)$ does not depend on W or $\mathrm{E}[\varepsilon(\phi_0(Z) - \phi_P(Z))|W] = 0$ for all $\phi_P \in \mathbf{\Phi}_0$, then $V_P - V_{P_\lambda}$ is positive semidefinite for all $0 \neq \lambda \in \mathbf{R}^{d_x}$. For the latter, the optimal penalty would choose $\phi_{P_\lambda} = \phi_0$ and successfully identify ϕ_0.

Lemma 3.4.1. *Suppose Assumptions 3.3.1(i)-(iv), 3.3.3(i) and 3.3.5(i) hold. Let $\sigma^2(\cdot)$ be fixed and $P_\lambda(\cdot)$ given in* (3.18)*. Then $P(\cdot) = P_\lambda(\cdot)$ satisfies Assumption 3.3.3(iii) and the resulting estimator of $\lambda'\beta_0$ has the smallest asymptotic variance.*

Remark 3.4.1. Unfortunately, Lemma 3.4.1 does not offer an easy solution to minimize $\lambda' V_P \lambda$ globally with respect to both $P(\cdot)$ and $\sigma^2(\cdot)$. By Lemma 3.9.10, minimizing $\lambda' V_P \lambda$ evaluated at $\sigma^2(\cdot) = \sigma_P^2(\cdot)$ with respect to $P(\cdot)$ gives the global minimal. However, in general $\lambda' V_P \lambda$ evaluated at $\sigma^2(\cdot) = \sigma_P^2(\cdot)$ depends on $P(\cdot)$ in a complicated and intractable way. In particular, plugging $\sigma^2(\cdot) = \sigma_P^2(\cdot)$ into $\lambda' V_P \lambda$ yields the asymptotic variance $\lambda'(\mathrm{E}[\mathrm{E}[X - \Phi^*(Z)|W](\mathrm{E}[X - \Phi^*(Z)|W])' \sigma_P^{-2}(W)])^{-1}\lambda$, which is not strictly convex in ϕ_P since $\mathrm{E}[X - \Phi^*(Z)|W]$ may depend on ϕ_P and $\sigma_P^2(\cdot)$ appears as its inverse. Therefore, no straightforward penalty satisfies Assumption 3.3.3(iii).

3.4.2 Two-Step Procedure

To obtain a feasible optimally penalized estimator, we have to estimate $P_\lambda(\cdot)$. We only need to estimate Γ and $\Sigma(\cdot)$. Note that a consistent estimator of $\sigma^2(\cdot)$ is given *a prior*. The idea is to replace the expectation with the sample average, the conditional mean with the series estimator and the unknown parameters with their estimators.

To estimate Γ and $\Sigma(\cdot)$, we have to first estimate Φ^*. For $j = 1, \ldots, d_x$, let $\hat{\phi}_j^*$ solve

$$\min_{\phi \in \mathbf{\Phi}_{J_n}} \frac{1}{n} \sum_{i=1}^{n} [\hat{\mathrm{E}}[X^{(j)} - \phi(Z)|W_i]]^2 \hat{\sigma}^{-2}(W_i). \tag{3.19}$$

Let $\hat{\Phi}^* \equiv (\hat{\phi}_1^*, \ldots, \hat{\phi}_{d_x}^*)'$. Then Γ can be estimated by

$$\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mathrm{E}}[X - \hat{\Phi}^*(Z)|W_i](\hat{\mathrm{E}}[X - \hat{\Phi}^*(Z)|W_i])' \hat{\sigma}^{-2}(W_i), \tag{3.20}$$

and $\Sigma(\cdot)$ can be estimated by $\hat{\Sigma}(\cdot)$ with

$$\hat{\Sigma}(\phi) = \frac{1}{n} \sum_{i=1}^{n} \hat{\mathrm{E}}[X - \hat{\Phi}^*(Z)|W_i](\hat{\mathrm{E}}[X - \hat{\Phi}^*(Z)|W_i])' \hat{\sigma}^{-4}(W_i)(Y_i - X_i'\hat{\beta}_1 - \phi(Z_i))^2, \tag{3.21}$$

where $\hat{\beta}_1$ is an initial estimator of β_0. Thus, $P_\lambda(\cdot)$ can be estimated by $\hat{P}_\lambda(\cdot)$ with

$$\hat{P}_\lambda(\phi) = \lambda' \hat{\Gamma}^{-1} \hat{\Sigma}(\phi) \hat{\Gamma}^{-1} \lambda. \tag{3.22}$$

To establish the consistency of $\hat{P}_\lambda(\cdot)$, we only need consistency of $\hat{\mathrm{E}}[X - \hat{\Phi}^*(Z)|W = \cdot]$ to $\mathrm{E}[X - \Phi^*(Z)|W = \cdot]$ that is guaranteed, though $\hat{\Phi}^*$ may not be consistent for Φ^* without the completeness condition (3.2).[3]

To establish the consistency of $\hat{P}_\lambda(\cdot)$, we establish the following proposition.

Proposition 3.4.1. *Suppose Assumptions 3.3.1(i)-(iv), 3.3.2, 3.3.3(i), 3.3.4 and 3.3.5(i) hold and* $\|\hat{\beta}_1 - \beta_0\| = o_p(1)$. *Let* $\hat{\Gamma}$ *and* $\hat{\Sigma}(\cdot)$ *be the estimators in* (3.20) *and* (3.21), *respec-*

[3] Alternatively, a penalty term can be attached to the objective function in (3.19) to ensure the consistency of $\hat{\Phi}^*$ by following the method in Section 3.3.1.

tively. If $J_n \leq k_n$ *and* $\max\{\frac{k_n}{n}, k_n^{-2\alpha_w}, J_n^{-2\alpha_z}\} = o(1)$*, then*

$$\sup_{\phi\in\mathbf{\Phi}} \|\hat{\Gamma}^{-1}\hat{\Sigma}(\phi)\hat{\Gamma}^{-1} - \Gamma^{-1}\Sigma(\phi)\Gamma^{-1}\| = o_p(1).$$

Proposition 3.4.1 implies that $\sup_{\phi\in\mathbf{\Phi}} \|\hat{P}_\lambda(\phi) - P_\lambda(\phi)\| = o_p(1)$, and so $\hat{P}_\lambda(\cdot)$ satisfies Assumption 3.3.3(ii) for $P_\lambda(\cdot)$. In addition, for any $\epsilon_n = o(n^{-1/2})$, we have $\sup_{\phi\in\mathbf{\Phi}} |\hat{P}_\lambda(\phi \pm \epsilon_n \Pi_n \phi_j^*) - \hat{P}_\lambda(\phi)| = O_p(\epsilon_n)$ by Lemma 3.9.11, Assumptions 3.3.1(i), (ii), (iv), 3.3.3(i) and 3.3.5(i). It implies that $P_\lambda(\cdot)$ satisfies Assumption 3.3.6(ii) whenever $\lambda_n = o(n^{-1/2})$. It together with Lemma 3.4.1 yields the following corollary.

Corollary 3.4.1. *Suppose the conditions of Theorem 3.3.3 hold and* $\|\hat{\beta}_1 - \beta_0\| = o_p(1)$*. Let* $\sigma^2(\cdot)$ *be fixed. For* $0 \neq \lambda \in \mathbf{R}^{d_x}$*, let* $\hat{P}_\lambda(\cdot)$ *be the estimator in* (3.22) *and* $(\hat{\beta}_P, \hat{\phi}_P)$ *be the estimator in* (3.7)*. Then* $\sqrt{n}\lambda'(\hat{\beta}_{P_\lambda} - \beta_0)$ *has no larger asymptotic variance than* $\sqrt{n}\lambda'(\hat{\beta}_P - \beta_0)$ *for any* $P(\cdot)$*.*

Corollary 3.4.1 implies that employing the penalty $\hat{P}_\lambda(\cdot)$ leads to the feasible optimal penalized estimator $\lambda'\hat{\beta}_{P_\lambda}$ with minimum asymptotic variance for a given $\sigma^2(\cdot)$. To implement the feasible optimally penalized estimator, an initial consistent estimator $\hat{\beta}_1$ is needed. As only consistency is required, $\hat{\beta}_1$ can be the SMD estimator, the PSMD estimator, or our penalized estimator. The feasible two-step procedure is summarized as follows.

1. For the identity weight, compute $\hat{\beta}_1$ as the SMD estimator, the PSMD estimator, or the estimator in (3.7) for $P(\cdot)$ being any L^2-type penalty.

2. For a given $\sigma^2(\cdot)$, compute $\hat{P}_\lambda(\cdot)$ according to (3.22) for $0 \neq \lambda \in \mathbf{R}^{d_x}$. For the same $\sigma^2(\cdot)$, compute $(\hat{\beta}_{P_\lambda}, \hat{\phi}_{P_\lambda})$ according to (3.7).

It is worth pointing out that the feasible penalized estimator only requires a consistent penalty estimator (i.e., Assumption 3.3.3(ii)), which is in contrast to the sufficiently fast convergence rate requirement on the weight estimator (i.e., Assumption 3.3.6(i)) for the optimally weighted estimator (Ai and Chen, 2003; Chen and Pouzo, 2009). In fact, the

optimally penalized estimator only requires an initial consistent estimator of β_0, whereas the optimally weighted estimator requires an initial estimator of (β_0, ϕ_P) with sufficiently fast convergence rate. Thus, the feasibility of the optimally penalized estimator is easier to achieve. Note that if $\sigma^2(\cdot) = \sigma_P^2(\cdot)$ in Step 2 for a given $P(\cdot)$, then our optimally penalized estimator not only delivers an efficient estimator for β_0 when ϕ_0 is identified but also reduces the asymptotic variance of the estimator of $\lambda'\beta_0$ when ϕ_0 is only partially identified. So our procedure can provide an estimator that is as efficient as those in Ai and Chen (2003) and Chen and Pouzo (2009) when ϕ_0 is identified, and an estimator with asymptotic normality and locally minimized variance for $\lambda'\beta_0$ when ϕ_0 is only partially identified.

3.5 Variance Estimation

For the purpose of inference, a consistent estimation of V_P is needed. The natural estimator for V_P is given by $\hat{V}_P \equiv \hat{\Gamma}^{-1}\hat{\Sigma}(\hat{\phi}_P)\hat{\Gamma}^{-1}$ with $\hat{\beta}_1 = \hat{\beta}_P$. Given Theorem 3.3.1 and Proposition 3.4.1, the consistency of $\hat{V}_P$ immediately follows by the continuous mapping theorem. This is given in the following corollary.

Corollary 3.5.1. *Suppose the conditions of Theorem 3.3.1 and Assumption 3.3.5(i) hold. Let* $(\hat{\beta}_P, \hat{\phi}_P)$ *be the estimator in* (3.7)*, and* $\hat{\Gamma}$ *and* $\hat{\Sigma}(\cdot)$ *be the estimators in* (3.20) *and* (3.21) *with* $\hat{\beta}_1 = \hat{\beta}_P$*.. Then* $\|\hat{V}_P - V_P\| = o_p(1)$*.*

Corollary 3.5.1 implies $\hat{V}_P^{-1/2}\sqrt{n}(\hat{\beta}_P - \beta_0) \xrightarrow{L} N(0, I_{d_x})$ provided Σ_P is positive definite, which holds if and only if $\sigma_P^2(\cdot)$ is nonzero. This result can be used for hypothesis testing and confidence set construction for β_0. Note that the variance matrix estimate is directly available in the two-step procedure for the optimally penalized estimator, so there is no extra computation cost for obtaining the variance matrix estimate.

3.6 Simulation Studies

In this section, we conduct small-scale Monte Carlo simulations to examine finite sample performance of our penalized estimators and show how the distribution of the SMD estimator in Ai and Chen (2003) deviates from normality.

We assume that $(X, Z, W) \in [0,1]^3$ are generated according to the density

$$f_{X,Z,W}(x,z,w) = \frac{2}{3}(x+z+w) \text{ for } (x,z,w) \in [0,1]^3. \tag{3.23}$$

By construction, $\mathrm{E}[6Z^2 - 6Z + 1|X, W] = 0$ which implies the completeness condition (3.2) is violated by $f_{X,Z,W}$. Let U be uniformly distributed on $[0,1]$ independent of (X, Z, W), $\beta_0 = 1$, $\phi_0(Z) = Z^2/2$ and Y be generated according to the relationship

$$Y = X\beta_0 + \phi_0(Z) + \varepsilon \text{ with } \varepsilon = \frac{U}{12}\left(\frac{1}{f_{Z|X,W}(Z|X,W)} - 1\right). \tag{3.24}$$

where $f_{Z|X,W}$ is the conditional density of Z given X and W. By construction, $\mathrm{E}[\varepsilon|X, W] = 0$ and $\mathrm{E}[\varepsilon|Z] = (1 - f_Z(Z))/(24 f_Z(Z))$. So (X, W) are exogenous while Z is endogenous. Note that $\mathrm{E}[X|X, W] = X$ and $\mathrm{E}[\phi(Z)|X, W]$ is a function of both X and W for any $\phi \in L^2(Z)$ by symmetry of $f_{X,Z,W}$, so Assumption 3.3.1(v) is satisfied.

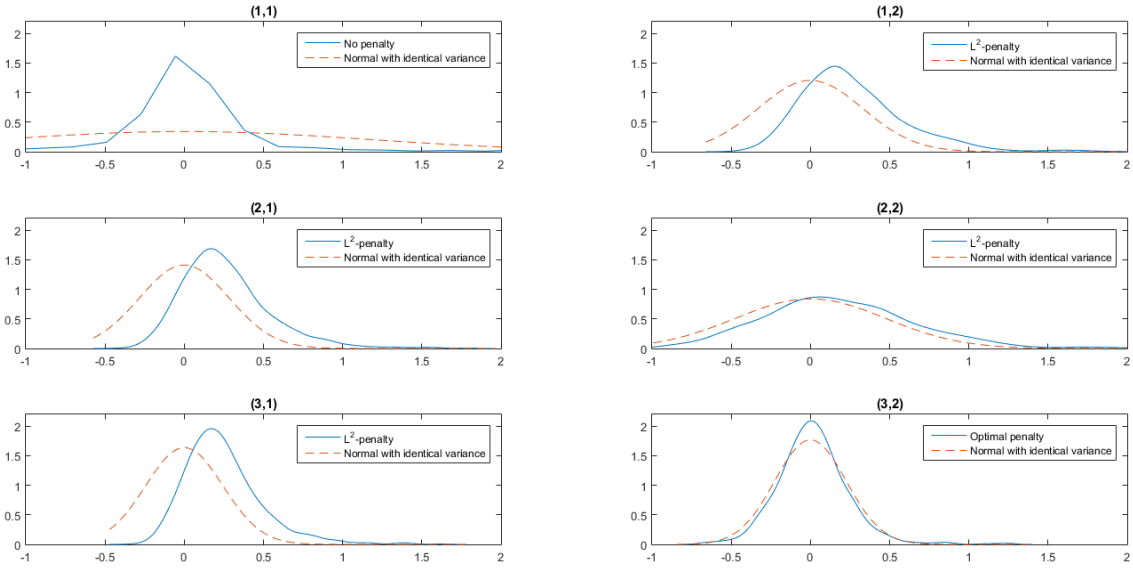
We employ $q_j(z) = z^{j-1}$ for $j = 1, 2, \ldots$ to construct $\mathbf{\Phi}_{J_n}$ and $p_1(x, w) = x$ and $\{p_k(x, w)\}_{k=2}^{\infty}$ being orthonormal polynomials of w for series estimators.[4] In constructing $\mathbf{\Phi}_{J_n}$, we impose $\|\phi\|_{\infty,2} \leq 100$. We let $\sigma^2(\cdot) = 1$. To investigate the effect of the penalty, we implement our penalized estimation for four L^2-type penalties, which differ in the assigned probability measure, and the optimal penalty in Corollary 3.4.1, which is denoted as $P_1(\cdot)$. Specifically, the four L^2-type penalties are given by $P(\phi) = \int_0^1 \phi^2(z)\nu(z)dz$, with $\nu(z) = f_Z(z), 1, 2z, 1/((1+z)\ln 2)$ being density functions. We use the SMD estimator as the initial estimator in our optimal penalized estimation. It is easy to show that all rate requirements

[4]The results for $\{q_j\}_{j=1}^{\infty}$ being splines of order 4 with equally spaced knots and $\{p_k\}_{k=1}^{\infty}$ being splines of order 2 with equally spaced knots are similar and are available upon request.

are satisfied by letting (k_n, J_n, λ_n) satisfy the conditions of Theorem 3.3.3. However, the theory offers little guidance as to how to select (k_n, J_n, λ_n). For simplicity, we let $k_n = \underline{n^{\frac{1}{5}}}+1$ and $J_n = k_n - 1$ and $\lambda_n = n^{-\frac{2}{3}}$, $n^{-\frac{4}{5}}$ and $n^{-\frac{8}{9}}$ to investigate the sensitivity of the results with respect to λ_n. The results are reported for $n = 100, 500, 1000$ and 2000 based on 1000 Monte Carlo replications. Figures 3.1-3.4 show the distribution of the SMD estimator and our penalized estimators.[5] For each case, the normal distribution with mean zero and identical variance is plotted for comparison. Tables 3.1 and 3.2 report the bias and standard deviation of our penalized estimators.

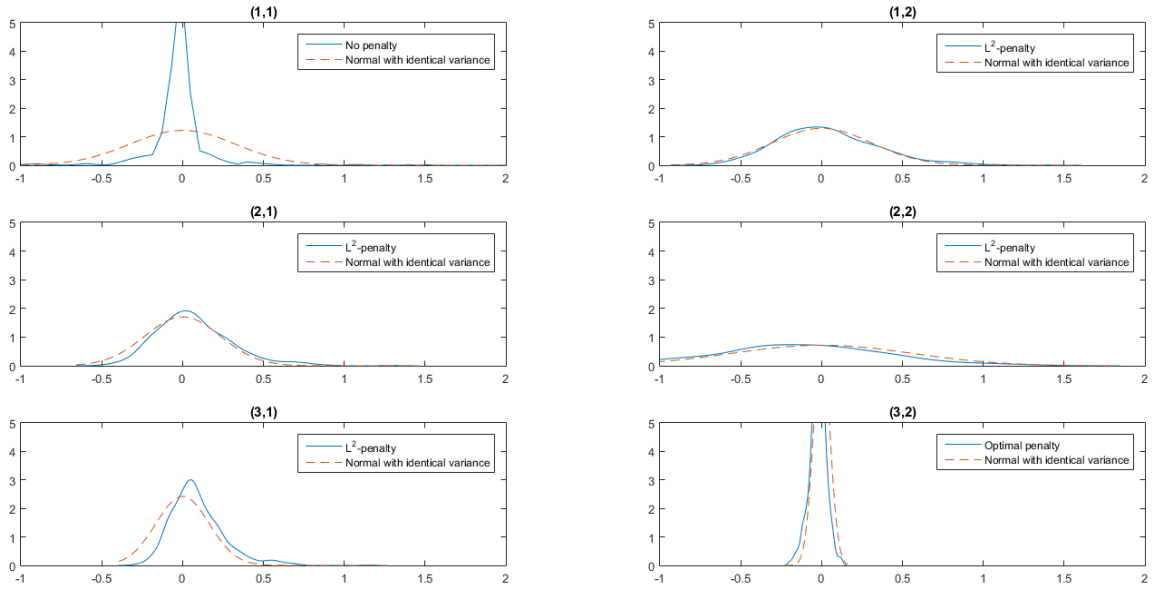
The top left graph in Figures 3.1-3.4 shows that the SMD estimator does not exhibit normality even for $n = 2000$. In particular, the shape of the graph is similar to shape of the graph in Figure 1 of Choi and Phillips (1992), so the distribution of the SMD estimator appears to be a mixture of normals. It suggests the necessity of penalization. The remaining graphs in Figures 3.1-3.4 show that the distribution of our penalized estimators with $\lambda_n = n^{-4/5}$ is close to normal for all n's and $P(\cdot)$. The results for $\lambda_n = n^{-2/3}$ and $\lambda_n = n^{-8/9}$ are similar and are available upon request. Table 3.1 implies that the bias of our penalized estimators is sensitive to the choice of λ_n when $n = 100$ and $n = 500$, and the sensitivity becomes less severe as n increases uniformly over all $P(\cdot)$. In particular, the bias is small and close to zero when $n = 2000$. Table 3.2 shows that the standard deviation of $\sqrt{n}(\hat{\beta}_P - \beta_0)$ is not sensitive to the choice of λ_n, but significantly sensitive to the choice of $P(\cdot)$. It confirms the theoretical finding in Theorem 3.3.3 and suggests the necessity of optimal penalization. The last row of Table 3.2 shows that using $P_1(\cdot)$ yields smaller variances than using the other four L^2 penalties for all cases except for $n = 100$. Overall, the performance of our penalized and optimally penalized estimators is encouraging in finite samples.

[5]The scale of y-axis in Figure 3.1 is different from those in Figures 3.2-3.4.



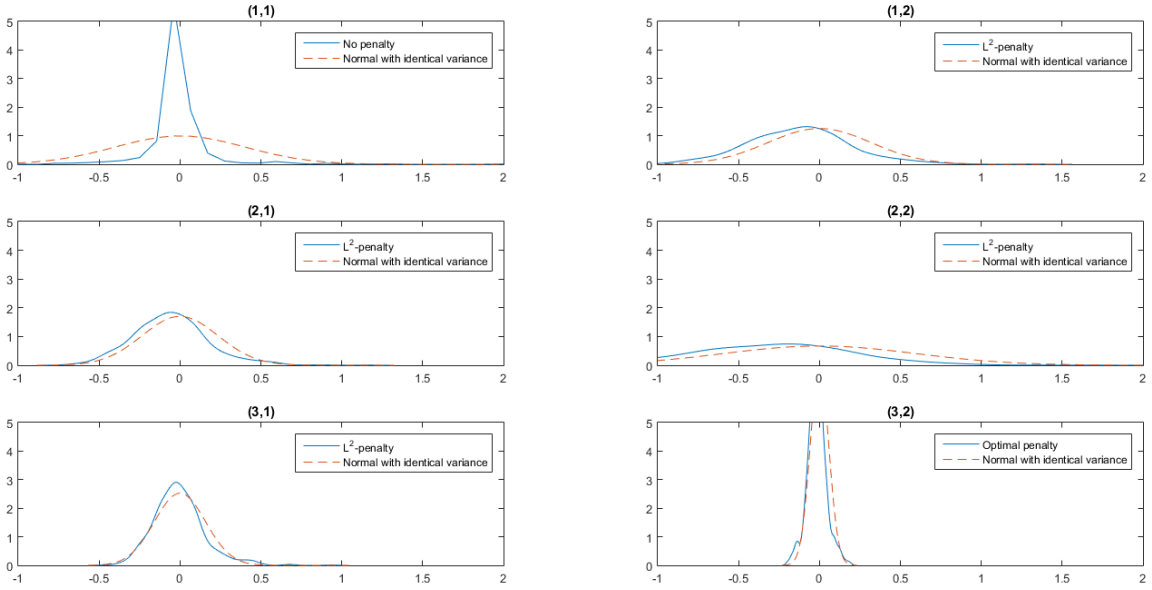
Plot (1,1) is for the SMD estimator (i.e., $P(\cdot) = 0$) while plots (1,2)-(3,2) are for our penalized estimators, whose order is the same as in Tables 3.1 and 3.2.

Figure 3.1: The distribution of $\sqrt{n}(\hat{\beta}_P - \beta_0)$ for various $P(\cdot)$ when $\lambda_n = n^{-4/5}$ and $n = 100$



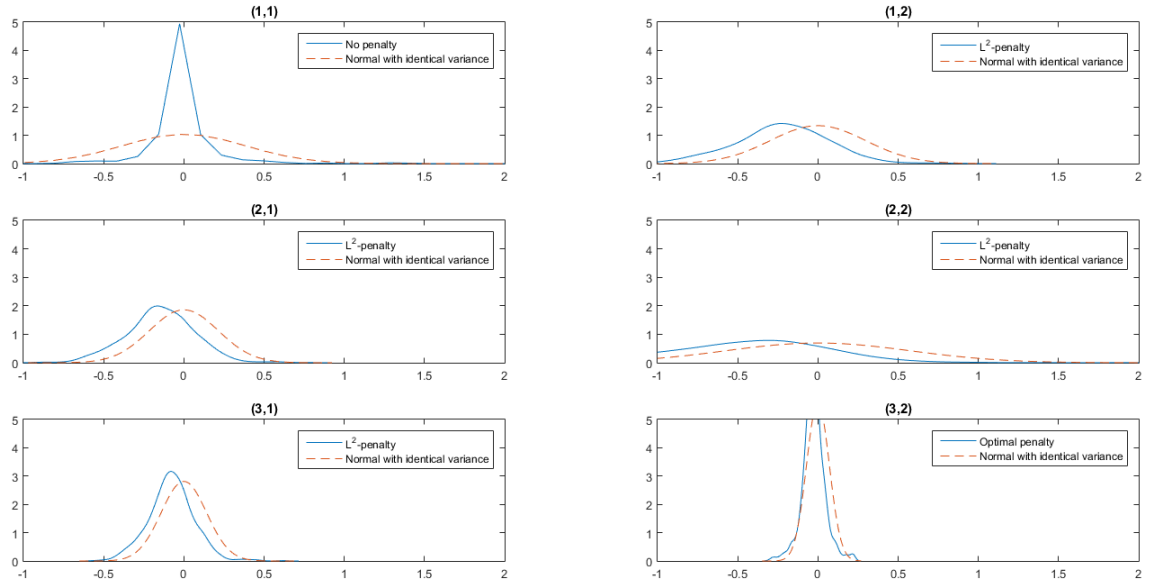
Plot (1,1) is for the SMD estimator (i.e., $P(\cdot) = 0$) while plots (1,2)-(3,2) are for our penalized estimators, whose order is the same as in Tables 3.1 and 3.2.

Figure 3.2: The distribution of $\sqrt{n}(\hat{\beta}_P - \beta_0)$ for various $P(\cdot)$ when $\lambda_n = n^{-4/5}$ and $n = 500$



Plot (1,1) is for the SMD estimator (i.e., $P(\cdot) = 0$) while plots (1,2)-(3,2) are for our penalized estimators, whose order is the same as in Tables 3.1 and 3.2.

Figure 3.3: The distribution of $\sqrt{n}(\hat{\beta}_P - \beta_0)$ for various $P(\cdot)$ when $\lambda_n = n^{-4/5}$ and $n = 1000$



Plot (1,1) is for the SMD estimator (i.e., $P(\cdot) = 0$) while plots (1,2)-(3,2) are for our penalized estimators, whose order is the same as in Tables 3.1 and 3.2.

Figure 3.4: The distribution of $\sqrt{n}(\hat{\beta}_P - \beta_0)$ for various $P(\cdot)$ when $\lambda_n = n^{-4/5}$ and $n = 2000$

Table 3.1: The bias of $\hat{\beta}_P$ for various $P(\cdot)$, λ_n and n

	$n=100$			$n=500$		
	$\lambda_n = n^{-\frac{2}{3}}$	$\lambda_n = n^{-\frac{4}{5}}$	$\lambda_n = n^{-\frac{8}{9}}$	$\lambda_n = n^{-\frac{2}{3}}$	$\lambda_n = n^{-\frac{4}{5}}$	$\lambda_n = n^{-\frac{8}{9}}$
$P(\phi) = \mathrm{E}[\phi^2(Z)]$	0.0468	0.0267	0.0178	0.0113	0.0016	−0.0013
$P(\phi) = \int_0^1 \phi^2(z)dz$	0.0454	0.0263	0.0178	0.0125	0.0031	0.0002
$P(\phi) = \int_0^1 \phi^2(z)2zdz$	0.0366	0.0179	0.0099	−0.0003	−0.0067	−0.0079
$P(\phi) = \int_0^1 \frac{\phi^2(z)}{(1+z)\ln 2}dz$	0.0418	0.0244	0.0167	0.0126	0.0041	0.0013
$P(\phi) = P_1(\phi)$	0.0026	0.0030	0.0033	−0.0011	−0.0011	−0.0011
	$n=1000$			$n=2000$		
	$\lambda_n = n^{-\frac{2}{3}}$	$\lambda_n = n^{-\frac{4}{5}}$	$\lambda_n = n^{-\frac{8}{9}}$	$\lambda_n = n^{-\frac{2}{3}}$	$\lambda_n = n^{-\frac{4}{5}}$	$\lambda_n = n^{-\frac{8}{9}}$
$P(\phi) = \mathrm{E}[\phi^2(Z)]$	0.0017	−0.0040	−0.0049	−0.0024	−0.0051	−0.0048
$P(\phi) = \int_0^1 \phi^2(z)dz$	0.0039	−0.0018	−0.0030	−0.0002	−0.0033	−0.0034
$P(\phi) = \int_0^1 \phi^2(z)2zdz$	−0.0099	−0.0119	−0.0110	−0.0125	−0.0110	−0.0088
$P(\phi) = \int_0^1 \frac{\phi^2(z}{(1+z)\ln 2}dz$	0.0052	−0.0002	−0.0015	0.0014	−0.0017	−0.0021
$P(\phi) = P_1(\phi)$	−0.0007	−0.0005	−0.0005	−0.0007	−0.0006	−0.0006

Table 3.2: The standard deviation of $\sqrt{n}\hat{\beta}_P$ for various $P(\cdot)$, λ_n and n

	$n=100$			$n=500$		
	$\lambda_n = n^{-\frac{2}{3}}$	$\lambda_n = n^{-\frac{4}{5}}$	$\lambda_n = n^{-\frac{8}{9}}$	$\lambda_n = n^{-\frac{2}{3}}$	$\lambda_n = n^{-\frac{4}{5}}$	$\lambda_n = n^{-\frac{8}{9}}$
$P(\phi) = \mathrm{E}[\phi^2(Z)]$	0.3360	0.3306	0.3224	0.3358	0.3075	0.2848
$P(\phi) = \int_0^1 \phi^2(z)dz$	0.2933	0.2824	0.2736	0.2627	0.2335	0.2131
$P(\phi) = \int_0^1 \phi^2(z)2zdz$	0.4697	0.4752	0.4667	0.5940	0.5557	0.5197
$P(\phi) = \int_0^1 \frac{\phi^2(z)}{(1+z)\ln 2}dz$	0.2582	0.2435	0.2346	0.1927	0.1646	0.1475
$P(\phi) = P_1(\phi)$	0.3178	0.3829	0.4396	0.0450	0.0597	0.0719
	$n=1000$			$n=2000$		
	$\lambda_n = n^{-\frac{2}{3}}$	$\lambda_n = n^{-\frac{4}{5}}$	$\lambda_n = n^{-\frac{8}{9}}$	$\lambda_n = n^{-\frac{2}{3}}$	$\lambda_n = n^{-\frac{4}{5}}$	$\lambda_n = n^{-\frac{8}{9}}$
$P(\phi) = \mathrm{E}[\phi^2(Z)]$	0.3511	0.3188	0.2948	0.3360	0.2973	0.2662
$P(\phi) = \int_0^1 \phi^2(z)dz$	0.2654	0.2345	0.2136	0.2482	0.2147	0.1895
$P(\phi) = \int_0^1 \phi^2(z)2zdz$	0.6449	0.5951	0.5552	0.6578	0.5754	0.5063
$P(\phi) = \int_0^1 \frac{\phi^2(z)}{(1+z)\ln 2}dz$	0.1852	0.1579	0.1416	0.1696	0.1418	0.1237
$P(\phi) = P_1(\phi)$	0.0499	0.0729	0.0948	0.0572	0.0911	0.0988

3.7 Conclusion

In this paper, we developed $\sqrt{n}$ asymptotically normal estimators for the slope coefficients in the PLIV model, which are robust to a possible lack of nonparametric identification. Since the model is not regular in the presence of nonparametric partial identification, we then developed a feasible two-step optimally penalized estimator with minimum asymptotic variance for a linear functional of the slope coefficients through employing an optimal penalty. In addition, we provided a consistent estimator for the asymptotic variance matrix. Monte Carlos simulations demonstrated good finite sample performance of our penalized estimators. Despite the focus on the PLIV model in this paper, the results can be extended to general semiparametric conditional moment restriction models following the same technique as long as sufficient structural information on the conditional moment function is imposed.

3.8 Acknowledgement

Chapter 3 "Robust and Optimal Estimation for Partially Linear Instrumental Variables Models with Partial Identification," in part, is currently being prepared for submission for publication of the material. Chen, Qihui. The dissertation author was the primary investigator and author of this material.

3.9 Appendix

3.9.1 Proofs of Main Results

For ease of reference, Table 3.3 presents simplified notation for random variables, conditional means and series estimators, which will be used throughout the appendix. Table 3.4 collects the sequences utilized in the text and the location of their introduction.

Table 3.3: List of simplified notation

$\rho(\beta, \phi, Y, X, Z)$	The random variable $Y - X'\beta - \phi(Z)$ for $\beta \in \mathbf{B}$ and $\phi \in \mathbf{\Phi}$.
$m(\beta, \phi, W)$	The conditional mean $\mathrm{E}[\rho(\beta, \phi, Y, X, Z)\|W]$ for $\beta \in \mathbf{B}$ and $\phi \in \mathbf{\Phi}$.
$\hat{m}(\beta, \phi, W)$	The series estimator $\hat{\mathrm{E}}[\rho(\beta, \phi, Y, X, Z)\|W]$ for $\beta \in \mathbf{B}$ and $\phi \in \mathbf{\Phi}$.
$h_j(\phi, X, Z)$	The random variable $X^{(j)} - \phi(Z)$ for $\phi \in \mathbf{\Phi}$, $j = 1, \ldots, d_x$.
$g_j(\phi, W)$	The conditional mean $\mathrm{E}[h_j(\phi, X, Z)\|W]$ for $\phi \in \mathbf{\Phi}$, $j = 1, \ldots, d_x$.
$\hat{g}_j(\phi, W)$	The series estimator $\hat{\mathrm{E}}[h_j(\phi, X, Z)\|W]$ for $\phi \in \mathbf{\Phi}$, $j = 1, \ldots, d_x$.

Table 3.4: List of sequences

λ_n	The penalty tuning parameter. See Display (3.6).
J_n	The dimension of the sieve space. See Display (3.8).
k_n	The number of sieves for the series estimator. See Display (3.9).
$J_n^{-\alpha_z}$	The rate of the sieve approximation error. See Assumption 3.3.2(ii).
$k_n^{-\alpha_w}$	The rate of the series approximation error. See Assumption 3.3.4(v).
$\delta_{P,n}$	The convergence rate of the penalty estimator. See Assumption 3.3.3(ii).
$\delta_{\delta,n}$	The convergence rate of the weight estimator. See Assumption 3.3.6(i).
$\delta_{m,n}$	The convergence rate of the series estimator. See Lemma 3.9.3.
ξ_n	The length of sieves for the series estimator. See Assumption 3.3.4(iii).
ϵ_n	The nuisance sequence for the penalty estimator. See Assumption 3.3.6(ii).

Proof of Theorem 3.3.1: For every $\epsilon > 0$, let $A(\epsilon) \equiv \{(\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi} : \|\beta - \beta_0\| + \|\phi - \phi_P\|_\infty \geq \epsilon\}$ and $A_n(\epsilon) \equiv \{(\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi}_{J_n} : \|\beta - \beta_0\| + \|\phi - \phi_P\|_\infty \geq \epsilon\}$. By the definition of $(\hat{\beta}_P, \hat{\phi}_P)$ in (3.7), we have for any $\epsilon > 0$

$$\mathrm{P}(\|\hat{\beta}_P - \beta_0\| + \|\hat{\phi}_P - \phi_P\|_\infty \geq \epsilon) \leq \mathrm{P}(\inf_{(\beta,\phi)\in A_n(\epsilon)} \frac{1}{n}\sum_{i=1}^n \hat{m}^2(\beta, \phi, W_i)\hat{\sigma}^{-2}(W_i) + \lambda_n \hat{P}(\phi) \leq \frac{1}{n}\sum_{i=1}^n \hat{m}^2(\beta_0, \Pi_n\phi_P, W_i)\hat{\sigma}^{-2}(W_i) + \lambda_n \hat{P}(\Pi_n\phi_P)). \tag{3.25}$$

By Lemma 3.9.3 and Assumption 3.3.3(ii), result (3.25) implies for any $\epsilon > 0$

$$\mathrm{P}(\|\hat{\beta}_P - \beta_0\| + \|\hat{\phi}_P - \phi_P\|_\infty \geq \epsilon) \leq \mathrm{P}(\min_{(\beta,\phi)\in A_n(\epsilon)} C\,\mathrm{E}[m^2(\beta, \phi, W)\sigma^{-2}(W)] + \lambda_n P(\phi) \leq C'\,\mathrm{E}[m^2(\beta_0, \Pi_n\phi_P, W)\sigma^{-2}(W)] + \lambda_n P(\Pi_n\phi_P) + o_p(\lambda_n) + O_p(\delta_{m,n}^2)). \tag{3.26}$$

Recalling that $m(\beta_0, \Pi_n\phi_P, W) = \mathrm{E}[\phi_P(Z) - \Pi_n\phi_P(Z)|W]$, we have

$$\begin{aligned}\mathrm{E}[m^2(\beta_0, \Pi_n\phi_P, W)\sigma^{-2}(W)] &= \mathrm{E}[(\mathrm{E}[\phi_P(Z) - \Pi_n\phi_P(Z)|W])^2\sigma^{-2}(W)] \\ &\leq \|\Pi_n\phi_P - \phi_P\|_\infty^2 \leq O(J_n^{-2\alpha_z}),\end{aligned} \tag{3.27}$$

where the first inequality follows by Assumption 3.3.3(i) and the second by Assumption 3.3.2(ii). Since $A_n(\epsilon) \subset A(\epsilon)$ for any $\epsilon > 0$ by Assumption 3.3.2 and $|P(\Pi_n\phi_P) - P(\phi_P)| = o(1)$ by Assumptions 3.3.2 and 3.3.3(iii), results (3.26) and (3.27) imply for any $\epsilon > 0$

$$\begin{aligned}\mathrm{P}(\|\hat{\beta}_P - \beta_0\| + \|\hat{\phi}_P - \phi_P\|_\infty \geq \epsilon) \leq \mathrm{P}(&\min_{(\beta,\phi)\in A(\epsilon)} C\,\mathrm{E}[m^2(\beta, \phi, W)\sigma^{-2}(W)] \\ &+ \lambda_n P(\phi) \leq O(J_n^{-2\alpha_z}) + \lambda_n P(\phi_P) + o_p(\lambda_n) + O_p(\delta_{m,n}^2)).\end{aligned} \tag{3.28}$$

Recall that $\delta_{n,m} = \max\{\sqrt{k_n/n}, k_n^{-\alpha_w}\}$ and $\max\{k_n/n, k_n^{-2\alpha_w}, J_n^{-2\alpha_z}\} = o(\lambda_n)$, so result (3.28) implies for any $\epsilon > 0$

$$\begin{aligned}\mathrm{P}(\|\hat{\beta}_P - \beta_0\| + \|\hat{\phi}_P - \phi_P\|_\infty \geq \epsilon) \leq \mathrm{P}(&\min_{(\beta,\phi)\in A(\epsilon)} C\,\mathrm{E}[m^2(\beta, \phi, W)\sigma^{-2}(W)] \\ &+ \lambda_n P(\phi) \leq \lambda_n P(\phi_P) + o_p(\lambda_n)).\end{aligned} \tag{3.29}$$

Note that $A(\epsilon)$ is compact with respect to $\|\cdot\| + \|\cdot\|_\infty$ by Assumptions 3.3.1(ii) and (iii) and $C\,\mathrm{E}[m^2(\beta, \phi, W)\sigma^{-2}(W)] + \lambda_n P(\phi)$ is continuous in (β, ϕ) with respect to $\|\cdot\| + \|\cdot\|_\infty$ by Lemma 3.9.4 and Assumption 3.3.3(iii), so there exists $(\beta_{\epsilon,n}, \phi_{\epsilon,n}) \in A(\epsilon)$ such that

$$\begin{aligned}&C\,\mathrm{E}[m^2(\beta_{\epsilon,n}, \phi_{\epsilon,n}, W)\sigma^{-2}(W)] + \lambda_n P(\phi_{\epsilon,n}) \\ &\qquad = \min_{(\beta,\phi)\in A(\epsilon)} C\,\mathrm{E}[m^2(\beta, \phi, W)\sigma^{-2}(W) + \lambda_n P(\phi).\end{aligned} \tag{3.30}$$

Thus, results (3.29) and (3.30) imply for any $\epsilon > 0$

$$\begin{aligned}\mathrm{P}(\|\hat{\beta}_P - \beta_0\| + \|\hat{\phi}_P - \phi_P\|_\infty \geq \epsilon) \leq \mathrm{P}(&C\,\mathrm{E}[m^2(\beta_{\epsilon,n}, \phi_{\epsilon,n}, W)\sigma^{-2}(W)] \\ &+ \lambda_n P(\phi_{\epsilon,n}) - \lambda_n P(\phi_P) \leq o_p(\lambda_n)).\end{aligned} \tag{3.31}$$

By the definition of lim inf, there exists a subsequence $\{n_k\}_{k=1}^{\infty}$ such that

$$\lim_{k\to\infty} \frac{C\,\mathrm{E}[m^2(\beta_{\epsilon,n_k},\phi_{\epsilon,n_k},W)\sigma^{-2}(W)] + \lambda_{n_k}P(\phi_{\epsilon,n_k}) - \lambda_{n_k}P(\phi_P)}{\lambda_{n_k}}$$
$$= \liminf_{n\to\infty} \frac{C\,\mathrm{E}[m^2(\beta_{\epsilon,n},\phi_{\epsilon,n},W)\sigma^{-2}(W)] + \lambda_n P(\phi_{\epsilon,n}) - \lambda_n P(\phi_P)}{\lambda_n}. \tag{3.32}$$

Since $A(\epsilon)$ is compact, it is without loss of generality to assume that $\{(\beta_{\epsilon,n},\phi_{\epsilon,n})\}_{k=1}^{\infty}$ is convergent with the limit, say $(\beta_\epsilon,\phi_\epsilon) \in A(\epsilon)$. Then it must be one of the cases: (i) $\mathrm{E}[m^2(\beta_\epsilon,\phi_\epsilon,W)\sigma^{-2}(W)] > 0$ or (ii) $\mathrm{E}[m^2(\beta_\epsilon,\phi_\epsilon,W)\sigma^{-2}(W)] = 0$. For case (ii), we have $m^2(\beta_\epsilon,\phi_\epsilon,W) = 0$ almost surely by Assumption 3.3.3(i), and thus $\beta_\epsilon = \beta_0$ and $\phi_\epsilon \in \mathbf{\Phi}_0$ by Assumption 3.3.1(v). This in turn implies that $\|\phi_\epsilon - \phi_P\|_\infty \geq \epsilon$ since $(\beta_\epsilon,\phi_\epsilon) \in A(\epsilon)$, and thus $P(\phi_\epsilon) > P(\phi_P)$ by Assumption 3.3.3(iii). For case (ii), we therefore have

$$\lim_{k\to\infty} \frac{C\,\mathrm{E}[m^2(\beta_{\epsilon,n_k},\phi_{\epsilon,n_k},W)\sigma^{-2}(W)] + \lambda_{n_k}P(\phi_{\epsilon,n_k}) - \lambda_{n_k}P(\phi_P)}{\lambda_{n_k}} > 0. \tag{3.33}$$

Note that (3.33) is obviously true for case (i) since $0 < \lambda_n = o(1)$. It combines with result (3.32) to yield

$$\liminf_{n\to\infty} \frac{C\,\mathrm{E}[m^2(\beta_{\epsilon,n},\phi_{\epsilon,n},W)\sigma^{-2}(W)] + \lambda_n P(\phi_{\epsilon,n}) - \lambda_n P(\phi_P)}{\lambda_n} > 0. \tag{3.34}$$

Combining results (3.31) and (3.34) yields for any $\epsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}(\|\hat{\beta}_P - \beta_0\| + \|\hat{\phi}_P - \phi_P\|_\infty \geq \epsilon) = 0, \tag{3.35}$$

which completes the proof of the theorem. ■

Proof of Theorem 3.3.2: By the definition of $(\hat{\beta}_P,\hat{\phi}_P)$ in (3.7), we have

$$\frac{1}{n}\sum_{i=1}^{n} \hat{m}^2(\hat{\beta}_P,\hat{\phi}_P,W_i)\hat{\sigma}^{-2}(W_i) + \lambda_n \hat{P}(\hat{\phi}_P)$$
$$\leq \frac{1}{n}\sum_{i=1}^{n} \hat{m}^2(\beta_0,\Pi_n\phi_P,W_i)\hat{\sigma}^{-2}(W_i) + \lambda_n \hat{P}(\Pi_n\phi_P)). \tag{3.36}$$

By Lemma 3.9.3 and Assumption 3.3.3(ii), result (3.36) implies

$$C\,\mathrm{E}[m^2(\hat{\beta}_P,\hat{\phi}_P,W)\sigma^{-2}(W)] + \lambda_n P(\hat{\phi}_P) \leq C'\,\mathrm{E}[m^2(\beta_0,\Pi_n\phi_P,W)\sigma^{-2}(W)] + \lambda_n P(\Pi_n\phi_P) + O_p(\delta_{P,n}\lambda_n) + O_p(\delta_{m,n}^2)). \tag{3.37}$$

By Theorem 3.3.1 and the continuous mapping theorem, there exists $\delta_{1,n} \downarrow 0$ such that $|P(\hat{\phi}_P) - P(\phi_P)| = O_p(\delta_{1,n})$. Let $\delta_{2,n} \equiv |P(\Pi_n\phi_P) - P(\phi_P)|$, then $\delta_{2,n} = o(1)$ by Assumptions 3.3.2 and 3.3.3(iii). Thus, results (3.27) and (3.37) imply

$$\mathrm{E}[m^2(\hat{\beta}_P,\hat{\phi}_P,W)\sigma^{-2}(W)] \leq O_p((\delta_{P,n}+\delta_{1,n}+\delta_{2,n})\lambda_n) + O_p(\delta_{m,n}^2 + J_n^{-2\alpha_z}). \tag{3.38}$$

Since $\delta_{P,n}+\delta_{1,n}+\delta_{2,n} = o(1)$ and $\delta_{m,n}^2+J_n^{-2\alpha_z} = o(\lambda_n)$ by recalling that $\delta_{n,m} = \max\{\sqrt{k_n/n}, k_n^{-\alpha_w}\}$, result (3.38) implies

$$\mathrm{E}[m^2(\hat{\beta}_P,\hat{\phi}_P,W)\sigma^{-2}(W)] \leq o_p(\lambda_n). \tag{3.39}$$

Note that $\|(\hat{\beta}_P,\hat{\phi}_P)-(\beta_0,\phi_P)\|_w^2 = \mathrm{E}[m^2(\hat{\beta}_P,\hat{\phi}_P,W)\sigma^{-2}(W)]$ by the definition of $\|\cdot\|_w$ in (3.11), so the result of the theorem follows by result (3.39). ■

Proof of Theorem 3.3.3: For $j = 1,\ldots,d_x$, let e_j be the jth column of the $d_x \times d_x$ identity matrix. By Theorem 3.3.1 and Assumptions 3.3.1(ii), 3.3.2, 3.3.5(i) and 3.3.7, we have $\hat{\beta}_P \pm \epsilon_n e_j \in \mathbf{B}$ and $\hat{\phi}_P \pm \epsilon_n\Pi_n\phi_j^* \in \mathbf{\Phi}_{J_n}$ for any $\epsilon_n = o(1)$ with probability approaching one. By the definition of $(\hat{\beta}_P,\hat{\phi}_P)$ in (3.7), it follows

$$\begin{aligned}
0 \leq{}& \frac{1}{2n}\sum_{i=1}^n \hat{m}^2(\hat{\beta}_P \mp \epsilon_n e_j, \hat{\phi}_P \pm \epsilon_n\Pi_n\phi_j^*, W_i)\hat{\sigma}^{-2}(W_i) + \frac{\lambda_n}{2}\hat{P}(\hat{\phi}_P \pm \epsilon_n\Pi_n\phi_j^*) \\
&- \frac{1}{2n}\sum_{i=1}^n \hat{m}^2(\hat{\beta}_P,\hat{\phi}_P,W_i)\hat{\sigma}^{-2}(W_i) - \frac{\lambda_n}{2}\hat{P}(\hat{\phi}_P) \\
={}& \frac{\pm\epsilon_n}{n}\sum_{i=1}^n \hat{g}_j(\Pi_n\phi_j^*,W_i)\hat{m}(\hat{\beta}_P,\hat{\phi}_P,W_i)\hat{\sigma}^{-2}(W_i) \\
&+ \frac{\epsilon_n^2}{2n}\sum_{i=1}^n \hat{g}_j^2(\Pi_n\phi_j^*,W_i)\hat{\sigma}^{-2}(W_i) + \frac{\lambda_n}{2}\hat{P}(\hat{\phi}_P \pm \epsilon_n\Pi_n\phi_j^*) - \frac{\lambda_n}{2}\hat{P}(\hat{\phi}_P)
\end{aligned} \tag{3.40}$$

for any $\epsilon_n = o(1)$ with probability approaching one. By Assumption 3.3.6(ii), there exists $\epsilon_n = o(n^{-1/2})$ such that

$$\frac{\lambda_n}{2}\hat{P}(\hat{\phi}_P \pm \epsilon_n \Pi_n \phi_j^*) - \frac{\lambda_n}{2}\hat{P}(\hat{\phi}_P) = o_p(\epsilon_n^2). \tag{3.41}$$

Since $\frac{1}{n}\sum_{i=1}^n \hat{g}_j^2(\Pi_n\phi_j^*, W_i)\hat{\sigma}^{-2}(W_i) = O_p(1)$ by Assumptions 3.3.1(ii), (iv), 3.3.2, 3.3.3(i) and 3.3.5(i), combining results (3.40) and (3.41) yields

$$\frac{1}{n}\sum_{i=1}^n \hat{g}_j(\Pi_n\phi_j^*, W_i)\hat{m}(\hat{\beta}_P, \hat{\phi}_P, W_i)\hat{\sigma}^{-2}(W_i) = o_p(n^{-1/2}). \tag{3.42}$$

By Theorem 3.3.2, Lemmas 3.9.5 and 3.9.6, Assumption 3.3.3(i), the triangle inequality and the Cauchy Schwartz inequality, result (3.42) implies

$$\frac{1}{n}\sum_{i=1}^n g_j(\phi_j^*, W_i)\hat{m}(\hat{\beta}_P, \hat{\phi}_P, W_i)\sigma^{-2}(W_i) = o_p(n^{-1/2}). \tag{3.43}$$

For $j = 1, \ldots, d_x$, let $\hat{g}_j^{(\sigma)}(\phi_j^*, W) \equiv \hat{\mathrm{E}}[g_j(\phi_j^*, W)\sigma^{-2}(W)|W]$. Then result (3.43) can be written as

$$\frac{1}{n}\sum_{i=1}^n \hat{g}_j^{(\sigma)}(\phi_j^*, W_i)\rho(\hat{\beta}_P, \hat{\phi}_P, Y_i, X_i, Z_i) = o_p(n^{-1/2}). \tag{3.44}$$

By Lemma 3.9.7 and Theorems 3.3.1 and 3.3.2, result (3.44) implies

$$\frac{1}{n}\sum_{i=1}^n g_j(\phi_j^*, W_i)\rho(\hat{\beta}_P, \hat{\phi}_P, Y_i, X_i, Z_i)\sigma^{-2}(W_i) = o_p(n^{-1/2}). \tag{3.45}$$

Let $G^*(W) \equiv \mathrm{E}[X - \Phi^*(Z)|W]$, which is equal to $(g_1(\phi_1^*, W), \ldots, g_{d_x}(\phi_{d_x}^*, W))'$ as well. For $\lambda \in \mathbb{S}^{d_x}$, let $\mathcal{F}_\lambda \equiv \{f : \mathbf{R} \times \mathbf{R}^{d_x} \times \mathcal{Z} \times \mathcal{W} \to \mathbf{R} : f(y, x, z, w) = \lambda'\Gamma^{-1}G^*(w)$ $\rho(\beta, \phi, y, x, z)\sigma^{-2}(w), (\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi}\}$. Then $\mathcal{F}_\lambda$ is Donsker by Lemma 3.9.8 since $\sup_{w\in\mathcal{W}}$ $|\lambda'\Gamma^{-1}G^*(w)| < \infty$ by Assumptions 3.3.1(ii), (iv), (v) and 3.3.5(i). Thus, the stochastic

continuity and Theorem 3.3.1 imply

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\lambda'\Gamma^{-1}G^*(W_i)\rho(\hat{\beta}_P,\hat{\phi}_P,Y_i,X_i,Z_i)\sigma^{-2}(W_i)\\
&\quad-\lambda'\Gamma^{-1}\,\mathrm{E}[G^*(W)\rho(\hat{\beta}_P,\hat{\phi}_P,Y,X,Z)\sigma^{-2}(W)]\\
&\quad-\frac{1}{n}\sum_{i=1}^{n}\lambda'\Gamma^{-1}G^*(W_i)\rho(\beta_0,\phi_P,Y_i,X_i,Z_i)\sigma^{-2}(W_i)\\
&\quad+\lambda'\Gamma^{-1}\,\mathrm{E}[G^*(W)\rho(\beta_0,\phi_P,Y,X,Z)\sigma^{-2}(W)]=o_p(n^{-1/2}). \qquad (3.46)
\end{aligned}$$

Note that $\mathrm{E}[G^*(W)\rho(\beta_0,\phi_P,Y,X,Z)\sigma^{-2}(W)]=0$, so combining results (3.15), (3.45) and (3.46) yields

$$\lambda'(\sqrt{n}(\hat{\beta}_P-\beta_0))=\lambda'\Gamma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}G^*(W_i)\rho(\beta_0,\phi_P,Y_i,X_i,Z_i)\sigma^{-2}(W_i)+o_p(1). \qquad (3.47)$$

Therefore, the result of the theorem follows by result (3.47), the central limit theorem and the Slutsky lemma.

Proof of Lemma 3.4.1: For $0\neq\lambda\in\mathbf{R}^{d_x}$, let $a(W)\equiv\lambda'\Gamma^{-1}\,\mathrm{E}[X-\Phi^*(Z)|W](\mathrm{E}[X-\Phi^*(Z)|W])'\sigma^{-4}(W)\Gamma^{-1}\lambda$, then $P_\lambda(\phi)=\mathrm{E}[a(W)(Y-X'\beta_0-\phi(Z))^2]$. By Assumptions 3.3.1(i)-(iv), 3.3.3(i) and 3.3.5(i), we have for any $\phi_1,\phi_2\in\mathbf{\Phi}$,

$$|P_\lambda(\phi_1)-P_\lambda(\phi_2)|\lesssim\|\phi_1-\phi_2\|_\infty, \qquad (3.48)$$

which implies that $P_\lambda(\cdot)$ is continuous in $\mathbf{\Phi}$. Obviously, $P_\lambda(\cdot)$ is strictly convex in $\mathbf{\Phi}$. Note that $\mathbf{\Phi}_0$ is convex and compact since $\mathbf{\Phi}$ is compact by Assumption 3.3.1(ii) and (iii), so $P_\lambda(\cdot)$ has a unique minimizer over $\mathbf{\Phi}_0$. This completes the proof of the lemma. ■

Proof of Proposition 3.4.1: The proof proceeds by showing $\|\hat{\Gamma}-\Gamma\|=o_p(1)$ and $\sup_{\phi\in\mathbf{\Phi}}\|\hat{\Sigma}(\phi)-\Sigma(\phi)\|=o_p(1)$, and concludes the result of the proposition by the triangle inequality. We first show $\|\hat{\Gamma}-\Gamma\|=o_p(1)$. By Lemma 3.9.11, Assumption 3.3.3(i) and the

triangle inequality, for $1 \leq j, k \leq d_x$,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{g}_j(\varphi, W_i)\hat{g}_k(\psi, W_i)\hat{\sigma}^{-2}(W_i) = \frac{1}{n}\sum_{i=1}^{n} g_j(\varphi, W_i)g_k(\psi, W_i)\sigma^{-2}(W_i) + o_p(1) \tag{3.49}$$

uniformly over $\varphi, \psi \in \mathbf{\Phi}$. By the similar argument as in the proof of Lemma 3.9.8, we have $\{f : \mathcal{W} \to \mathbf{R} : f(w) = g_j(\varphi, w)g_k(\psi, w)\sigma^{-2}(w), \varphi, \psi \in \mathbf{\Phi}\}$ is Glivenko Cantelli. So

$$\frac{1}{n}\sum_{i=1}^{n} g_j(\varphi, W_i)g_k(\psi, W_i)\sigma^{-2}(W_i) = \mathrm{E}[g_j(\phi, W)g_k(\varphi, W)\sigma^{-2}(W)] + o_p(1) \tag{3.50}$$

uniformly over $\varphi, \psi \in \mathbf{\Phi}$. Since $\mathrm{E}[g_j(\phi_j^*, W)\,\mathrm{E}[\varphi(Z)|W]\sigma^{-2}(W)] = 0$ for all $\varphi \in \mathbf{\Phi}$,

$$\begin{aligned}\mathrm{E}[g_j(\varphi, W)g_k(\psi, W)\sigma^{-2}(W)] &= \mathrm{E}[g_j(\phi_j^*, W)g_k(\phi_k^*, W)\sigma^{-2}(W)] \\ &\quad + \mathrm{E}[(g_j(\varphi, W) - g_j(\phi_j^*, W))(g_k(\psi, W) - g_j(\phi_k^*, W))\sigma^{-2}(W)].\end{aligned} \tag{3.51}$$

Combining results results (3.49), (3.50) and (3.51) yields

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \hat{g}_j(\varphi, W_i)\hat{g}_k(\psi, W_i)\hat{\sigma}^{-2}(W_i) &= \mathrm{E}[g_j(\phi_j^*, W)g_k(\phi_k^*, W)\sigma^{-2}(W)] \\ &\quad + \mathrm{E}[(g_j(\varphi, W) - g_j(\phi_j^*, W))(g_k(\psi, W) - g_j(\phi_k^*, W))\sigma^{-2}(W)] + o_p(1)\end{aligned} \tag{3.52}$$

uniformly over $\varphi, \psi \in \mathbf{\Phi}$. Setting $\varphi = \psi$ and $j = k$ in result (3.52) leads to

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \hat{g}_j^2(\varphi, W_i)\hat{\sigma}^{-2}(W_i) &= \mathrm{E}[g_j^2(\phi_j^*, W)\sigma^{-2}(W)] \\ &\quad + \mathrm{E}[(g_j(\varphi, W) - g_j(\phi_j^*, W))^2\sigma^{-2}(W)] + o_p(1)\end{aligned} \tag{3.53}$$

uniformly over $\varphi \in \mathbf{\Phi}$. Since $\frac{1}{n}\sum_{i=1}^{n} \hat{g}_j^2(\hat{\phi}_j^*, W_i)\hat{\sigma}^{-2}(W_i) \leq \frac{1}{n}\sum_{i=1}^{n} \hat{g}_j^2(\Pi_n\phi_j^*, W_i)\hat{\sigma}^{-2}(W_i)$ by the definition of $\hat{\phi}_j^*$ in (3.19), applying result (3.53) to $\varphi = \hat{\phi}_j^*$ and $\Pi_n\phi_j^*$ leads to

$$\begin{aligned}&\mathrm{E}[(g_j(\hat{\phi}_j^*, W) - g_j(\phi_j^*, W))^2\sigma^{-2}(W)] \\ &\qquad \leq \mathrm{E}[(g_j(\Pi_n\phi_j^*, W) - g_j(\phi_j^*, W))^2\sigma^{-2}(W)] + o_p(1) = o_p(1),\end{aligned} \tag{3.54}$$

where the equality follows since $\mathrm{E}[(g_j(\Pi_n\phi_j^*, W) - g_j(\phi_j^*, W))^2\sigma^{-2}(W)] = o(1)$ due to $\|\Pi_n\phi_j^* - \phi_j^*\|_\infty = o(1)$. By the Cauchy Schwartz inequality, result (3.54) implies

$$\mathrm{E}[(g_j(\hat{\phi}_j^*, W) - g_j(\phi_j^*, W))(g_k(\hat{\phi}_k^*, W) - g_k(\phi_k^*, W))\sigma^{-2}(W)] = o_p(1). \tag{3.55}$$

Combing results (3.52) and (3.55) yields

$$\frac{1}{n}\sum_{i=1}^{n} \hat{g}_j(\hat{\phi}_j^*, W_i)\hat{g}_k(\hat{\phi}_k^*, W_i)\hat{\sigma}^{-2}(W_i) = \mathrm{E}[g_j(\phi_j^*, W)g_k(\phi_k^*, W)\sigma^{-2}(W)] + o_p(1), \tag{3.56}$$

which implies $\|\hat{\Gamma} - \Gamma\| = o_p(1)$. We next show $\sup_{\phi\in\mathbf{\Phi}} \|\hat{\Sigma}(\phi) - \Sigma(\phi)\| = o_p(1)$. By Lemma 3.9.11, $\|\hat{\beta}_1 - \beta_0\| = o_p(1)$, Assumptions 3.3.1(i)-(iv) and the triangle inequality, for $1 \leq j, k \leq d_x$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \hat{g}_j(\varphi, W_i)\hat{g}_k(\psi, W_i)\hat{\sigma}^{-4}(W_i)\rho^2(\hat{\beta}_1, \phi, Y_i, X_i, Z_i) \\
&\qquad = \frac{1}{n}\sum_{i=1}^{n} g_j(\varphi, W_i)g_k(\psi, W_i)\sigma^{-4}(W_i)\rho^2(\beta_0, \phi, Y_i, X_i, Z_i) + o_p(1)
\end{aligned} \tag{3.57}$$

uniformly over $\varphi, \psi, \phi \in \mathbf{\Phi}$. By the similar argument as in the proof of Lemma 3.9.8, we have $\{f : \mathbf{R} \times \mathbf{R}^{d_x} \times \mathcal{Z} \times \mathcal{W} \to \mathbf{R} : f(y, x, z, w) = g_j(\varphi, w)g_k(\psi, w)\sigma^{-4}(w)\ \rho^2(\beta_0, \phi, y, x, z), \varphi, \psi, \phi \in \mathbf{\Phi}\}$ is Glivenko Cantelli. So

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} g_j(\varphi, W_i)g_k(\psi, W_i)\sigma^{-4}(W_i)\rho^2(\beta_0, \phi, Y_i, X_i, Z_i) \\
&\qquad = \mathrm{E}[g_j(\varphi, W)g_k(\psi, W)\sigma^{-4}(W)\rho^2(\beta_0, \phi, Y, X, Z)] + o_p(1)
\end{aligned} \tag{3.58}$$

uniformly over $\varphi, \psi, \phi \in \mathbf{\Phi}$. Combining results (3.57) and (3.58) yields

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \hat{g}_j(\hat{\phi}_j^*, W_i)\hat{g}_k(\hat{\phi}_k^*, W_i)\hat{\sigma}^{-4}(W_i)\rho^2(\hat{\beta}_P, \phi, Y_i, X_i, Z_i) \\
&\qquad = \mathrm{E}[g_j(\hat{\phi}_j^*, W)g_k(\hat{\phi}_k^*, W)\sigma^{-4}(W)\rho^2(\beta_0, \phi, Y, X, Z)] + o_p(1)
\end{aligned} \tag{3.59}$$

uniformly over $\phi \in \mathbf{\Phi}$. By the triangle inequality and the Cauchy Schwartz inequality,

results (3.54) and (3.59) imply

$$\frac{1}{n}\sum_{i=1}^{n} \hat{g}_j(\hat{\phi}_j^*, W_i)\hat{g}_k(\hat{\phi}_k^*, W_i)\hat{\sigma}^{-4}(W_i)\rho^2(\hat{\beta}_P, \phi, Y_i, X_i, Z_i)$$
$$= \mathrm{E}[g_j(\phi_j^*, W)g_k(\phi_k^*, W)\sigma^{-4}(W)\rho^2(\beta_0, \phi, Y, X, Z)] + o_p(1) \tag{3.60}$$

uniformly over $\phi \in \mathbf{\Phi}$, which implies $\sup_{\phi\in\mathbf{\Phi}} \|\hat{\Sigma}(\phi) - \Sigma(\phi)\| = o_p(1)$. ■

Proof of Corollary 3.5.1: By Theorem 3.3.1, we have $\|\hat{\beta}_P - \beta_0\| = o_p(1)$ and $\|\hat{\phi}_P - \phi_P\|_\infty = o_p(1)$. Since $\hat{\beta}_1 = \hat{\beta}_P$ and $\hat{\phi}_P \in \mathbf{\Phi}$, Proposition 3.4.1 implies

$$\|\hat{\Gamma}^{-1}\hat{\Sigma}(\hat{\phi}_P)\hat{\Gamma}^{-1} - \Gamma^{-1}\Sigma(\hat{\phi}_P)\Gamma^{-1}\| = o_p(1). \tag{3.61}$$

Note that $\|\Sigma(\phi_1) - \Sigma(\phi_2)\| \lesssim \|\phi_1 - \phi_2\|_\infty$ for any $\phi_1, \phi_2 \in \mathbf{\Phi}$ by Assumptions 3.3.1(i)-(iv), so $\Sigma(\cdot)$ is continuous in $\mathbf{\Phi}$ with respect to $\|\cdot\|_\infty$. By the continuous mapping theorem, it together with $\|\hat{\phi}_P - \phi_P\|_\infty = o_p(1)$ implies

$$\|\Gamma^{-1}\Sigma(\hat{\phi}_P)\Gamma^{-1} - \Gamma^{-1}\Sigma(\phi_P)\Gamma^{-1}\| = o_p(1). \tag{3.62}$$

Combine results (3.61) and (3.62) to conclude the result of the corollary by noting that $\Sigma(\phi_P) = \Sigma_P$.

3.9.2 Useful Lemmas

Lemma 3.9.1. *Suppose $\{Y_i, X_i, Z_i, W_i\}_{i=1}^n$ is a set of independently and identically distributed observations according to*

$$Y = X\beta_0 + Z\gamma_0 + \varepsilon \text{ with } X = W'b + u \text{ and } Z = W'c + v,$$

where $Y \in \mathbf{R}$ is the dependent variable, $X \in \mathbf{R}$ and $Z \in \mathbf{R}$ are potentially endogenous variables, and $W \in \mathbf{R}^2$ are IVs such that $\mathrm{E}[W(\varepsilon, u, v)] = 0$. Suppose $\mathrm{E}[WW'] = I$ and $\|\mathrm{E}[WW'\epsilon^2]\| < \infty$ for $\epsilon \in \{\varepsilon, u, v\}$. If $b = (1, 0)'$ and $c = 0$, then the IV estimator for β_0 is

not necessarily asymptotically normally distributed.

PROOF: The IV estimator for β_0 is given by

$$\hat{\beta} = \beta_0 + \frac{\hat{c}_2\hat{e}_1 - \hat{c}_1\hat{e}_2}{\hat{c}_2\hat{b}_1 - \hat{c}_1\hat{b}_2}, \tag{3.63}$$

where $\hat{b} \equiv (\hat{b}_1, \hat{b}_2)' \equiv \frac{1}{n}\sum_{i=1}^n W_i X_i$, $\hat{c} \equiv (\hat{c}_1, \hat{c}_2)' \equiv \frac{1}{n}\sum_{i=1}^n W_i Z_i$ and $\hat{e} \equiv (\hat{e}_1, \hat{e}_2)' \equiv \frac{1}{n}\sum_{i=1}^n W_i \varepsilon_i$. Since $\|\operatorname{E}[WW'\epsilon^2]\| < \infty$ for $\epsilon \in \{\varepsilon, v\}$ and $c = 0$, $(\sqrt{n}\hat{e}', \sqrt{n}\hat{c}')' \xrightarrow{d} N(0, \operatorname{E}[((\varepsilon, v)(\varepsilon, v)') \otimes WW'])$ by the central limit theorem, where $\otimes$ denotes Kronecker product. Since $\operatorname{E}[WW'] = I$, $\|\operatorname{E}[WW'u^2]\| < \infty$ and $b = (1, 0)'$, $\hat{b} \xrightarrow{p} (1, 0)'$ by the law of large numbers. It follows from (3.63) by the Slutsky lemma that

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} \frac{\mathcal{Z}_4\mathcal{Z}_1 - \mathcal{Z}_3\mathcal{Z}_2}{\mathcal{Z}_4}, \tag{3.64}$$

where $\mathcal{Z} \equiv (\mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3, \mathcal{Z}_4)'$ is a zero mean Gaussian random vector with covariance matrix $\operatorname{E}[((\varepsilon, v)(\varepsilon, v)') \otimes WW']$. The result of the lemma follows by noting that the right hand side of (3.64) is not normally distributed if $\operatorname{E}[W_1W_2v^2] = 0$ and $\operatorname{E}[W_2^2\varepsilon v] = 0$. ■

Lemma 3.9.2. *Suppose $\mathbf{\Phi} = C_M^{\gamma_z}$ for some $M > 0$ and $\gamma_z > 0$, Assumptions 3.3.1(i) and (iii) hold. Then we have*

$$\sup_{\phi \in \mathbf{\Phi}} \left| \frac{1}{n}\sum_{i=1}^n \phi^2(Z_i) - \operatorname{E}[\phi^2(Z)] \right| = o_p(1)$$

PROOF: Let $\mathcal{G} \equiv \{f : \mathcal{Z} \to \mathbf{R} : f(z) = \phi^2(z), \phi \in \mathbf{\Phi}\}$. For any $\phi_1, \phi_2 \in \mathbf{\Phi}$, $|\phi_1^2(z) - \phi_2^2(z)| \leq 2M\|\phi_1 - \phi_2\|_\infty$. So $\mathcal{G}$ is lipschitz in $\mathbf{\Phi}$. Theorem 2.7.1 and 2.7.11 of van der Vaart and Wellner (1996b) imply for every $\epsilon > 0$,

$$N_{[]}(\epsilon, \mathcal{G}, L^1(\mathrm{P})) \leq N(\epsilon/4M, \mathbf{\Phi}, \|\cdot\|_\infty) \lesssim \exp((4M/\epsilon)^{d_z/\gamma_z}) < \infty. \tag{3.65}$$

Result (3.65) implies $\mathcal{G}$ is Glivenko-Cantelli by Theorem 2.4.1 in van der Vaart and Wellner (1996b), which gives the result of the lemma. ■

Lemma 3.9.3. *Suppose Assumptions 3.3.1(i)-(iv), 3.3.3(i) and 3.3.4 hold. Let* $\delta_{n,m} \equiv \max\{\sqrt{k_n/n}, k_n^{-\alpha_w}\}$. *Then there exists finite constants* $C, C' > 0$ *such that*

$$C\,\mathrm{E}[m^2(\beta,\phi,W)\sigma^{-2}(W)] - O_p(\delta_{m,n}^2) \leq \frac{1}{n}\sum_{i=1}^{n} \hat{m}^2(\beta,\phi,W_i)\hat{\sigma}^2(W_i)$$
$$\leq C'\,\mathrm{E}[m^2(\beta,\phi,W)\sigma^{-2}(W)] + O_p(\delta_{m,n}^2)$$

uniformly over $(\beta,\phi) \in \mathbf{B}\times\mathbf{\Phi}$.

Proof: The proof proceeds by verifying the conditions of Lemma C.2(ii) of Chen and Pouzo (2012a). For $j = 1,\ldots,k_n$, let $\mathcal{O}_j \equiv \{f : \mathbf{R}\times\mathbf{R}^{d_x}\times\mathcal{Z}\times\mathcal{W}\to\mathbf{R} : f(y,x,z,w) = p_j(w)\rho(\beta,\phi,y,x,z), (\beta,\phi)\in\mathbf{B}\times\mathbf{\Phi}\}$. Note that $\max_{1\leq j\leq k_n}\mathrm{E}[p_j^2(W)] \lesssim 1$ by Assumption 3.3.4(iii), so Lemma 3.9.8 implies

$$\max_{1\leq j\leq k_n} J_{[]}(1,\mathcal{O}_j,\|\cdot\|_{L^2(\mathrm{P})}) \lesssim 1. \tag{3.66}$$

Thus, result (3.66) implies Assumption C.2(iii) of Chen and Pouzo (2012a) is satisfied with $C_n \lesssim 1$. Note that $|\rho(\beta,\phi,Y,X,Z)| \lesssim |Y| + \|X\| + 1$ and $\mathrm{E}[(|Y| + \|X\| + 1)^2|W] \lesssim 1$ by Assumptions 3.3.1(i), (ii) and (iv), so Assumption C.2(i) of Chen and Pouzo (2012a) is satisfied. Assumptions 3.3.4(v) and (vi) imply that for $(\beta,\phi)\in\mathbf{B}\times\mathbf{\Phi}$, there is $\pi_{\beta,\phi}\in\mathbf{R}^{k_n}$ such that $\sup_{w\in\mathcal{W}}|m(\beta,\phi,w) - p^{k_n}(w)'\pi_{\beta,\phi}| = O(k_n^{-\alpha_w})$ uniformly over $(\beta,\phi)\in\mathbf{B}\times\mathbf{\Phi}$, so Assumption C.2(ii) of Chen and Pouzo (2012a) is satisfied. In addition, Assumption C.1 of Chen and Pouzo (2012a) is satisfied by Assumptions 3.3.1(i), 3.3.3(i) and 3.3.4(i)-(iv). Thus, the result of the lemma follows by Lemma C.2(ii) of Chen and Pouzo (2012a). ■

Lemma 3.9.4. *Suppose Assumptions 3.3.1(i), (ii), (iv) and 3.3.3(i) hold. Then we have for any* $(\beta_1,\phi_1),(\beta_2,\phi_2)\in\mathbf{B}\times\mathbf{\Phi}$,

$$|\,\mathrm{E}[m^2(\beta_1,\phi_1,W)\sigma^{-2}(W)] - \mathrm{E}[m^2(\beta_2,\phi_2,W)\sigma^{-2}(W)]| \lesssim \|\beta_1-\beta_2\| + \|\phi_1-\phi_2\|_\infty.$$

PROOF: We have for any $(\beta_1, \phi_1), (\beta_2, \phi_2) \in \mathbf{B} \times \mathbf{\Phi}$,

$$\begin{aligned} &|\,\mathrm{E}[m^2(\beta_1, \phi_1, W)\sigma^{-2}(W)] - \mathrm{E}[m^2(\beta_2, \phi_2, W)\sigma^{-2}(W)]| \\ &\qquad \lesssim (\mathrm{E}[m^2(\beta_1, \phi_1, W)] + \mathrm{E}[m^2(\beta_2, \phi_2, W)])^{1/2} \\ &\qquad\qquad \times (\mathrm{E}[m(\beta_1, \phi_1, W) - m(\beta_2, \phi_2, W)]^2)^{1/2}. \end{aligned} \tag{3.67}$$

by the Cauchy Schwartz inequality and Assumption 3.3.3(i). Note that $m(\beta, \phi, W) = \mathrm{E}[X'(\beta_0 - \beta)|W] + \mathrm{E}[\phi_0(Z) - \phi(Z)|W]$, so for any $(\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi}$,

$$\mathrm{E}[m^2(\beta, \phi, W)] \lesssim 1. \tag{3.68}$$

by Assumption 3.3.1(ii) and (iv). Note that $m(\beta_1, \phi_1, W) - m(\beta_2, \phi_2, W) = \mathrm{E}[X'(\beta_2 - \beta_1)] + \mathrm{E}[\phi_2(Z) - \phi_1(Z)|W]$, so for any $(\beta_1, \phi_1), (\beta_2, \phi_2) \in \mathbf{B} \times \mathbf{\Phi}$,

$$\mathrm{E}[m(\beta_1, \phi_1, W) - m(\beta_2, \phi_2, W)]^2 \lesssim (\|\beta_1 - \beta_2\| + \|\phi_1 - \phi_2\|)^2 \tag{3.69}$$

by Assumption 3.3.1(ii) and (iv). Combine results (3.67)-(3.69) to conclude the result of the lemma. ■

Lemma 3.9.5. *Suppose Assumptions 3.3.1(i)-(iv) and 3.3.4 (i)-(iv) hold. Then we have*

$$\frac{1}{n}\sum_{i=1}^{n} \hat{m}^2(\beta, \phi, W_i) = o_p(n^{-1/2}) + O_p(k_n/n)$$

uniformly over $(\beta, \phi) \in \{(\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi} : \|(\beta, \phi) - (\beta_0, \phi_P)\|_w = o_p(n^{-1/4})\}$.

PROOF: For any $(\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi}$, let $u(\beta, \phi, Y, X, Z, W) \equiv \rho(\beta, \phi, Y, X, Z) - m(\beta, \phi, W)$ and $\hat{u}(\beta, \phi, W) \equiv \hat{\mathrm{E}}[u(\beta, \phi, Y, X, Z, W)|W]$. Then by the Cauchy Schwartz inequality and $P(P'P)^- P'$ is idempotent, we have

$$\frac{1}{n}\sum_{i=1}^{n} \hat{m}^2(\beta, \phi, W_i) \leq \frac{2}{n}\sum_{i=1}^{n} m^2(\beta, \phi, W_i) + \frac{2}{n}\sum_{i=1}^{n} \hat{u}^2(\beta, \phi, W_i). \tag{3.70}$$

Following the same argument as in the proof Lemma 3.9.8, we have $\{f : \mathcal{W} \to \mathbf{R} : f(w) =$

$m^2(\beta,\phi,w),(\beta,\phi)\in\mathbf{B}\times\mathbf{\Phi}\}$ is Donsker. Note that $\mathrm{E}[(m^2(\beta,\phi,W)-m^2(\beta_0,\phi_P,W))^2]\lesssim\|(\beta,\phi)-(\beta_0,\phi_P)\|_w^2$ and $m(\beta_0,\phi_P,W)=0$, so by the stochastic equicontinuity it follows

$$\frac{1}{n}\sum_{i=1}^{n}m^2(\beta,\phi,W_i)-\mathrm{E}[m^2(\beta,\phi,W)]=o_p(n^{-1/2}) \tag{3.71}$$

uniformly over $(\beta,\phi)\in\{(\beta,\phi)\in\mathbf{B}\times\mathbf{\Phi}:\|(\beta,\phi)-(\beta_0,\phi_P)\|_w=o_p(1)\}$. Recall that $\mathrm{E}[m^2(\beta,\phi,W)]\lesssim\|(\beta,\phi)-(\beta_0,\phi_P)\|_w^2$, so result (3.71) implies

$$\frac{1}{n}\sum_{i=1}^{n}m^2(\beta,\phi,W_i)\lesssim\|(\beta,\phi)-(\beta_0,\phi_P)\|_w^2+o_p(n^{-1/2})=o_p(n^{-1/2}) \tag{3.72}$$

uniformly over $(\beta,\phi)\in\{(\beta,\phi)\in\mathbf{B}\times\mathbf{\Phi}:\|(\beta,\phi)-(\beta_0,\phi_P)\|_w=o_p(n^{-1/4})\}$. By the proof of Lemma 3.9.3, Assumption C.1, C.2(i) and (iii) of Chen and Pouzo (2012a) are satisfied with $C_n\lesssim 1$ and Lemma C.1(ii) of Chen and Pouzo (2012a) implies

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}^2(\beta,\phi,W_i)=O_p(k_n/n) \tag{3.73}$$

uniformly over $(\beta,\phi)\in\{(\beta,\phi)\in\mathbf{B}\times\mathbf{\Phi}:\|(\beta,\phi)-(\beta_0,\phi_P)\|_w=o_p(n^{-1/4})\}$. Combine results (3.70), (3.72) and (3.73) to conclude the result of the lemma. ■

Lemma 3.9.6. *Suppose Assumptions 3.3.1(i), (ii), (iv), 3.3.2(ii), 3.3.4(v), (vi), 3.3.5(i) and (ii) hold. Then we have*

$$\max_{1\le j\le d_x}\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j(\Pi_n\phi_j^*,W_i)-g_j(\phi_j^*,W_i))^2=O_p(\max\{J_n^{-2\alpha_z},\delta_{m,n}^2\}).$$

Proof: By the Cauchy Schwartz inequality, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j(\Pi_n\phi_j^*,W_i)-g_j(\phi_j^*,W_i))^2\le{}&\frac{2}{n}\sum_{i=1}^{n}(\hat{g}_j(\Pi_n\phi_j^*,W_i)-\hat{g}_j(\phi_j^*,W_i))^2\\&+\frac{2}{n}\sum_{i=1}^{n}(\hat{g}_j(\phi_j^*,W_i)-g_j(\phi_j^*,W_i))^2\end{aligned} \tag{3.74}$$

Note that $P(P'P)^{-}P'$ is idempotent, so by Assumptions 3.3.2(ii) and 3.3.5(i) we have

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j(\Pi_n\phi_j^*, W_i) - \hat{g}_j(\phi_j^*, W_i))^2 \leq \frac{1}{n}\sum_{i=1}^{n}(\phi_j^*(Z_i) - \Pi_n\phi_j^*(Z_i))^2 = O(J_n^{-2\alpha_z}). \tag{3.75}$$

Let $G_j^* \equiv (g_j(\phi_j^*, W_1), \ldots, g_j(\phi_j^*, W_n))'$ and $H_j^* \equiv (h_j(\phi_j^*, X_1, Z_1), \ldots, h_j(\phi_j^*, X_n, Z_n))'$. Then we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j(\phi_j^*, W_i) - g_j(\phi_j^*, W_i))^2 = \frac{1}{n}(G_j^*)'(I - P(P'P)^{-}P')G_j^* \\ + \frac{1}{n}(H_j^* - G_j^*)'P(P'P)^{-}P'(H_j^* - G_j^*).\end{aligned} \tag{3.76}$$

Note that $\sup_{w\in\mathcal{W}} \mathrm{E}[(h_j(\phi_j^*, X, Z))^2|W = w] \lesssim 1$ by Assumptions 3.3.1(ii), (iv) and 3.3.5(i), so $\frac{1}{n}(H_j^* - G_j^*)'P(P'P)^{-}P'(H_j^* - G_j^*) = O_p(k_n/n)$ by the Markov inequality. By Assumptions 3.3.4(v), (vi) and 3.3.5(i), $\frac{1}{n}(G_j^*)'(I - P(P'P)^{-}P')G_j^* = O(k_n^{-2\alpha_w})$. Hence, result (3.76) implies

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j(\phi_j^*, W_i) - g_j(\phi_j^*, W_i))^2 = O_p(\delta_{m,n}^2). \tag{3.77}$$

Combine results (3.74), (3.75) and (3.77) to conclude the result of the lemma. ■

Lemma 3.9.7. *Suppose Assumptions 3.3.1(i)-(iv), 3.3.3(i), 3.3.4(iii) and 3.3.5(iii) hold. Then we have*

$$\max_{1\leq j\leq d_x}\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))\rho(\beta, \phi, Y_i, X_i, Z_i) = o_p(n^{-1/2}) + O_p(k_n^{-2\alpha_w})$$

uniformly over $(\beta, \phi) \in \{(\beta, \phi) \in \mathbf{B}\times\mathbf{\Phi} : \|(\beta, \phi) - (\beta_0, \phi_P)\|_w = o_p(n^{-1/4}), \|\beta - \beta_0\| + \|\phi - \phi_P\|_\infty = o_p(1)\}$.

Proof: Recalling that $u(\beta, \phi, Y, X, Z, W) = \rho(\beta, \phi, Y, X, Z) - m(\beta, \phi, W)$ for $(\beta, \phi) \in$

$\mathbf{B} \times \mathbf{\Phi}$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))\rho(\beta, \phi, Y_i, X_i, Z_i) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))m(\beta, \phi, W_i) \\
&\quad\quad + \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))u(\beta, \phi, Y_i, X_i, Z_i, W_i). \qquad (3.78)
\end{aligned}$$

For $j = 1, \ldots, d_x$, let $D_j \equiv (g_j(\phi_j^*, W_1)\sigma^{-2}(W_1), \ldots, g_j(\phi_j^*, W_n)\sigma^{-2}(W_n))'$. Then we have

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))^2 = \frac{1}{n}D_j'(I - P(P'P)'P')D_j. \qquad (3.79)$$

Note that $\frac{1}{n}D_j'(I - P(P'P)^-P')D_j \lesssim \sup_{w\in\mathcal{W}} \|G^*(w)\sigma^{-2}(w) - \Pi_\sigma^* p^{k_n}(w)\| = O(k_n^{-2\alpha_w})$ by Assumption 3.3.5(ii), so result (3.79) implies

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))^2 = O_p(k_n^{-2\alpha_w}). \qquad (3.80)$$

By the Cauchy Schwartz inequality, results (3.72) and (3.80) imply

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))m(\beta, \phi, W_i) = o_p(n^{-1/2}) + O_p(k_n^{-2\alpha_w}) \qquad (3.81)$$

uniformly over $(\beta, \phi) \in \{(\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi} : \|(\beta, \phi) - (\beta_0, \phi_P)\|_w = o_p(n^{-1/4})\}$. Due to result (3.80), $\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))u(\beta, \phi, Y_i, X_i, Z_i, W_i)$ is stochastically equicontinuous by Lemma 3.9.8. Thus, it follows

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))u(\beta, \phi, Y_i, X_i, Z_i, W_i) \\
&- \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))u(\beta_0, \phi_P, Y_i, X_i, Z_i, W_i) = o_p(n^{-1/2}) \qquad (3.82)
\end{aligned}$$

uniformly over $(\beta, \phi) \in \{(\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi} : \|\beta - \beta_0\| + \|\phi - \phi_P\|_\infty = o_p(1)\}$. Now by

Assumptions 3.3.1(i), (ii) and (iv), we have

$$\mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))u(\beta_0, \phi_P, Y_i, X_i, Z_i, W_i)]^2$$
$$\lesssim \mathrm{E}[\frac{1}{n^2}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))^2] \tag{3.83}$$

By the Markov inequality, results (3.80) and (3.83) imply

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))u(\beta_0, \phi_P, Y_i, X_i, Z_i, W_i)$$
$$= O_p(n^{-1/2}k_n^{-\alpha_w}). \tag{3.84}$$

Combining results (3.82) and (3.84) yields

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}_j^{(\sigma)}(\phi_j^*, W_i) - g_j(\phi_j^*, W_i)\sigma^{-2}(W_i))u(\beta, \phi, Y_i, X_i, Z_i, W_i)$$
$$= o_p(n^{-1/2}) + O_p(k_n^{-2\alpha_w}) \tag{3.85}$$

uniformly over $(\beta, \phi) \in \{(\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi} : \|\beta - \beta_0\| + \|\phi - \phi_P\|_\infty = o_p(1)\}$. Combine results (3.78), (3.81) and (3.85) to conclude the result of the lemma. ■

Lemma 3.9.8. *Suppose Assumptions 3.3.1(ii) and (iii) hold. Let* $c : \mathcal{W} \to \mathbf{R}$ *with* $\mathrm{E}[c^2(W)] < \infty$ *and* $\mathcal{F} \equiv \{f : \mathbf{R}\times\mathbf{R}^{d_x}\times\mathcal{Z}\times\mathcal{W} \to \mathbf{R} : f(y, x, z, w) = c(w)\rho(\beta, \phi, y, x, z), (\beta, \phi) \in \mathbf{B}\times\mathbf{\Phi}\}$. *Then there exits* $K > 0$ *such that* $F(y, x, z) \equiv K|c(w)|(|y|+\|x\|+1)$ *is an envelope for* $\mathcal{F}$ *and*

$$J_{[]}(1, \mathcal{F}, \|\cdot\|_{L^2(\mathrm{P})}) \lesssim 1.$$

If in addition Assumptions 3.3.1(i) and (iv) hold, then $\mathcal{F}$ *is Donsker.*

Proof: Since $\mathbf{B} \times \mathbf{\Phi}$ is bounded by Assumptions 3.3.1(ii) and (iii), so there exists $K > 0$ such that $|c(w)\rho(\beta, \phi, y, x, z)| \leq K|c(w)|(|y| + \|x\| + 1)$ for any $(\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi}$, which gives

the first result of the lemma. We have for any $(\beta_1, \phi_1), (\beta_2, \phi_2) \in \mathbf{B} \times \mathbf{\Phi}$,

$$\begin{aligned} &|c(w)\rho(\beta_1, \phi_1, y, x, z) - c(w)\rho(\beta_2, \phi_2, y, x, z)| \\ &\qquad \leq H(x, w)(\|\beta_1 - \beta_2\| + \|\phi_1 - \phi_2\|_\infty), \end{aligned} \tag{3.86}$$

where $H(x, w) \equiv c(w)(\|x\| + 1)$. Result (3.86) implies $\mathcal{F}$ is lipschitz continuous in $(\beta, \phi) \in \mathbf{B} \times \mathbf{\Phi}$. By Theorem 2.7.11 of van der Vaart and Wellner (1996b), for every $\epsilon > 0$

$$\begin{aligned} N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{L^2(\mathrm{P})}) &\leq N(\epsilon/(2\|H\|_{L^2(\mathrm{P})}), \mathbf{B} \times \mathbf{\Phi}, \|\cdot\| + \|\cdot\|_\infty) \\ &\leq N(\epsilon/4(\|H\|_{L^2(\mathrm{P})}), \mathbf{B}, \|\cdot\|) \times N(\epsilon/(4\|H\|_{L^2(\mathrm{P})}), \mathbf{\Phi}, \|\cdot\|_\infty). \end{aligned} \tag{3.87}$$

where the second inequality follows by $N(\epsilon, \mathbf{B} \times \mathbf{\Phi}, \|\cdot\| + \|\cdot\|_\infty) \leq N(\epsilon/2, \mathbf{B}, \|\cdot\|) \times N(\epsilon/2, \mathbf{\Phi}, \|\cdot\|_\infty)$. By Assumptions 3.3.1(ii) and (iii), Theorem 2.7.1 of van der Vaart and Wellner (1996b) implies

$$\log N(\epsilon, \mathbf{\Phi}, \|\cdot\|_\infty) \lesssim (\frac{1}{\epsilon})^{d_z/\gamma_z}. \tag{3.88}$$

Note that $N(\epsilon, \mathbf{B}, \|\cdot\|) \lesssim (2/\epsilon)^{d_x}$ and $\|H\|_{L^2(\mathrm{P})} \lesssim 1$ by Assumptions 3.3.1(iv), so results (3.87) and (3.88) imply

$$\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{L^2(\mathrm{P})}) \lesssim \log(\frac{2}{\epsilon})^{d_x} + (\frac{1}{\epsilon})^{d_z/\gamma_z} \tag{3.89}$$

for $\epsilon < 1$. By Assumption 3.3.1(ii), $\gamma_z > d_z/2$ and then result (3.89) implies

$$J_{[]}(1, \mathcal{F}, \|\cdot\|_{L^2(\mathrm{P})}) \lesssim 1, \tag{3.90}$$

which gives the second result of the lemma. Note that $\mathrm{E}[F^2(Y, X, Z)] < \infty$ by Assumptions 3.3.1(i), (ii) and (iv), so the last result of the lemma follows by Theorem 2.5.6 of van der Vaart and Wellner (1996b). ■

Lemma 3.9.9. *Suppose* $Y, X, Z \in \mathbf{R}$ *and* $W \in \mathbf{R}^2$ *satisfy* $\|\mathrm{E}[W(Y, X, Z)]\| \leq \infty$ *and*

$\mathrm{E}[W(Y - X\beta_0 - Z\gamma_0)] = 0$ *for a unique* $\beta_0 \in \mathbf{R}$ *and some* $\gamma_0 \in \mathbf{R}$. *Let* $\Theta \in \mathbf{R}^6$ *be the parameter space for* $(\mathrm{E}[YW'], \mathrm{E}[XW'], \mathrm{E}[ZW'])'$. *Then* $\beta_0(\theta)$ *is discontinuous in* Θ.

PROOF: Since β_0 is identified, it is either $\theta_3\theta_6 - \theta_4\theta_5 \neq 0$ or $\theta_5 = \theta_6 = 0, \theta_3^2 + \theta_4^2 \neq 0$ and $\theta_1\theta_4 = \theta_2\theta_3$, which corresponds to γ_0 being identified and not being identified, respectively. It follows

$$\beta_0(\theta) = \begin{cases} \frac{\theta_6\theta_1 - \theta_5\theta_2}{\theta_3\theta_6 - \theta_4\theta_5} & \text{if } \theta_3\theta_6 - \theta_4\theta_5 \neq 0 \\ \frac{\theta_1}{\theta_3} \text{ or } \frac{\theta_2}{\theta_4} & \text{if } \theta_5 = \theta_6 = 0, \theta_3^2 + \theta_4^2 \neq 0, \theta_1\theta_4 = \theta_2\theta_3. \end{cases} \tag{3.91}$$

Consider the path $\theta(c) \equiv (1, 0, 1, c, c, 0) \in \Theta$ for $c \in \mathbf{R}$, then (3.91) implies

$$\beta_0(\theta(c)) = \begin{cases} 0 & \text{if } c \neq 0 \\ 1 & \text{if } c = 0. \end{cases} \tag{3.92}$$

Result (3.92) implies $\beta_0(\theta)$ is not continuous at $\theta = (1, 0, 1, 0, 0, 0) \in \Theta$, so the result of the lemma follows. ■

Lemma 3.9.10. *Let* V_P *and* $\sigma_P^2(\cdot)$ *be given in Theorem 3.3.3. Then* $V_P - V_P^*$ *is positive semidefinite for a given* $P(\cdot)$, *where* V_P^* *denotes* V_P *evaluated* $\sigma^2(\cdot) = \sigma_P^2(\cdot)$.

PROOF: As $\mathrm{E}[X - \Phi^*(Z)|W]$ and Γ may depend on $\sigma^2(\cdot)$, we write $\mathrm{E}[X - \Phi_\sigma^*(Z)|W]$ and Γ_σ instead of $\mathrm{E}[X - \Phi^*(Z)|W]$ and Γ. Note that $\mathrm{E}[\mathrm{E}[X - \Phi_\sigma^*(Z)|W]\sigma^{-2}(W)(\mathrm{E}[X - \Phi_{\sigma_P}^*(Z)|W])'] = \Gamma_\sigma$ by result (3.13), so we have

$$V_P - V_P^* = \mathrm{E}[A(W)\sigma_P^2(W)(A(W))'], \tag{3.93}$$

where $A(W) \equiv \Gamma_\sigma^{-1}\,\mathrm{E}[X - \Phi_\sigma^*(Z)|W]\sigma^{-2}(W) - \Gamma_{\sigma_P}^{-1}\,\mathrm{E}[X - \Phi_{\sigma_P}^*(Z)|W]\sigma_P^{-2}(W)$. The result of the lemma follows by noting that the right hand side of (3.93) is positive semidefinite. ■

Lemma 3.9.11. *Suppose Assumptions 3.3.1(i)-(iv) and 3.3.4 hold and* $\delta_{m,n} = o(1)$. *Then we have*

$$\max_{1 \leq j \leq d_x} \sup_{w \in \mathcal{W}} |\hat{g}_j(\phi, w) - g_j(\phi, w)| = o_p(1)$$

uniformly over $\phi \in \mathbf{\Phi}$.

PROOF: For $\phi \in \mathbf{\Phi}$ and $j = 1, \ldots, d_x$, let $\hat{\pi}_{j,\phi} \equiv (P'P)^- \sum_{i=1}^n p^{k_n}(W)' h_j(\phi, X_i, Z_i)$. Thus, $\hat{g}_j(\phi, W) = p^{k_n}(W)'\hat{\pi}_{j,\phi}$. By Assumptions 3.3.4(v) and (vi), for $\phi \in \mathbf{\Phi}$ and $j = 1, \ldots, d_x$ there is $\pi_{j,\phi} \in \mathbf{R}^{k_n}$ such that

$$\sup_{w \in \mathcal{W}} |g_j(\phi, w) - p^{k_n}(w)'\pi_{j,\phi}| = O(k_n^{-\alpha_w}) \tag{3.94}$$

uniformly over $\phi \in \mathbf{\Phi}$. By the Cauchy Schwartz inequality and Assumption 3.3.4(iii), it follows

$$\mathrm{E}[(\hat{g}_j(\phi, W) - g_j(\phi, W))^2] \leq 2\|\hat{\pi}_{j,\phi} - \pi_{j,\phi}\|^2 + 2 \sup_{w \in \mathcal{W}} |g_j(\phi, w) - p^{k_n}(w)'\pi_{j,\phi}|^2. \tag{3.95}$$

Following Newey (1997)(p.162), by Assumptions 3.3.1(i), 3.3.4(iii) and (iv), with probability approaching one $P'P/n$ is invertible and

$$\begin{aligned} \|\hat{\pi}_{j,\phi} - \pi_{j,\phi}\| \leq O_p(1) \sup_{w \in \mathcal{W}} |g_j(\phi, w) - p^{k_n}(w)'\pi_{j,\phi}| \\ + O_p(1)\|\frac{1}{n}\sum_{i=1}^n p^{k_n}(W_i)'(h_j(\phi, X_i, Z_i) - g_j(\phi, W_i))\| \end{aligned} \tag{3.96}$$

uniformly over $\phi \in \mathbf{\Phi}$. Following Chen and Pouzo (2012b)(p.17-20), by Assumptions 3.3.1(i)-(iv) and 3.3.4, we have (in particular, $C_n \lesssim 1$ there following the similar argument as in the proof of Lemma 3.9.3)

$$\|\frac{1}{n}\sum_{i=1}^n p^{k_n}(W_i)'(h_j(\phi, X_i, Z_i) - g_j(\phi, W_i))\| = O_p(\sqrt{k_n/n}) \tag{3.97}$$

uniformly over $\phi \in \mathbf{\Phi}$. Combining results (3.94)-(3.97) yields

$$\mathrm{E}[(\hat{g}_j(\phi, W) - g_j(\phi, W))^2] = O_p(\delta_{m,n}^2). \tag{3.98}$$

uniformly over $\phi \in \mathbf{\Phi}$. Note that the density of W is bounded and bounded away from over

$\mathcal{W}$ that is compact by Assumptions 3.3.4(i) and (ii), so result (3.98) implies

$$\sup_{w \in \mathcal{W}} |\hat{g}_j(\phi, w) - g_j(\phi, w)| = o_p(1) \tag{3.99}$$

uniformly over $\phi \in \mathbf{\Phi}$, since $\delta_{m,n} = o(1)$. This completes the proof of the lemma. ■

References

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71** 1795–1843.

Ai, C. and Chen, X. (2007). Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, **141** 5–43.

Al-Sadoon, M. M. (2015). A general theory of rank testing. Working paper.

Aliprantis, C. D. and Border, K. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. 3rd ed. Springer Verlag.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, **22** 327–351.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. 3rd ed. Wiley New York.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, **68** 399–405.

Andrews, D. W. K. (2002). Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators. *Econometrica* 119–162.

Andrews, D. W. K. and Guggenberger, P. (2009a). Hybrid and size-corrected subsampling methods. *Econometrica*, **77** 721–762.

Andrews, D. W. K. and Guggenberger, P. (2009b). Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities. *Econometric Theory*, **25** 669–709.

Andrews, D. W. K. and Guggenberger, P. (2010). Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory*, **26** 426–468.

Andrews, D. W. K. and Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, **81** 609–666.

Andrews, D. W. K. and Shi, X. (2014). Nonparametric inference based on conditional moment inequalities. *Journal of Econometrics*, **179** 31 – 45.

Andrews, D. W. K. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, **78** 119–157.

Aoki, M. (1990). *State Space Modeling of Time Series*. 2nd ed. Springer.

Babu, G. J. (1984). Bootstrapping statistics with linear combinations of Chi-squares as weak limit. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, **46** 85–93.

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

Bai, J. and Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics*, **25** 52–60.

Barnett, W. A. and Serletis, A. (2008). Measuring consumer preferences and estimating demand systems. In *Quantifying Consumer Preferences*. 1–35.

Benkwitz, A., Neumann, M. H. and Lütekpohl, H. (2000). Problems related to confidence intervals for impulse responses of autoregressive processes. *Econometric Reviews*, **19** 69–103.

Beran, R. (1984). Jackknife approximations to bootstrap estimates. *The Annals of Statistics*, **12** 101–118.

Berkes, I., Horváth, L. and Rice, G. (2016). On the asymptotic normality of kernel estimators of the long run covariance of functional time series. *Journal of Multivariate Analysis*, **144** 150–175.

Bertail, P., Politis, D. N. and Romano, J. P. (1999). On subsampling estimators with unknown rate of convergence. *Journal of the American Statistical Association*, **94** 569–579.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, **9** 1196–1217.

Bickel, P. J., Götze, F. and van Zwet, W. R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica*, **7** 1–31.

Bickel, P. J., Klaassen, C. A., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.

Bickel, P. J., Klaassen, C. A., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.

Bierens, H. (1997). Nonparametric cointegration analysis. *Journal of Econometrics*, **77** 379–404.

Blundell, R., Chen, X. and Kristensen, D. (2007). Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, **75** 1613–1669.

Bogachev, V. I. (1998). *Gaussian Measures*. American Mathematical Society, Providence.

Bogachev, V. I. (2007). *Measure Theory*. Springer-Verlag.

Bonnans, J. F. and Shapiro, A. (2000). *Perturbation Analysis of Optimization Problems*. Springer.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. Springer.

Brillinger, D. (1981). *Time Series: Data Analysis and Theory*. Holden-Day.

Bura, E. and Yang, J. (2011). Dimension estimation in sufficient dimension reduction: a unifying approach. *Journal of Multivariate Analysis*, **102** 130–142.

Camba-Mendez, G. and Kapetanios, G. (2009a). Statistical tests and estimators of the rank of a matrix and their applications in econometric modelling. *Econometric Reviews*, **28** 581–611.

Camba-Mendez, G. and Kapetanios, G. (2009b). Statistical tests and estimators of the rank of a matrix and their applications in econometric modelling. *Econometric Reviews*, **28** 581–611.

Canay, I., Santos, A. and Shaikh, A. (2013). On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, **81** 2535–2559.

Caner, M. (2009). Lasso-type GMM estimator. *Econometric Theory*, **25** 270–290.

Canova, F. and Sala, L. (2009). Back to square one: Identification issues in DSGE models. *Journal of Monetary Economics*, **56** 431 – 449.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, **14** 1171–1179.

Cavaliere, G., Rahbek, A. and Taylor, A. (2012). Bootstrap determination of the co-integration rank in vector autoregressive models. *Econometrica*, **80** 1721–1740.

Chang, Y., Park, J. Y. and Song, K. (2006). Bootstrapping cointegrating regressions. *Journal of Econometrics*, **133** 703–739.

Chen, Q. and Fang, Z. (2015). Inference on functionals under first order degeneracy. Working paper.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, **6** 5549–5632.

Chen, X., Chernozhukov, V., Lee, S. and Newey, W. (2014a). Local identification of nonparametric and semiparametric models. *Econometrica*, **82** 785–809.

Chen, X., Chernozhukov, V., Lee, S. and Newey, W. K. (2014b). Local identification of nonparametric and semiparametric models. *Econometrica*, **82** 785–809.

Chen, X. and Pouzo, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, **152** 46–60.

Chen, X. and Pouzo, D. (2012a). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, **80** 277–321.

Chen, X. and Pouzo, D. (2012b). Supplement to "estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals". *Econometrica*, **80** 277–321.

Chen, X., Pouzo, D. and Tamer, E. (2011a). Sieve QLR inference on partially identified semi-nonparametric conditional moment models. Working paper, Yale University.

Chen, X., Tamer, E. and Torgovitsky, A. (2011b). Sensitivity analysis in semiparametric likelihood models. Working paper, Yale University.

Cheng, X. and Liao, Z. (2015). Select the valid and relevant moments: An information-based LASSO for GMM with many moments. *Journal of Econometrics*, **186** 443–464.

Chernozhukov, V., Hong, H. and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, **75** 1243–1284.

Chernozhukov, V., Newey, W. and Santos, A. (2015). Constrained conditional moment restriction models. Working paper, CEMMP.

Chesher, A. (2003). Identification in nonseparable models. *Econometrica*, **71** 1405–1441.

Choi, I. and Phillips, P. C. B. (1992). Asymptotic and finite sample distribution theory for iv estimators and tests in partially identified structural equations. *Journal of Econometrics*, **51** 113–150.

Cragg, J. G. and Donald, S. G. (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory*, **9** 222–240.

Cragg, J. G. and Donald, S. G. (1996). On the asymptotic properties of ldu-based tests of the rank of a matrix. *Journal of the American Statistical Association*, **91** 1301–1309.

Cragg, J. G. and Donald, S. G. (1997). Inferring the rank of a matrix. *Journal of Econometrics*, **76** 223 – 250.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramér-von Mises tests. *The Annals of Mathematical Statistics*, **28** 823–838.

Datta, S. (1995). On a modified bootstrap for certain asymptotically nonnormal statistics. *Statistics & Probability Letters*, **24** 91 – 98.

Davydov, Y. A., Lifshits, M. A. and Smorodina, N. V. (1998). *Local Properties of Distributions of Stochastic Functionals*. American Mathematical Society.

Dehling, H., Mikosch, T. and Sørensen, M. (2002). *Empirical Process Techniques for Dependent Data*. Springer.

Delgado, M. A., Rodríguez-Poo, J. M. and Wolf, M. (2001). Subsampling inference in cube root asymptotics with an application to manski's maximum score estimator. *Economics Letters*, **73** 241 – 250.

Dovonon, P. and Gonçalves, S. (2014). Bootstrapping the GMM overidentification test under first-order underidentification. Working paper.

Dovonon, P. and Renault, E. (2013). Testing for common conditionally heteroskedastic factors. *Econometrica*, **81** 2561–2586.

Dudley, R. (1990). Nonlinear functionals of empirical measures and the bootstrap. In *Probability in Banach Spaces 7* (E. Eberlein, J. Kuelbs and M. Marcus, eds.), vol. 21 of *Progress in Probability*. Birkhäuser Boston, 63–82.

Dudley, R. M. (1966). Convergence of Baire measures. *Studia Mathematica*, **27** 251–268.

Dudley, R. M. (1968). Distances of probability measures and random variables. *The Annals of Mathematical Statistics*, **39** 1563–1572.

Dugundji, J. (1951). An extension of Tietze's theorem. *Pacific Journal of Mathematics*, **1** 353–367.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, **95** 125–140.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, **7** 1–26.

Engle, R. (1984). Wald, likelihood ratio, and lagrange multiplier tests in econometrics. In *Handbook of Econometrics* (Z. Griliches and M. D. Intriligator, eds.), vol. 2 of *Handbook of Econometrics*, chap. 13. Elsevier, 775–826.

Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, **55** 251–276.

Engle, R. F. and Kozicki, S. (1993). Testing for common features. *Journal of Business & Economic Statistics*, **11** 369–380.

Engle, R. F., Ng, V. K. and Rothschild, M. (1990). Asset pricing with a factor-ARCH covariance structure: Empirical estimates for treasury bills. *Journal of Econometrics*, **45** 213–237.

Engle, R. F. and Susmel, R. (1993). Common volatility in international equity markets. *Journal of Business & Economic Statistics*, **11** 167–176.

Fang, Z. (2016). Optimal plug-in estimators of directionally differentiable functionals. Working paper.

Fang, Z. and Santos, A. (2015). Inference on directionally differentiable functions. Working paper.

Fisher, F. M. (1961). Identifiability criteria in nonlinear systems. *Econometrica*, **29** 574–590.

Fisher, F. M. (1966). *The Identification Problem in Econometrics*. McGraw-Hill.

Florens, J., Johannes, J. and Van Bellegem, S. (2011). Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory*, **27** 472–496.

Florens, J., Johannes, J. and Van Bellegem, S. (2012). Instrumental regression in partially linear models. *The Econometrics Journal*, **15** 304–324.

Folland, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*. 2nd ed. Wiley & Sons.

Gill, L. and Lewbel, A. (1992). Testing the rank and definiteness of estimated matrices with applications to factor, state-space and arma models. *Journal of the American Statistical Association*, **87** 766–776.

Giurcanu, M. C. (2012). Bootstrapping in non-regular smooth function models. *Journal of Multivariate Analysis*, **111** 78 – 93.

Gorman, W. M. (1981). Some Engel curves. In *Essays in the Theory and Measurement of Consumer Behaviour: In Honour of Sir Richard Stone*. Cambridge University Press, 7–29.

Granas, A. and Dugundji, J. (2003). *Fixed Point Theory*. Springer.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer.

Hall, P. and Horowitz, J. L. (1996). Bootstrap critical values for tests based on Generalized-Method-of-Moments estimators. *Econometrica*, **64** 891–916.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, New Jersey.

Hannan, E. (1970). *Multiple Time Series*. Wiley.

Hansen, B. E. (2015). Regression kink with an unknown threshold. *Journal of Business and Economic Statistics*. Forthcoming.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50** 1029–1054.

Harville, D. A. (2008). *Matrix Algebra From a Statistician's Perspective*. Springer.

Hirano, K. and Porter, J. (2012). Impossibility results for nondifferentiable functionals. *Econometrica*, **80** 1769–1790.

Hoderlein, S. and Winter, J. (2010). Structural measurement errors in nonseparable models. *Journal of Econometrics*, **157** 432–440.

Hong, H. and Li, J. (2015). The numerical Delta method. Working paper.

Hong, S. (2012). Inference in semiparametric conditional moment models with partial identification. Working paper, Tsinghua University.

Horn, R. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*. Cambridge University Press.

Horowitz, J. L. (2001). The bootstrap. In *Handbook of Econometrics V* (J. J. Heckman and E. Leamer, eds.). Elsevier, 3159–3228.

Horowitz, J. L. (2002). Bootstrap critical values for tests based on the smoothed maximum score estimator. *Journal of Econometrics*, **111** 141 – 167.

Huang, J. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics*, **26** 242–272.

Huang, J. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, **31** 1600–1635.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, **72** 1845–1857.

Jagannathan, R., Skoulakis, G. and Skoulakis, Z. (2002). Generalized methods of moments: Applications in finance. *Journal of Business & Economic Statistics*, **20** 470–481.

Jagannathan, R. and Wang, Z. (1996). The conditional CAPM and the cross-section of expected returns. *The Journal of Finance*, **51** 3–53.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, **59** 1551–1580.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press.

Kaido, H. and Santos, A. (2014). Asymptotically efficient estimation of models defined by convex moment inequalities. *Econometrica*, **82** 387–413.

Kleibergen, F. and Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, **133** 97 – 126.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of statistics*, **28** 1356–1378.

Komunjer, I. and Ng, S. (2011). Dynamic identification of dynamic stochastic general equilibrium models. *Econometrica*, **79** 1995–2032.

Kosorok, M. (2008a). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Kosorok, M. R. (2008b). Bootstrapping the Grenander estimator. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (N. Balakrishnan, E. A. Peña and M. J. Silvapulle, eds.), vol. 1. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 282–292.

Kress, R. (2013). *Linear Integral Equations*, vol. 82 of *Applied Mathematical Sciences*. 3rd ed. Springer, New York.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, **17** 1217–1241.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. **40** 694–726.

Lee, J. H. and Liao, Z. (2014). On standard inference for GMM with seeming local identification failure. Working paper.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer Verlag.

Lemenant, A., Milakis, E. and Spinolo, L. V. (2014). On the extension property of Reifenberg-flat domains. *Annales Academi Scientiarum Fennicæ*, **39** 51–71.

Lewbel, A. (1991). The rank of demand systems: Theory and nonparametric estimation. *Econometrica*, **59** 711–730.

Lewbel, A. (2006). Engel curves. In *The New Palgrave Dictionary of Economics*, vol. 2.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86** 316–327.

Liao, Y. and Jiang, W. (2011). Posterior consistency of nonparametric conditional moment restricted models. *The Annals of Statistics*, **39** 3003–3031.

Liao, Z. (2013). Adaptive GMM shrinkage estimation with consistent moment selection. *Econometric Theory*, **29** 857–904.

Liao, Z. and Phillips, P. (2015). Automated estimation of vector error correction models. *Econometric Theory*, **31** 581–646.

Linton, O., Song, K. E. and Whang, Y.-J. (2010). An improved bootstrap test of stochastic dominance. *Journal of Econometrics*, **154** 186 – 202.

Magnus, J. R. and Neudecker, H. (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. 3rd ed. John Wiley & Sons.

Matzkin, R. L. (2008). Identification in nonparametric simultaneous equations models. *Econometrica*, **76** 945–978.

Newey, W. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, **79** 147–168.

Newey, W. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, vol. 4. 2111–2245.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, **71** 1565–1578.

Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces*. Academic Press.

Patra, R. K., Seijo, E. and Sen, B. (2015). A consistent bootstrap procedure for the maximum score estimator. Working paper.

Phillips, P. C. B. (1989). Partially identified econometric models. *Econometric Theory*, **5** 181–240.

Ploberger, W. and Phillips, P. C. (2012). Optimal estimation under nonstandard conditions. *Journal of Econometrics*, **169** 258 – 265.

Politis, D. and Romano, J. (1992). A general resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectral density estimation. *The Annals of Statistics* 1985–2007.

Politis, D. and Romano, J. (1993). Nonparametric resampling for homogeneous strong mixing random fields. *Journal of Multivariate Analysis*, **47** 301–328.

Politis, D., Romano, J. and Lai, T. (1992). Bootstrap confidence bands for spectra and cross-spectra. *IEEE Transactions on Signal Processing*, **40** 1206–1215.

Portier, F. and Delyon, B. (2014). Bootstrap testing of the rank of a matrix via least-squared constrained estimation. *Journal of the American Statistical Association*, **109** 160–172.

Priestley, M. (1981). *Spectral Analysis and Time Series*. Adademic Press.

Rao, J. N. K. and Wu, C. F. J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, **80** 620–630.

Robin, J.-M. and Smith, R. J. (2000). Tests of rank. *Econometric Theory*, **null** 151–175.

Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica*, **56** 931–954.

Roehrig, C. S. (1988). Conditions for identification in nonparametric and parametric models. *Econometrica*, **56** 433–447.

Romano, J. P. and Shaikh, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, **138** 2786 – 2807.

Romano, J. P. and Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica*, **78** 169–211.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, **13** 341–360.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, **39** 577–591.

Santos, A. (2011). Instrumental variable methods for recovering continuous linear functionals. *Journal of Econometrics*, **161** 129–146.

Santos, A. (2012). Inference in nonparametric instrumental variables with partial identification. *Econometrica*, **80** 213–275.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, **26** 393–415.

Sargan, J. D. (1959). The estimation of relationships with autocorrelated residuals by the use of instrumental variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **21** 91–105.

Sen, B., Banerjee, M. and Woodroofe, M. (2010). Inconsistency of bootstrap: The Grenander estimator. *The Annals of Statistics*, **38** 1953–1977.

Severini, T. A. and Tripathi, G. (2012). Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *Journal of Econometrics*, **170** 491–498.

Shao, J. (1991). Second-order differentiability and jackknife. *Statistica Sinica*, **1** 185–202.

Shao, J. (1994). Bootstrap sample size in nonregular cases. In *Proceedings of the American Mathematical Society*, vol. 122. 1251–1262.

Shao, J. and Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics*, **17** 1176–1197.

Shapiro, A. (1990). On concepts of directional differentiability. *Journal of Optimization Theory and Applications*, **66** 477–487.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research*, **30** 169–186.

Shapiro, A. (2000). Statistical inference of stochastic optimization problems. In *Probabilistic Constrained Optimization* (S. P. Uryasev, ed.), vol. 49 of *Nonconvex Optimization and Its Applications*. Springer, 282–307.

Shintani, M. (2001). A simple cointegrating rank test without vector autoregression. *Journal of Econometrics*, **105** 337–362.

Song, K. (2014). Local asymptotic minimax estimation of nonregular parameters with translation-scale equivariant maps. *Journal of Multivariate Analysis*, **125** 136 – 158.

Stein, C. (1956). Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. 187–195.

Swensen, A. R. (2006). Bootstrap algorithms for testing and determining the cointegration rank in VAR models1. *Econometrica*, **74** 1699–1714.

Tao, J. (2014). Inference for point and partially identified semi-nonparametric conditional moment models. Working paper, University of Wisconsin-Madison.

Tao, T. (2012). *Topics in Random Matrix Theory*. American Mathematical Society.

Vakhania, N., Tarieladze, V. and Chobanyan, S. (1987). *Probability Distributions on Banach Spaces*. Dordrecht: Reidel.

van der Vaart, A. W. (1991). On differentiable functionals. *The Annals of Statistics*, **19** 178–204.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1990). Prohorov and continuous mapping theorems in the Hoffmann-Jørgensen weak convergence theory, with application to convolution and asymptotic minimax theorems. Tech. Rep. 157, Department of Statistics, University of Washington, Seattle.

van der Vaart, A. W. and Wellner, J. A. (1996a). *Weak Convergence and Empirical Processes*. Springer Verlag.

van der Vaart, A. W. and Wellner, J. A. (1996b). *Weak Convergence and Empirical Processes: With Applications to Statistics*. 1st ed. Springer Series in Statistics, Springer, New York.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54** 426–482.

West, K. (1997). Another heteroskedasticity-and autocorrelation-consistent covariance matrix estimator. *Journal of Econometrics*, **76** 171–191.