Essays on Asymptotic Methods in Econometrics

by

Tetsuya Kaji

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics and Statistics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

© Tetsuya Kaji, MMXVIII. All rights reserved.

The author hereby grants to MIT permission to reproduce and to
distribute publicly paper and electronic copies of this thesis document
in whole or in part in any medium now known or hereafter created.

Author .................................................................
Department of Economics
May 15, 2018

Certified by ..............................................................
Anna Mikusheva
Associate Professor of Economics
Thesis Supervisor

Certified by ..............................................................
Victor Chernozhukov
Ford International Professor of Economics
Thesis Supervisor

Accepted by .............................................................
Ricardo Caballero
Ford International Professor of Economics
Chairman, Departmental Committee on Graduate Studies
Essays on Asymptotic Methods in Econometrics

by

Tetsuya Kaji

Submitted to the Department of Economics
on May 15, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics and Statistics

Abstract

This thesis consists of three essays that contribute to statistical methods in econometrics.

Chapter 1 develops new theory of integrable empirical processes and applies it to outlier robustness analysis. A frequent concern in empirical research is to ensure that a handful of outlying observations have not driven the key empirical findings. This chapter constructs a formal statistical test of outlier robustness and provides its theoretical foundation. The key is to observe that statistics related to outlier robustness analyses are represented as $L$-statistics—integrals of empirical quantile functions with respect to sample selection measures—and to consider these elements in appropriate normed spaces. We characterize the asymptotic distribution of $L$-statistics and prove the validity of nonparametric bootstrap. An empirical application shows the utility of the proposed test.

Chapter 2 establishes the theory of weak identification in semiparametric models and provides an efficiency concept for weakly identified parameters. We first formulate the defining feature of weak identification as weak regularity, the asymptotic dependence of a parameter on the model score. While this feature deems consistent and equivariant estimation of a weakly regular parameter impossible, we show that there exists an underlying parameter that is regular and fully characterizes the weakly regular parameter. Using the minimal sufficient underlying regular parameter, we define weak efficiency for a weakly regular parameter through local asymptotic Rao-Blackwellization. Simulation shows that efficiency of popular estimators in linear IV models can be improved under heteroskedasticity.

Chapter 3 provides a method to account for estimation error in financial risk control. As accuracy of estimated risk is subject to estimation error, risk control based on estimated risk may fail to control the true, unobservable risk. We show that risk measures that give bounds to the probabilities of bad events can be effectively controlled by the Bonferroni inequality when the distributions of their estimators are known or estimable. We call such risk measures tail risk measures and show that they subsume Value-at-Risk and expected shortfall. An empirical application to portfolio risk management shows that a multiplier of 1.3 to 1.9 can control the true
risk probability of expected shortfall at 10%.

Thesis Supervisor: Anna Mikusheva
Title: Associate Professor of Economics

Thesis Supervisor: Victor Chernozhukov
Title: Ford International Professor of Economics
Acknowledgments

I would like to express my sincere gratitude to my advisor Anna Mikusheva for her continuous support, patience, encouragement, and deep insights. Her guidance helped me overcome numerous obstacles I faced through my research. Truthfully, I could not have imagined having a better advisor and mentor for my intellectual venture at MIT.

Besides my advisor, I would like to thank Isaiah Andrews and Victor Chernozhukov, as well as other faculty members at the Department of Economics, who provided me with candid opinions and inspirations whenever I needed. I am also grateful to my fellow students for their feedback, cooperation, enlightenment, and of course, friendship. Rachael Meager inspired me on many open problems in econometrics and educated me with her enormous knowledge on economics and statistics. I benefited substantially from honest and uncompromising discussions with Hyungjune Kang, who became a coauthor of the third chapter. Friends from the skating community brought and shared with me a countless number of enjoyable and memorable moments into an otherwise monotonous, stressful graduate life.

In addition, I would like to express my gratitude to the staff of the Department for their immense support. A special gratitude goes to Emily Gallagher, who kindly offered thorough emendation of my early drafts during her busiest time. The financial support from the Ito Foundation that made my academic pursuit possible is also gratefully acknowledged.

Lastly, and most importantly, none of this could have happened without my family. My mother, who never ceased to trust me and support me, with exuberant encouragement and unconditional love throughout my life. My father and my sister, who, in the most challenging time of our lives, made every effort to maintain a secure environment for me to concentrate on my study. This thesis was made possible because of you.
## Contents

1 Switching to the New Norm: From Heuristics to Formal Tests using Integrable Empirical Processes .................................................. 15

1.1 Introduction ............................................................................................................. 16

1.2 Setup and Motivation ............................................................................................... 22

1.2.1 \( L \)-Statistics ........................................................................................................ 22

1.2.2 Outlier Robustness Analysis .................................................................................. 24

1.2.3 Notes on the Setup ............................................................................................... 27

1.3 Overview of Main Results ......................................................................................... 28

1.3.1 Step 1: Convergence of Empirical Processes ....................................................... 30

1.3.2 Step 2: Convergence of Quantile Processes ......................................................... 34

1.3.3 Step 3: Convergence of \( L \)-Statistics .................................................................. 35

1.4 Application to Outlier Robustness Analysis ............................................................ 41

1.4.1 Test of Robustness to Outliers ............................................................................ 41

1.4.2 Empirical Application to the Effect of Democracy on Growth ........................... 42

1.5 Applications to Other Econometric Problems ......................................................... 50

1.5.1 Multiple Testing with Dependence ...................................................................... 50

1.5.2 Testing Higher Degree Lorenz Dominance ......................................................... 51

1.5.3 Controlling Tail Risk with Estimation Error ....................................................... 55

1.5.4 Assessing Outcome-Dependent Heterogeneity in Treatment Effects .................. 56

1.6 Literature Review .................................................................................................... 56

1.6.1 Empirical and Quantile Processes ...................................................................... 57
1.6.2 L-Statistics .................................................. 59
1.6.3 Outlier Detection and Robust Estimation .................. 59
1.7 Conclusion ..................................................... 60

Appendices ......................................................... 63
1.A Figures and Tables ............................................. 63
1.B Proofs of Results Stated in the Main Text ................. 69
1.C Mathematical Preliminaries .................................. 69
1.D Tightness of Bounded Integrable Processes ............... 72
1.E Differentiability of the Inverse Map ....................... 84
1.F Differentiability of L-Statistics ............................. 89
1.G Validity of Nonparametric Bootstrap ...................... 98
1.H Glivenko-Cantelli Type Results .......................... 104

Bibliography ....................................................... 111

2 Theory of Weak Identification in Semiparametric Models .... 119
2.1 Introduction .................................................. 120
2.2 Weak Identification in Economics ......................... 125
2.3 Overview of Main Results using Linear IV Models ....... 130
2.4 Weak Identification in Semiparametric Models .......... 135
  2.4.1 Weakly Regular Parameters ......................... 137
  2.4.2 Fundamental Impossibility ........................... 143
  2.4.3 Underlying Regular Parameters ...................... 144
2.5 Minimal Sufficient Underlying Regular Parameters ... 147
  2.5.1 Nuisance Tangent Spaces ......................... 148
  2.5.2 Sufficiency and Minimality of Underlying Regular Parameters 151
2.6 Weak Efficiency for Weakly Regular Parameters ....... 155
  2.6.1 Local Asymptotic Rao-Blackwellization ............ 161
  2.6.2 Weakly Efficient Estimators of Weakly Regular Parameters 163
2.7 Simulation of Weak Efficiency in Linear IV Models ...... 165
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.A.1</td>
<td>Empirical quantile functions and sample selection functions for OLS and IV estimators</td>
<td>64</td>
</tr>
<tr>
<td>1.A.2</td>
<td>Distributions of residuals of OLS and IV regressions</td>
<td>65</td>
</tr>
<tr>
<td>1.A.3</td>
<td>Joint distributions of full-sample and outlier-removed OLS and IV estimators</td>
<td>67</td>
</tr>
<tr>
<td>1.A.4</td>
<td>Histogram of numbers of removal that countries had in 10,000 iterations of bootstrap</td>
<td>68</td>
</tr>
<tr>
<td>2-1</td>
<td>Simulated distributions of 2SLS and Rao-Blackwellized 2SLS estimators under heteroskedasticity and availability of feasible GLS</td>
<td>134</td>
</tr>
<tr>
<td>2-1</td>
<td>Distributions of the 2SLS, GMM, and Fuller estimators, and their Rao-Blackwellization under heteroskedasticity (model A)</td>
<td>169</td>
</tr>
<tr>
<td>2-2</td>
<td>Distributions of the unbiased, Rao-Blackwellized unbiased, 2SLS, Fuller, Rao-Blackwellized Fuller, and HFUL estimators under heteroskedasticity (model B)</td>
<td>170</td>
</tr>
<tr>
<td>3-1</td>
<td>The daily returns of Bank of America Corp., Morgan Stanley, and the Dow Jones Industrial Averages</td>
<td>210</td>
</tr>
<tr>
<td>3-2</td>
<td>Expected shortfall of level 95% of the daily returns of Bank of America Corp., Morgan Stanley, and the Dow Jones Industrial Averages with historical simulation</td>
<td>212</td>
</tr>
<tr>
<td>3-3</td>
<td>The fitted daily volatility in the GARCH model for Bank of America Corp., Morgan Stanley, and the Dow Jones Industrial Averages</td>
<td>214</td>
</tr>
</tbody>
</table>
3-4 Expected shortfall of level 95% of the daily returns of Bank of America Corp., Morgan Stanley, and the Dow Jones Industrial Averages with filtered historical simulation 216
3.A.1 Distributions of estimated risk and estimated risk probability 223
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.A.1</td>
<td>Estimates and comparison of formal and heuristic $p$-values for the outlier robustness test</td>
<td>66</td>
</tr>
<tr>
<td>2.1</td>
<td>Mean absolute errors (MAEs) and mean squared errors (MSEs)</td>
<td>171</td>
</tr>
<tr>
<td>3.1</td>
<td>Expected shortfall of level 95% of the daily returns of Bank of America Corp., Morgan Stanley, and the Dow Jones Industrial Averages with historical simulation</td>
<td>211</td>
</tr>
<tr>
<td>3.2</td>
<td>Expected shortfall of level 95% of the daily returns of Bank of America Corp., Morgan Stanley, and the Dow Jones Industrial Averages with filtered historical simulation</td>
<td>216</td>
</tr>
</tbody>
</table>
Chapter 1

Switching to the New Norm: From Heuristics to Formal Tests using Integrable Empirical Processes

A frequent concern in empirical research is whether a handful of outlying observations have driven the key empirical findings. The current widespread practice in economics is to redo the statistical analysis adjusting for outliers and see if we obtain similar results, checking “robustness to outliers.” However, such empirical practices have little theoretical justification, and researchers have had to rely on heuristic arguments. This chapter constructs a formal statistical test of outlier robustness that accommodates many empirical settings. The key is to observe that statistics related to outlier robustness analysis are represented as $L$-statistics—integrals of empirical quantile functions with respect to sample selection measures—and to consider them in spaces equipped with appropriate norms. In particular, we characterize weak convergence of empirical distribution functions in the space of bounded integrable functions, establish the delta method for their inverses (empirical quantile functions) as maps from this space into the space of integrable functions, characterize weak convergence of random sample selection measures in the space of bounded integrable Lipschitz functions, and derive the delta method for $L$-statistics as maps from those spaces into a Euclidean space. As an empirical application, we revisit the outlier robustness analysis in Acemoglu et al.
and demonstrate that our test can detect sensitivity of a parameter that was otherwise indiscernible had we relied on existing heuristics. Our theory of $L$-statistics is new and of independent interest; we propose other applications, including multiple testing problems and tests of higher-order Lorenz dominance.

1.1 Introduction

In empirical research in economics, a common concern is whether a handful of outlying observations may have driven crucial empirical findings. In estimating the effect of microcredit with experimental data, Augsburg et al. (2015) report that trimming 1% of the observations makes the effect of the loan program on business profits significant that is otherwise insignificant. The analysis of Acemoglu et al. (2001) on the effect of institutions on economic performance using differences in mortality rates prompted extensive discussion of whether outliers undermined the validity of mortality rates as instruments (Albouy, 2012; Acemoglu et al., 2012). Herndon et al. (2014) discuss whether the exclusion of some observations invalidates the findings in Reinhart and Rogoff (2010). Guthrie et al. (2012) find that a result in Chhaochharia and Grinstein (2009) is driven by outliers. De Long and Summers (1991) and de Long et al. (1992) find that machinery and equipment investment have a strong connection with economic growth, which is followed by discussion of whether outliers drove their findings (Auerbach et al., 1994; de Long and Summers, 1994). Toda and Walsh (2015) report that removing 200 outliers from 410,788 observations drastically changes the estimate of the risk aversion coefficient.

It is thus considered an important characteristic of valid empirical findings that a small number of outliers do not affect the conclusion of analysis to a nonnegligible degree (Young, 2017). The common practice in empirical research is to carry out robustness checks by redoing the analyses on the sample that is adjusted for outliers (such as removal or winsorization) and comparing the results from the original ones relative to standard errors (e.g., Acemoglu et al., 2001, 2016, 2017; Agarwal et al., 2010; Alatas et al., 2016; Fabrizio et al., 2007). While such heuristic practices lack for-
mal justification (as explained in Section 1.2.2), it is technically demanding to obtain the joint distribution of the estimates necessary to formalize the outlier robustness checks as statistical tests.

The main contribution of this chapter is to develop a method to derive the joint distribution of full-sample and outlier-adjusted estimators for a wide range of sample selection procedures, including removal or winsorization at cutoffs that depend on the entire sample. With our results, we can test whether the parameter of interest changes its value significantly before and after such sample selection, enabling formal statistical investigation of outlier robustness checks. Many statistics related to outlier robustness analysis are represented as $L$-statistics—integrals of transformations of empirical quantile functions with respect to random sample selection measures. We develop a new empirical process theory tailored for these statistics, with an important innovation related to the choice of appropriate norms. Despite the long tradition of empirical process techniques in establishing asymptotic properties of $L$-statistics [Shorack and Wellner 1986; Van der Vaart and Wellner 1996; Koul 2002], the literature has confined attention to empirical processes under the “uniform norm,” which has imposed severe limitations to the range of applications; in particular, it did not cover some essential $L$-statistics that appear in outlier robustness analysis. In contrast, our theory employs appropriate norms and allows us to cover a very general form of $L$-statistics including them.

Our theoretical contribution consists of three key results: we consider empirical distribution functions in the space of bounded integrable functions and characterize weak convergence therein (Section 1.3.1); we consider empirical quantile functions in the space of integrable functions and establish the functional delta method for the map from distribution functions to quantile functions in these spaces (Section 1.3.2); we consider random sample selection measures in the space of bounded integrable Lipschitz functions and establish a functional delta method for $L$-statistics from these spaces (Section 1.3.3). Lastly, we obtain the formula for the joint asymptotic distribution of $L$-statistics and establish the validity of bootstrap for computing the asymptotic distribution.
The challenge in deriving asymptotic distributions of outlier-adjusted statistics is that the sample selection procedure often depends on the whole of the sample in ways that render classical multivariate central limit theorems inapplicable. As the first step of our analysis, we observe that many statistics related to outlier robustness analysis are given by $L$-statistics. An $L$-statistic is a quantity given by

$$\int_0^1 m(Q_n(u))dK_n(u),$$

where $m$ is a known continuously differentiable function, $Q_n : (0, 1) \to \mathbb{R}$ an empirical quantile function of some random variable $X_i$, and $K_n : (0, 1) \to \mathbb{R}$ a Lipschitz function that is possibly random. Here, $K_n$ represents the sample selection procedure (such as outlier removal or winsorization) that can be heavily dependent on the quantile function and other observations.

As a toy example, let us consider the problem of deriving the joint distribution of two sample means: the full sample mean and the $\alpha$-trimmed sample mean, the mean that drops $\alpha \in (0, 1)$ portions of observations from both tails,

$$\frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{and} \quad \frac{1}{n - 2[\alpha n]} \sum_{i=[\alpha n]+1}^{n-\lfloor\alpha n\rfloor} X_{(i)},$$

where $X_{(i)}$ denotes the $i$th smallest order statistic of $X_1, \ldots, X_n$. Surprisingly, deriving the joint distribution of these statistics is a nontrivial problem, as the order statistics are highly dependent, and the trimmed mean cannot be represented as a simple sum of i.i.d. (or stationary) random variables, preventing the use of familiar central limit theorems. We tackle this problem by transforming them into the integral forms:

$$\int_0^1 Q_n(u)du \quad \text{and} \quad \frac{n}{n - 2[\alpha n]} \int_{[\alpha n]/n}^{1-\lfloor\alpha n\rfloor/n} Q_n(u)du,$$

where $Q_n$ is the empirical quantile function of $X_i$.

---

1. The empirical quantile function is a generalized inverse of the empirical distribution function, in particular, $Q_n(u) := \inf\{x \in \mathbb{R} : F_n(x) \geq u\}$ where $F_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\}$.

2. Note that the trimmed mean can be further written as $\int_0^1 Q_n dK_n$ for $K_n$ Lipschitz. See Section 1.2.
functional delta methods, once we know “weak convergence” of the empirical quantile process $\sqrt{n}(Q_n - Q)$, where $Q : (0, 1) \to \mathbb{R}$ is the population quantile function of $X_i$.

However, now we face a major difficulty; the empirical quantile process thus defined does not converge in the standard sense. If $X_i$ is supported on the whole of $\mathbb{R}$, the true quantile function $Q$ is unbounded on $(0, 1)$. On the other hand, the empirical quantile function $Q_n$ is by construction bounded on $(0, 1)$ since for each $n$ there are only finitely many values $Q_n$ can take (in particular, $X_1$ through $X_n$). It is clear, then, that the maximum distance between $Q_n$ and $Q$ is infinity for every $n$, implying that $Q_n$ does not converge to $Q$ uniformly. Corresponding to this point, Van der Vaart (1998, p. 317) notes that the functional delta method for $L$-statistics “is preferable in that it applies to more general statistics, but it...does not cover the simplest $L$-statistic: the sample mean.” To circumvent this issue, the literature on empirical processes has often confined itself to bounded or truncated quantile functions (Van der Vaart and Wellner, 1996, Chapter 3.9) or weighted empirical quantile processes that suitably down-weight the tails (Csörgő and Horváth, 1993). However, none of these methods work for our purpose, as we neither want to limit attention to bounded random variables nor can we expect that the random variables are weighted in such a nice way.

We solve this problem by considering the quantile functions in the $L_1$ space instead of in the traditional $L_\infty$ space. The important point is to realize that what we truly need is the convergence of integrals of quantile functions; uniform convergence of quantile functions, as often considered, is neither necessary nor sufficient for our purpose. In light of this, we first characterize weak convergence of empirical distribution processes in the space of bounded and integrable functions, and then establish the functional delta method for the map from such distribution functions $F$ to quantile functions $Q = F^{-1}$; this establishes weak convergence of empirical quantile processes in $L_1$. The key intuition in the proof is to observe that the $L_1$ norm is compatible with Fubini’s theorem.

Given weak convergence of empirical quantile processes, we now proceed to weak convergence of the possibly random sample selection function $K_n$ and the functional
delta method for \((Q, K) \mapsto \int QdK\). Note that

\[
\int Q_n d\mathbb{K}_n - \int QdK = \int (Q_n - Q)d\mathbb{K}_n + \int Qd(\mathbb{K}_n - K).
\]

For the first term to converge whenever the sample average \(\int Q_n du\) does, we need \(\mathbb{K}_n\) to be uniformly Lipschitz. For the second to be well-defined, we need that \(\mathbb{K}_n\) converges to \(K\) in \(L_\infty\). Then by integration by parts, the second term can be approximately written as \(-\int (\mathbb{K}_n - K)dQ\), meaning that \(\mathbb{K}_n\) needs to converge to \(K\) in \(L_Q\), the space of functions integrable with respect to \(Q\). This exercise reveals that the appropriate convergence of the sample selection function \(\mathbb{K}_n\) can be established in, again, the space of bounded and integrable functions. Finally, we establish the functional delta method for the \(L\)-statistic, \((Q, K) \mapsto \int QdK\) (more precisely, we allow transformations of quantile functions, \(\int m(Q)dK\)).

The utility of our functional delta method approach is not only the generality it brings but also that it implies the validity of the nonparametric bootstrap. This allows researchers to avoid making strong distributional assumptions to derive the asymptotic distributions of their estimators.

This chapter characterizes the asymptotic distributions of many \(L\)-statistics in the form of Gaussian distributions. Note, however, that not all \(L\)-statistics converge weakly to Gaussian distributions; for example, the largest order statistic, appropriately scaled, often converges to some extreme value distribution \cite{deHaanFerreira_2006}. In this sense, this chapter can be seen as establishing the conditions under which a general form of \(L\)-statistics converges to a Gaussian distribution. The key to convergence toward Gaussian distributions is that the sample selection mechanisms become less and less dependent on \(n\) as \(n\) tends to infinity; in the outlier removal example, the threshold \(\alpha\) does not approach 0 as \(n\) tends to infinity. This assumption, however, may not be plausible in some applications. We note that our delta method results are potentially susceptible to generalizations to other nonstandard distributions; see Section 1.3.3.

Applying the theory developed thus far, we propose a test of outlier robustness
that takes into account natural comovement of the two estimators. As an empirical application, we revisit the outlier robustness analysis discussed in Acemoglu et al. (2017) and carry out a formal statistical test as proposed in this chapter. They estimate the effect of democracy on the GDP growth, and examine the sensitivity of their results to the removal of outliers on the residuals, in particular examining whether removing the extreme values in GDP growth would significantly change their findings. For all but one coefficient, the test is not rejected at 5% level, meaning that they exhibit robustness to such outlier removal. For the one rejected—persistence of the GDP growth—we show that the rejection would not have been “detectable” had we relied on the heuristic testing procedure commonly practiced in the literature.

The theory of $L$-statistics developed in this chapter is itself new and of independent interest; it can be used to solve other econometric problems aside from outlier robustness analyses. Chapter 3 defines a class of risk measures subsuming Value-at-Risk and expected shortfall that can incorporate estimation errors into the risks being estimated. The asymptotic results in their paper use the theory developed in this chapter. Kaji (2017) interprets quantile treatment effects as individual treatment effects that attain the lower bound of the total absolute treatment effect and proposes a variant of subgroup treatment effects to assess the heterogeneity of treatment effects. Again, the asymptotic properties follow from the results of the present chapter. In the main text, we also discuss applications to multiple testing problems and tests of higher-order Lorenz dominance.

The rest of this chapter is organized as follows. Section 1.2 defines the class of $L$-statistics considered in this chapter and discusses how it subsumes many statistics widely used in economics. Section 1.2 also elaborates on the outlier robustness analysis and explains how outlier robustness can be tested using the asymptotic distributions of $L$-statistics. Section 1.3 describes the main theoretical contribution of the chapter; it develops the asymptotic theory of $L$-statistics using integrable empirical processes and functional delta methods. The exposition is aimed to be minimal and intuitive, leaving most of the details to Appendices. Section 1.4 discusses an approach for testing outlier robustness and revisits the outlier robustness analysis of
Section 1.5 applies the asymptotic theory of $L$-statistics to other econometric problems: multiple testing problems, tests of higher-order Lorenz dominance, tail risk measures in Chapter 3, and heterogeneous treatment effects by Kaji (2017). Section 1.6 reviews the related literature on empirical processes, $L$-statistics, and robust estimation. Finally, Section 1.7 concludes. All figures, tables, and proofs appear in the Appendices.

1.2 Setup and Motivation

1.2.1 $L$-Statistics

This chapter concerns statistics that are averages of functions of independent and identically distributed (i.i.d.) random variables, where each observation may be omitted or weighted differently from other observations. To fix this idea, let $X_i$ be an i.i.d. scalar random variable and $w_i$ a possibly random weight whose distribution is assumed to be bounded but is allowed to depend on all of $X_i$. Consider a statistic of the form

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} m(X_i)w_i,$$

where $m$ is some continuously differentiable function. For example, the sample average is such a statistic where $m$ is an identity and $w_i$ is identically one; the sample average from 1st to 99th percentiles (excluding the bottom and top 1% of observations) is also such a statistic for which $m$ is an identity and $w_i$ is the indicator of whether $X_i$ falls between the 1st and 99th percentiles of $X_1, \ldots, X_n$. Thus, $w_i$ captures the idea that each $X_i$ may be weighted (or excluded) in an interdependent way.

Note that rearranging the summands does not affect the sum itself. In particular, let $X_{(i)}$ be the order statistic of $X_i$; $X_{(1)}$ represents the smallest observation, $X_{(2)}$ the second smallest, and so on. Denote by $w_{(i)}$ the weight corresponding to $X_{(i)}$ (so $w_i$ is sorted according to the order of $X_i$). Then, one may rewrite the average without loss
of generality as
\[ \hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} m(X_{(i)})w_{(i)}. \]

This formulation is known as an \textit{L-statistic}, where “L” stands for the fact that \( \hat{\beta} \) is a \textit{linear} combination of functions of order statistics \( X_{(i)} \). Many statistics commonly used in economics are \( L \)-statistics, as will be shown in examples below.

We develop a method to derive the distribution of \( \hat{\beta} \) using empirical quantile functions. Let \( Q_n(u), u \in (0, 1) \), be the empirical \( u \)-quantile of \( X_i \), that is,

\[
Q_n(u) := \begin{cases} 
X(1) & u \in (0, \frac{1}{n}], \\
X(i) & u \in (\frac{i-1}{n}, \frac{i}{n}], \\
X(n) & u \in (\frac{n-1}{n}, 1]. 
\end{cases}
\]

Using this, one can write
\[
\hat{\beta} = \int_0^1 m(Q_n(u)) dK_n(u),
\]

where \( K_n \) is the measure that assigns density \( w_{(i)} \) to \( u \in (\frac{i-1}{n}, \frac{i}{n}] \). Although the two representations are mathematically equivalent, the first representation as a sum of order statistics evokes the multivariate central limit theorems, while the second representation as an integral evokes the functional central limit theorems and functional delta methods. Correspondingly, there are two methods to derive the asymptotic distribution of \( L \)-statistics—the Hájek projection and the functional delta method—each of which covers nonoverlapping quantities \cite{van_derVaart1998} Chapter 22). For example, the Hájek projection covers the full sample average, while the functional delta method covers plug-ins of estimated quantile functions. However, nonoverlapping coverage can be problematic when we want the \textit{joint} distribution of various \( L \)-statistics, as in the outlier robustness analysis. While we leave further comparison of the two methods to Section 1.6.2, this chapter achieves substantial generalization of the second method that is enough to accommodate quite general forms of \( L \)-statistics useful for outlier robustness analyses and other problems.
To wrap up, our objective is to derive the joint distribution of finitely many statistics of the form $\hat{\beta} = \int_0^1 m(Q_n) dK_n$; in particular, $\hat{\beta}_j = \int_0^1 m_j(Q_{n,j}) dK_{n,j}$, $j = 1, \ldots, d$.

1.2.2 Outlier Robustness Analysis

This section clarifies the motivation of outlier robustness analysis and explains why $L$-statistics are useful for this purpose.

What is the problem in current heuristic practice? Let $\hat{\beta}_1$ be the key estimator on which our empirical findings are based. When we want to claim that $\hat{\beta}_1$ is “not the consequence of only a few outlying observations,” we often compute another estimator $\hat{\beta}_2$ from the sample that excludes some outliers and then argue that their difference is not too large compared to the standard error estimated for $\hat{\beta}_1$. However, comparing the difference $\hat{\beta}_1 - \hat{\beta}_2$ to the marginal standard error of $\hat{\beta}_1$ does not make much sense from a statistical point of view. Naturally, $\hat{\beta}_1$ and $\hat{\beta}_2$ are based on almost identical sets of observations. Therefore, even if the contribution of outliers is fairly large, the difference of the two estimators can be, by construction, much smaller than the marginal standard error of $\hat{\beta}_1$. Moreover, it so happens that the asymptotic distribution of an efficient estimator is independent of its difference from another estimator; then, such empirical practices may not be susceptible to an interpretation as a meaningful statistical testing procedure of some hypothesis.

What does a researcher want to investigate by checking “robustness” to outliers? As an example, consider the problem of estimating the treatment effect of microcredit provision on households’ business profits in rural villages in some country. Let $\beta_1$ be the true average treatment effect and suppose that its estimate $\hat{\beta}_1$ is significantly positive. We may then suggest policy implications such as “Since $\hat{\beta}_1$ is significantly positive, we recommend to expand availability of microcredit to all villages in this country.” However, we are worried that such a finding may be mostly driven by some “outlying” observations. For example, we are concerned about the possibility that the treatment effects are largely positive for above-the-poverty-line households while they can be negative for poor or extremely poor households, aggregating to a modestly
positive average treatment effect. If this is the case, despite the average effect being positive, we may not wish to implement microcredit as it may exacerbate economic inequalities. In another scenario, we may be concerned that some extreme data points are not representative of the true population; for example, some respondents with limited literacy may have mistakenly answered their incomes as unreasonably high (or low) figures, and that may be driving the treatment effect unreasonably positive and significant. If so, again, we may not wish to base our policy recommendations on such imprecise measures.

In this setting, let \( X_i \) be (a part of) household \( i \)'s characteristics; \( X_i \) can be a regressor or can be a dependent variable. We are worried about the robustness of our findings to outliers of \( X_i \); let \( \beta_2 \) be the true average treatment effect on the population that excludes the outlying portion of \( X_i \), e.g., \( \mathbb{E}[Y_{i1} - Y_{i0} \mid X_i \leq c] \). Concerns about the first scenario can be formulated as “the average effect \( \beta_1 \) does not represent the average effect among ‘typical’ individuals, \( \beta_2 \).” Then, the null hypothesis subject to be tested in the outlier robustness analysis can be formulated as

\[
H_0 : |\beta_1 - \beta_2| \leq h
\]

for some \( h \geq 0 \). In the second scenario, we are concerned that outliers may not be from the true data generating process of interest, and they may be affecting the estimate too much. However, if outliers affect the findings of the statistical analyses only to a negligible degree, then we may say that our findings are robust to such possibilities. Then, the null hypothesis we want to test is, again, \( H_0 : |\beta_1 - \beta_2| \leq h \) for some \( h \geq 0 \).

The choice of \( h \) in the null hypothesis is an important practical question, but we treat it as given in this chapter. This \( h \) should be based on how much error can be tolerated in applying the empirical findings, and hence should be determined on a case-by-case basis.\(^3\) That being said, we list a few special cases later in this section where the choice of \( h \) is necessarily determined by the characteristics of the model.

\(^3\)In the empirical application in Section 1.4, we use the severest null \( h = 0 \). If one cannot reject the hypothesis with \( h = 0 \), that can be considered a “strong” indicator of robustness.
To summarize, if we develop a way to test the above hypothesis, we can formalize many heuristic arguments carried out in empirical research.\footnote{One may wish to “test” whether the outliers affect the significance of the estimates. However, significance depends by construction on data and hence is not solely determined by the population characteristics; therefore, bringing it up in the null hypothesis is difficult to justify from a statistical point of view.}

To relate $L$-statistics to our context, consider the regression equation

$$y_i = x_i \beta + \varepsilon_i, \quad \mathbb{E}[x_i \varepsilon_i] = 0,$$

where we estimate $\beta$ by the ordinary least squares (OLS) regression. Let us first consider cases where we compare another OLS estimator that excludes outliers of $y_i$ as in Acemoglu et al. (2016) or Banerjee et al. (2014). Now we have two estimators:

$$\hat{\beta}_1 = \left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i y_i, \quad \hat{\beta}_2 = \left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 w_i \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i y_i w_i, \quad (1.1)$$

where $w_i = 1 \{ y(\lfloor \tau n \rfloor + 1) \leq y_i \leq y(n - \lfloor \tau n \rfloor) \}$. Denote by $F_n$ and $Q_n$ the empirical distribution and empirical quantile functions of $x_i y_i$. Then, they can also be written as

$$\hat{\beta}_1 = \int_0^1 Q_n(u) dK_{n,1}(u), \quad \hat{\beta}_2 = \int_0^1 Q_n(u) dK_{n,2}(u),$$

where $K_{n,1}$ and $K_{n,2}$ are random measures that assign, respectively, density $(\frac{1}{n} \sum_{i=1}^{n} x_i^2)^{-1}$ to $(0, 1)$ and density $(\frac{1}{n} \sum_{i=1}^{n} x_i^2 w_i)^{-1} w_i$ to $u \in (F_n(x_i y_i) - 1/n, F_n(x_i y_i)]$. Along the same line, we can represent the two-stage least squares (2SLS) estimators as $L$-statistics as well.

We might instead think that outlying observations have some information and want to winsorize $x_i$ as in Acemoglu et al. (2012). Here, winsorization of $x_i$ at quantile $\tau$ means replacing every $x(i)$ for $i = 1, \ldots, \lfloor \tau n \rfloor$ by $x(\lfloor \tau n \rfloor + 1)$, and every $x(i)$ for $i = n - \lfloor \tau n \rfloor + 1, \ldots, n$ by $x(n - \lfloor \tau n \rfloor)$. Thus, winsorization replaces “outliers” with the closest value that is considered non-outlier. Then, we would have

$$\hat{\beta}_1 = \left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i y_i, \quad \hat{\beta}_2 = \left( \frac{1}{n} \sum_{i=1}^{n} x_i^2 w_i^2 \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i y_i w_i, \quad (1.2)$$
where

\[
 w_i = \begin{cases} 
 x_{(\lfloor \tau_n \rfloor + 1)}/x_i & x_i < x_{(\lfloor \tau_n \rfloor + 1)} < 0, \\
 x_{(n-\lfloor \tau_n \rfloor)}/x_i & x_i > x_{(n-\lfloor \tau_n \rfloor)} > 0, \\
 1 & \text{otherwise}.
\end{cases}
\]

These can be written as

\[
 \hat{\beta}_1 = \int_0^1 Q_n(u) d\mathbb{K}_{n,1}(u), \quad \hat{\beta}_2 = \int_0^1 Q_n(u) d\mathbb{K}_{n,2}(u),
\]

where \( \mathbb{K}_{n,1} \) and \( \mathbb{K}_{n,2} \) are, again, random measures that assign density \( (\frac{1}{n} \sum_{i=1}^n x_i^2)^{-1} \) and density \( (\frac{1}{n} \sum_{i=1}^n x_i^2 w_i^2)^{-1} w_i \) to \( u \in (F_n(x_i y_i) - 1/n, F_n(x_i y_i)) \). If we can derive the joint distribution of the involved \( L \)-statistics, we are able to formally test our hypothesis about the outlier robustness.

Now let us look at a few special cases of linear regression models in which outlier removal will not cause the coefficient to change. First, if we have \( \mathbb{E}[\varepsilon_i | x_i] = 0 \), then any sample selection conditional on \( x_i \) does not change the value of \( \beta \). Therefore, outlier removal based on \( x_i \) is harmless, and we can use \( h = 0 \). Second, if the conditional distribution of \( \varepsilon_i \) conditional on \( x_i \) is symmetric around zero and we remove samples symmetrically by \( \varepsilon_i \), it will not cause any bias on \( \beta \) (in reality, we remove by \( \hat{\varepsilon}_i \), which consistently estimates \( \varepsilon \)). Third, if \( \varepsilon_i \) is independent of \( x_i \), then the sample selection based on \( \varepsilon_i \) does not introduce bias on \( \beta \) except for the intercept. However, if we select samples based on \( y_i \), the true value of \( \beta \) will almost always change.\[5\]

### 1.2.3 Notes on the Setup

The \( L \)-statistics introduced so far share an important feature that the random measure \( \mathbb{K}_n \) is “well-behaved” (in the sense defined precisely in the next section). The intuition is that the selection or weighting mechanism does not depend on the sample size \( n \), at least asymptotically. The results of this chapter apply in such contexts. The next

---

\[5\] Or, if one regards \( \beta \) as a fixed structural parameter, then it can be put as “the plim of popular estimators does not coincide with the structural \( \beta \) any more.”
Example (Extreme order statistics). The minimum of the observations $X_1, \ldots, X_n$ can be written as

$$X_{(1)} = \frac{1}{n} \sum_{i=1}^{(1/n)n} nX(i) = \int_0^1 Q_n(u) dK_n(u),$$

where $Q_n$ is the empirical quantile of $X_i$, and $K_n$ assigns density $n$ on $(0, 1/n]$ and zero elsewhere. Then $K_n$ “converges” to the measure that assigns mass 1 to $u = 0$ and zero elsewhere, which is not absolutely continuous with respect to the Lebesgue measure.

1.3 Overview of Main Results

This section describes the key ideas and theoretical contributions of this chapter. The formal mathematical development and proofs are given in the Appendices.

We recall our setup from Section 1.2.1. Let $(X_{i,1}, \ldots, X_{i,d})$ be an i.i.d. random vector and $(w_{i,1}, \ldots, w_{i,d})$ vector of possibly random weights whose distribution is bounded but allowed to depend on all of $X_{i,j}$. Denote by $F_{n,j}$ the (marginal) empirical distribution function of $X_{1,j}, \ldots, X_{n,j}$. We want to know the joint distribution of

$$\hat{\beta}_j = \int_0^1 m_j(Q_{n,j}(u)) dK_{n,j}(u), \quad j = 1, \ldots, d,$$

where $Q_{n,j} := F_{n,j}^{-1}$ denotes the generalized inverse of $F_{n,j}$, $m_j : \mathbb{R} \to \mathbb{R}$ are continuously differentiable functions, and $K_{n,j}$ possibly random measures. We derive this using the asymptotic distributions of $F_{n,j}$ and $K_{n,j}$ and applying the corresponding functional delta methods for the map $(F, K) \mapsto \int_0^1 m(F^{-1})dK$.

What would be a plausible derivative formula for the delta method? Let $Q$ and $K$ be the population counterparts of $Q_n$ and $K_n$ and suppress dependence on $j$. Informal

---

6This does not mean that extension to such cases is impossible. See the end of Section 1.3.
calculation suggests that
\[
\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}\left(\int_{0}^{1} m(Q_n) dK_n - \int_{0}^{1} m(Q) dK\right)
\]
\[
= \int_{0}^{1} \sqrt{n}[m(Q_n) - m(Q)] dK_n + \int_{0}^{1} m(Q) d\left(\sqrt{n}(K_n - K)\right)
\]
\[
\approx \int_{0}^{1} \sqrt{n}[m(Q_n) - m(Q)] dK - \int_{0}^{1} \sqrt{n}(K_n - K) dm(Q)
\]
\[
\approx \int_{0}^{1} m'(Q) \sqrt{n}(Q_n - Q) dK - \int_{0}^{1} m'(Q) \sqrt{n}(K_n - K) dQ,
\]
where the third “equality” follows from integration by parts and the fourth from a
delta method. One of the main goals of this chapter is to give the conditions under
which this derivation can be justified. The purpose of this section is to provide an
accessible introduction to the issues involved, while the rigorous treatment is left to
the Appendices.

We proceed in three steps:

Step 1. Explore in what sense the empirical distribution function \( F_n \) must converge,
and give sufficient conditions for such convergence. Along the way, we will
also find the right notion of convergence for the empirical quantile function
\( Q_n := \bar{F}_n^{-1} \).

Step 2. Under the stated conditions, show that functions of the empirical quantile
function \( Q_n \) do indeed converge in the required sense, and characterize its
asymptotic distribution. The key is the functional delta method for \( F \mapsto m(Q) = m(F^{-1}) \).

Step 3. Formulate a proper convergence notion for \( K_n \). Combining these results,
show that our \( L \)-statistics converge to a normal random vector, and obtain
its formula. The key is the functional delta method for \( (Q, K) \mapsto \int_{0}^{1} m(Q) dK \).
1.3.1 Step 1: Convergence of Empirical Processes

The empirical process literature (Shorack and Wellner, 1986; Van der Vaart and Wellner, 1996; Kosorok, 2008; Dudley, 2014) shows that the classical empirical process \( \sqrt{n}(F_n - F) \) converges to a Gaussian process in \( L_\infty \). Due to the choice of this norm (the uniform norm), such results are referred to as uniform central limit theorems. To proceed with our agenda, however, such classical results turn out to be insufficient.

To understand the difficulty we face, consider the empirical quantile process in analogy with the empirical process for distribution functions,

\[
\sqrt{n}(Q_n(u) - Q(u)), \quad u \in (0,1).
\]

If the support of the underlying distribution \( F \) is unbounded (which is necessary to accommodate many empirically relevant problems in economics), the true quantile function \( Q \) is an unbounded function on \((0,1)\), while the empirical quantile function \( Q_n \) is bounded for every \( n \) by construction. Therefore, it immediately follows that the empirical quantile process, with no restrictions on its range, never converges in the traditional uniform sense. This is why the previous literature has restricted its attention to convergence of quantile processes of truncated or bounded random variables (Van der Vaart, 1998; Van der Vaart and Wellner, 1996) or of weighted versions of quantile processes so they are effectively bounded (Csörgő and Horváth, 1993).

The first key idea of this chapter is to switch to a new norm on the space of quantile functions. Recall that our eventual target is the statistics represented by the integral of the empirical quantile functions; we are not interested in any kind of inference that requires uniform convergence of \( Q_n \) (such as Kolmogorov-Smirnov type tests or uniform confidence bands around a quantile function). Then, the appropriate space for our quantile functions would naturally be the space of integrable functions, and the corresponding notion of convergence be \( L_1 \); the classical uniform norm \( L_\infty \) appears neither appropriate nor desirable. Thus, we give up the uniform convergence and seek the conditions under which the empirical quantile process \( \sqrt{n}(Q_n - Q) \)
converges weakly in $L_1$. In light of this, define the following space.

**Definition.** Let $\mathbb{B}$ be the Banach space of measurable functions $z$ from $(0, 1)$ to $\mathbb{R}$ with the norm

$$
\|z\|_{\mathbb{B}} := \int_0^1 |z(u)| du.
$$

Not all probability distributions have a quantile function that is integrable. Precisely, a quantile function is integrable if and only if the corresponding probability distribution has a finite first moment (Lemma \[1.C.1\]). One sees therefore that even if the empirical distribution function $F_n$ converges to the true distribution function $F$ in the uniform sense (which is indeed the case for every probability distribution regardless of how many moments it has), it might not be the case that the empirical quantile function $Q_n$ converges to the true quantile function $Q$ in $L_1$. In other words, the inverse map $F \mapsto Q := F^{-1}$, when viewed as a map from $L_\infty$ to $L_1$, is not even continuous, let alone differentiable. This is why the classical uniform central limit theorems are not suitable for our purpose; we need to make use of a sufficiently strong norm on the space of distribution functions that ensures the existence of at least the first moment. Put together, the norm must be strong enough that the inverse map $F \mapsto Q$ be differentiable, but not so strong that it excludes many distributions of our potential interest.

The second key idea of this chapter is to observe that the integrability of quantile functions is equivalent to the integrability of distribution functions by integration by parts. In particular, we require the distribution function $F$ to be “integrable” in the sense that its modification

$$
\tilde{F}(x) = \begin{cases} 
F(x) & x < 0 \\
F(x) - 1 & x \geq 0 
\end{cases}
$$

is integrable. The adequacy of this norm is intuitively understood by observing that the quantile function is integrable if and only if the modification of the distribution function is integrable (Lemma \[1.C.1\]).

The precise definition of the norm is as follows.
**Definition.** Let $-\infty \leq a < c < b \leq \infty$ and $\mu$ be a positive Lebesgue-Stieltjes measure on $(a, b)$. Define the space $L_\mu$ of $\mu$-measurable functions $z : (a, b) \to \mathbb{R}$ with limits $z(a) := \lim_{x \to a} z(x)$ and $z(b) := \lim_{x \to b} z(x)$, and the norm
\[
\|z\|_{L_\mu} := \|z\|_\infty \vee \|	ilde{z}\|_\mu := \left( \sup_{x \in (a, b)} |z(x)| \right) \vee \left( \int_a^b |\tilde{z}(x)| d\mu(x) \right)
\]
where
\[
\tilde{z}(x) := \begin{cases} 
  z(x) - z(a) & x < c, \\
  z(x) - z(b) & x \geq c.
\end{cases}
\]

**Definition.** Let $L$ be the special case of $L_\mu$ where $(a, b, c) = (-\infty, \infty, 0)$ and $\mu$ be equal to the Lebesgue measure. The space of distribution functions is the subset $L_\phi$ of $L$ of functions $z$ that are monotone and cadlag with $z(-\infty) = 0$ and $z(+\infty) = 1$.

Note that we still require the distribution function to converge uniformly (the $L_\infty$ part of the norm); this ensures that the “inverse function” is well defined. Being the intersection of the familiar spaces $L_\infty$ and $L_1$, weak convergence in $L$ implies convergence in both.

Henceforth we will focus on distributions $F$ that are members of $L$ and prove weak convergence of empirical processes $\sqrt{n}(F_n - F)$ in $L$. Eventually, we want to show that this convergence of the empirical processes implies convergence of the empirical quantile processes $\sqrt{n}(Q_n - Q)$ in $\mathbb{R}$. Since convergence in $L$ is a stronger requirement than convergence in $L_\infty$, we cannot rely on classical results to show convergence in our norm; now we develop the conditions for our convergence. The next theorem gives the complete characterization of weak convergence in $L_\mu$.

**Theorem 1.1** (Characterization of weak convergence in $L_\mu$). The sequence of processes $X_n : \Omega \to \mathbb{R}$ converges weakly in $L_\mu$ if and only if all of the following three conditions are met. (We denote $X_n(\omega)(t)$ by $X_n(t)$.)

(i) Every finite marginal $(X_n(t_1), \ldots, X_n(t_k))$ converges weakly in $\mathbb{R}^k$ for every $k$.

(ii) There exists a semimetric $\rho_1$ on $(a, b)$ such that $(a, b)$ is totally bounded in $\rho_1$ and $X_n$ is asymptotically uniformly $\rho_1$-equicontinuous in probability, that is, for
every $\varepsilon, \eta > 0$ there exists $\delta > 0$ such that

$$\limsup_{n \to \infty} P \left( \sup_{\rho_1(s,t) < \delta} |X_n(s) - X_n(t)| > \varepsilon \right) < \eta.$$

(iii) There exists a semimetric $\rho_2$ on $(a,b)$ such that $(a,b)$ is totally bounded in $\rho_2$ and $X_n$ is asymptotically $(\rho_2, \mu)$-equiintegrable in probability, that is, for every $\varepsilon, \eta > 0$ there exists $\delta > 0$ such that

$$\limsup_{n \to \infty} P \left( \sup_{t \in \mathbb{R}} \int_{0 < \rho_2(s,t) < \delta} |\tilde{X}_n(s)| d\mu(s) > \varepsilon \right) < \eta.$$ 

Remark. The classical empirical process literature shows that weak convergence in $L_\infty$ is connected to the Arzelá-Ascoli theorem [Van der Vaart and Wellner, 1996, Chapter 1.5]. This is to say that for the sequence of stochastic processes to converge weakly uniformly, the elements of the sequence must be equally uniformly continuous. The upshot of the above theorem is that, in order for weak convergence in $L_1$ to take place additionally, the elements of the sequence must be “equally integrable” as well. This insight is reminiscent of the Dunford-Pettis theorem in functional analysis.

Despite its technical complexity, the conditions of the theorem are not necessarily difficult to check. It is known that the empirical process $\sqrt{n}(F_n - F)$ satisfies conditions (i) and (ii) [Van der Vaart and Wellner 1996, Examples 2.1.3 and 2.5.4]. The following proposition shows that if $F$ has slightly more than variance, then it also satisfies condition (iii).

Proposition 1.2 (Convergence of empirical processes). Let $F$ be a probability distribution function on $\mathbb{R}$ with a $(2 + \varepsilon)$th moment for some $\varepsilon > 0$. Then the empirical process $\sqrt{n}(F_n - F)$ converges weakly in $L_1$ to a Gaussian process with mean zero and covariance function $\text{Cov}(x, y) = F(x \wedge y) - F(x)F(y)$.

7The classical central limit theorems only require finite variance. This marginal gap between the classical central limit theorems and the $L_1$ functional central limit theorem is mentioned in Barrio et al. (1999). This is the “cost of generality” we pay in this chapter. In some cases, however, it is possible to show that the second moment is sufficient, e.g., as in Shorack and Wellner (1986, Chapter 19) and Mason and Shorack (1992).
1.3.2 Step 2: Convergence of Quantile Processes

Now we proceed on to weak convergence of the empirical quantile process \( \sqrt{n}(Q_n - Q) \) in \( \mathcal{B} \). This is established by showing that the inverse map \( F \mapsto Q = F^{-1} \) is Hadamard differentiable as a map from \( \mathcal{L} \) to \( \mathcal{B} \). Weak convergence of the empirical quantile process then follows by the functional delta method.

The next theorem establishes Hadamard differentiability of the inverse map.

**Theorem 1.3 (Differentiability of the inverse map).** Let \( F \in \mathcal{L}_\phi \) be a distribution function that has at most finitely many jumps and is otherwise continuously differentiable with a strictly positive density. Then the inverse map \( \phi : \mathcal{L}_\phi \to \mathcal{B}, \phi(F) = Q \), is Hadamard differentiable at \( F \) tangentially to the set \( \mathcal{L}_0 \) of all continuous functions in \( \mathcal{L} \). The derivative map, \( \phi'_F : \mathcal{L}_0 \to \mathcal{B}, \) is given by

\[
\phi'_F(z)(u) = -z(Q(u))Q'(u), \quad u \in (0, 1).
\]

Importantly, the derivative formula, we find, is the same as the one known in the literature for the uniform norm (Van der Vaart and Wellner 1996, Section 3.9.4.2). Note that, although they are both about the “same” operator \( \phi : F \mapsto Q \), the derivative formula need not be the same as we have changed the norm. The delta method states that the distribution of a function of a statistic is characterized by the derivative and the distribution of the statistic. Then, that the derivative formula stays unchanged reveals a relieving fact that we do not need to worry about the unboundedness of the quantile functions when it comes to integrating them; we may continue using the same old formula.

We summarize the main conclusion of this section.

**Proposition 1.4 (Convergence of quantile processes).** Let \( m : \mathbb{R} \to \mathbb{R} \) be a continuously differentiable function. For a distribution function \( F \) on \( \mathbb{R} \) that has at most finitely many jumps and is otherwise continuously differentiable with strictly positive density such that \( m(X) \) has a \((2 + \varepsilon)\)th moment for \( X \sim F \) and some \( \varepsilon > 0 \), the process \( \sqrt{n}(m(Q_n) - m(Q)) \) converges weakly in \( \mathcal{B} \) to a Gaussian process with mean
zero and covariance function \( \text{Cov}(s,t) = m'(Q(s))Q'(s)m'(Q(t))Q'(t)(s \land t - st) \).

In addition to inversion, the proposition allows for transformation \( m \). While we leave the formal treatment to the Appendices, we provide an intuitive discussion of this generalization in the remainder of this section.

Assume for simplicity that \( m \) is increasing. Observe that by integration by parts,

\[
\int m(x)dF = -\int \tilde{F} dm(x).
\]

This indicates that existence of the expectation of the random variable \( m(X) \) is equivalent to the distribution function \( F \) of \( X \) belonging to the space \( \mathbb{L}_m \) of functions that are integrable with respect to \( m \). Meanwhile, by the change of variables \( u = F \circ m^{-1}(x) \),

\[
\int m(x)dF = \int x dF \circ m^{-1} = \int (F \circ m^{-1})^{-1} du = \int m(Q)du.
\]

Combine the results as follows. If \( X \) is such that \( m(X) \) has a finite first moment, then \( F \) belongs to \( \mathbb{L}_m \). This is equivalent to saying that \( F \circ m^{-1} \) belongs to \( \mathbb{L} \). Now we invoke Theorem 1.3 to find that its inverse \( (F \circ m^{-1})^{-1} \) is in \( \mathbb{B} \). Since \( (F \circ m^{-1})^{-1} = m(Q) \), it follows that \( m(Q) \) is in \( \mathbb{B} \). Finally, if, in addition, \( m(X) \) has (slightly more than) a variance, then its “empirical distribution function” \( \mathbb{F}_n \circ m^{-1} \) converges weakly in \( \mathbb{L} \) by Proposition 1.2, and hence the result follows by the delta method just established.

### 1.3.3 Step 3: Convergence of \( L \)-Statistics

The last step is to show that the \( L \)-statistics of the form \( \int m_j(Q_{n,j})d\mathbb{K}_n, j = 1, \ldots, d \), jointly converge weakly to a normal vector. Again, this is achieved by proving that the \( L \)-statistics, when seen as a map, are Hadamard differentiable. But for this, we need to take care of the randomness that arises from the measure \( \mathbb{K}_n \).

By the informal exercise in (1.3), the appropriate notion of convergence for \( \mathbb{K}_n \) is expected to involve integrability. It turns out that the norm developed in Section 1.3.1
does the right job. Here we recall the definition with specialization to the unit interval.

**Definition.** For a quantile function $Q : (0, 1) \to \mathbb{R}$, denote by $L_Q$ the Banach space of functions $\kappa : (0, 1) \to \mathbb{R}$ with the norm

$$
\|\kappa\|_{L_Q} := \left( \sup_{u \in (0, 1)} |\kappa(u)| \right) \lor \left( \int_0^1 |\tilde{\kappa}(u)| dQ(u) \right),
$$

where $\tilde{\kappa}(u) := \kappa(u) - \kappa(0) \mathbb{1}_{\{u \leq 1/2\}} - \kappa(1) \mathbb{1}_{\{u > 1/2\}}$. Define by $L_{Q,M}$ the subset of $L_Q$ of Lipschitz functions whose Lipschitz constants are uniformly bounded by $M$.

Now we are ready to show Hadamard differentiability of $L$-statistics. Fortunately, the derivative formula in the next theorem confirms our intuition in equation (1.3).

**Theorem 1.5 (Differentiability of $L$-statistics).** For each $M$, the maps $\lambda : \mathcal{B} \times L_{Q,M} \to \mathbb{R}$ and $\tilde{\lambda} : \mathcal{B} \times L_{Q,M} \to L_\infty(0, 1)^2$,

$$
\lambda(Q, K) = \int_0^1 Q(u) dK(u) \quad \text{and} \quad \tilde{\lambda}(Q, K)(s, t) = \int_s^t Q(u) dK(u),
$$

are Hadamard differentiable at $(Q, K) \in \mathcal{B}_0 \times L_{Q,M}$ uniformly over $L_{Q,M}$ tangentially to the set $\mathcal{B} \times L_{Q,0}$ where $L_{Q,0}$ is the subset of $L_Q$ of continuous functions $\kappa$ such that $Q(u)\kappa(u) \to 0$ as $u \to \{0, 1\}$. The derivative is given by

$$
\lambda_{Q,K}'(z, \kappa) = \int_0^1 Q(u) d\kappa(u) + \int_0^1 z(u) dK(u),
$$

$$
\tilde{\lambda}_{Q,K}'(z, \kappa)(s, t) = \int_s^t Q(u) d\kappa(u) + \int_s^t z(u) dK(u),
$$

where $\int Q d\kappa$ is defined via integration by parts if $\kappa$ is of unbounded variation.

Thus, for Hadamard differentiability of $L$-statistics, we require that the random “distribution function” $K_n$ be uniformly Lipschitz, that is, when seen as a measure, $K_n$ has a uniformly bounded density with respect to the Lebesgue measure.

Do selection measures such as outlier removal or winsorization satisfy this condition? If so, how can we verify it? Are there more primitive conditions that are easy to check? To answer these questions, consider the randomly weighted sum

$$
\frac{1}{n} \sum_{i=1}^n X_i w_i,
$$

36
or equivalently, $\frac{1}{n} \sum_{i=1}^{n} X(i)w(i)$ where $w(i)$ is sorted according to the order of $X_i$.

Using the empirical quantile function $Q_n$ of $X_i$, we write this sum as an integral:

$$\int_{0}^{1} Q_n dK_{n,0} \quad \text{where} \quad K_{n,0}(u) := \frac{1}{n} \sum_{i=1}^{n} w(i) \times \begin{cases} 0 & u < \frac{i}{n}, \\ \frac{i}{n} & \frac{i}{n} \leq u. \end{cases}$$

This function $K_{n,0}$ is simple enough but the results developed in this chapter require that this function be Lipschitz. We accomplish this by linearly interpolating $K_{n,0}$, as $Q_n$ is piecewise constant on $1/n$ intervals. In particular, we can replace the integral by

$$\int_{0}^{1} Q_n dK_n \quad \text{where} \quad K_n(u) := \frac{1}{n} \sum_{i=1}^{n} w(i) \times \begin{cases} 0 & u < \frac{i-1}{n}, \\ \frac{i-1}{n} & \frac{i-1}{n} \leq u < \frac{i}{n}, \\ \frac{i}{n} & \frac{i}{n} \leq u. \end{cases}$$

Since $F_n(X(i)) = i/n$, we can write $K_n(u)$ as

$$\frac{1}{n} \sum_{i=1}^{n} w_i \mathbb{1} \{ 0 \vee (nu - nF_n(X_i) + 1) \wedge 1 \}.$$

Therefore, as long as $w_i$ is bounded by some constant, this $K_n$ is Lipschitz almost surely.

**Proposition 1.6 (Convergence of selection measures).** Let $U_1, \ldots, U_n$ be independent uniformly distributed random variables on $(0,1)$ and $w_{1,n}, \ldots, w_{n,n}$ random variables bounded by some constant $M$ whose distribution can depend on $U_1, \ldots, U_n$ and $n$. Define

$$F_n(u) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \{ U_i \leq u \}, \quad G_n(u) := \frac{1}{n} \sum_{i=1}^{n} w_{i,n} \mathbb{1} \{ U_i \leq u \}.$$

Let $I(u) := u$ and assume that $K(u) := \lim_{n \to \infty} \mathbb{E}[G_n(u)]$ exists and is Lipschitz and differentiable. If $\sqrt{n}(G_n - K)$ weakly converges jointly with $\sqrt{n}(F_n - I)$ in $L_\infty$, then
for the “selection” measure

\[ \mathbb{K}_n(u) := \frac{1}{n} \sum_{i=1}^{n} w_i \mathbb{1} \{ 0 \lor (nu - nF_n(U_i) + 1) \land 1 \}, \]

we have \( \sqrt{n}(\mathbb{K}_n - K) \) converge weakly in \( \mathbb{L}_Q \) for every quantile function \( Q \) whose distribution has a \((2 + c)\)th moment for some \( c > 0 \).

This means that most “well-behaved” sample selection measures converge in \( \mathbb{L}_Q \); roughly speaking, if the empirical distribution of the selected sample \( X_{1,n}, \ldots, X_{m,n} \) converges in the traditional uniform sense together with that of the entire sample, then the selection measure \( \mathbb{K}_n \) as defined in Section 1.2 converges in \( \mathbb{L}_Q \). This can be verified as follows.

**Example** (Outlier robustness analysis). Let \( F \) be the true distribution of \( x_iy_i \). In this example, \( \mathbb{F}_n \) and \( \mathbb{G}_n \) in Proposition 1.6 are

\[ \mathbb{F}_n(u) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \{ F(x_iy_i) \leq u \}, \quad \mathbb{G}_n(u) := \frac{1}{n} \sum_{i=1}^{n} w_{i,n} \mathbb{1} \{ F(x_iy_i) \leq u \}. \]

Since both \( y(\lfloor \tau n \rfloor + 1) \) and \( y(n - \lfloor \tau n \rfloor) \) converge almost surely to \( Q_y(\tau) \) and \( Q_y(1 - \tau) \), we have that for outlier removal at \( \tau \)- and \( (1 - \tau) \)-quantiles \([1.1]\), \( w_{i,n} \) converges almost surely to \( \mathbb{1} \{ Q_y(\tau) \leq y \leq Q_y(1 - \tau) \} \) and for winsorization at \( \tau \)- and \( (1 - \tau) \)-quantiles \([1.2]\), to

\[ w_i = \begin{cases} Q_y(\tau)/y_i & y_i < Q_y(\tau) < 0, \\ Q_y(1 - \tau)/y_i & y_i > Q_y(1 - \tau) > 0, \\ 1 & \text{otherwise}. \end{cases} \]

Then, \( \mathbb{F}_n \) and \( \mathbb{G}_n \) jointly converge uniformly, respectively to an identity function and \( \mathbb{E}[w_i \mid F(x_iy_i) = u] \). \( \square \)

Now we are ready to state the main result of this chapter: the joint asymptotic distribution of general \( L \)-statistics.

**Proposition 1.7** (Convergence of \( L \)-statistics). Let \( m_1, m_2 : \mathbb{R} \to \mathbb{R} \) be continu-
such that the empirical distributions of $X_L$ normalize uniformly jointly to continuously differentiable distribution functions. Then, the constant $\omega$ let independent and identically distributed random variables $m$ continuously differentiable with strictly positive marginal densities such that marginal distributions $F$ uniformly differentiable functions and $\mathbb{E}$ converge weakly in $\mathbb{R}$. Together, we see that $\mathbb{E}$ and $\mathbb{E}_n$ for $n = 1$ can depend on $n$ and all of $X_1, \ldots, X_n$ and $X_1, \ldots, X_n$ such that the empirical distributions of $X_{i,1}, X_{i,2}, w_{i,n,1}X_{i,1}$, and $w_{i,n,2}X_{i,2}$ converge uniformly jointly to continuously differentiable distribution functions. Then, the normalized $L$-statistics

$$\sqrt{n} \left( \frac{\mathbb{E}[m_1(X_{i,1})w_{i,n,1}] - \mathbb{E}[m_1(X_{i,1})w_{i,n,1}]}{\mathbb{E}[m_2(X_{i,2})w_{i,n,2}] - \mathbb{E}[m_2(X_{i,2})w_{i,n,2}]} \right) = \sqrt{n} \left( \int_0^1 m_1(Q_{n,1})dK_{n,1} - \int_0^1 m_1(Q_1)dK_1 \right)$$

where

$$K_{n,j}(u) := \frac{1}{n} \sum_{i=1}^n w_{i,n,j} \mathbb{I} \left\{ 0 \lor \left( u - \mathbb{E}_{n,j}(X_j) + \frac{1}{n} \right) \land \frac{1}{n} \right\},$$

$$K_j(u) := \lim_{n \to \infty} \mathbb{E}[w_{i,n,j} | F_j(X_{i,j}) \leq u],$$

converge weakly in $\mathbb{R}^2$ to a normal vector $(\xi_1, \xi_2)$ with mean zero and (co)variance

$$\text{Cov}(\xi_j, \xi_k) = \int_0^1 \int_0^1 m_j(Q_j(s))Q_j'(s)m_k(Q_k(t))Q_k'(t) \times \left( [F_{jk}(s, t) - st] + [K_{jk}(s, t)F_{jk}(s, t) - stK_j(s)K_k(t)] \right. \left. - K_j(s)[F_{jk}(s, t) - st] - K_k(t)[F_{jk}(s, t) - st] \right) dsdt,$$

where $K_{jk}(s, t) := \lim_{n \to \infty} \mathbb{E}[w_{i,n,j}w_{i,n,k} | X_{i,j} \leq Q_j(s), X_{i,k} \leq Q_k(t)]$ and $F_{jk}(s, t) := \text{Pr}(X_{i,j} \leq Q_j(s), X_{i,k} \leq Q_k(t))$. In applications, one can compute the distribution either analytically, by paramet-
ric bootstrap, or by nonparametric bootstrap. Nonparametric bootstrap does not require distributional assumptions and can be quite convenient when one iteration of the estimation does not consume much time. The following is a procedure for the nonparametric bootstrap, stated for completeness.

**Proposition 1.8 (Validity of nonparametric bootstrap).** In the assumptions stated in Proposition 1.7 assume further that \( w_{i,n,j} \) represents sample selection based on a fixed number of empirical quantiles\(^8\) Then, the joint distribution of \((\hat{\beta}_1, \ldots, \hat{\beta}_d)\) can be computed by nonparametric bootstrap. The algorithm is as follows. Here, \( X_i \) denotes a vector \((X_{i,1}, \ldots, X_{i,d})\).

i. Bootstrap \( n \) (or fewer) random observations from \( X_1, \ldots, X_n \) with replacement.

ii. Compute the statistics \((\hat{\beta}_1^*, \ldots, \hat{\beta}_d^*)\) for the bootstrapped sample.

iii. Repeat the above steps \( S \) times.

iv. Use the empirical distribution of \((\hat{\beta}_1^*, \ldots, \hat{\beta}_d^*)\) as the approximation to the theoretical asymptotic distribution of \((\hat{\beta}_1, \ldots, \hat{\beta}_d)\).

We have hitherto assumed that \( X_i \) of interest is univariate. Multivariate cases, as in regressions, can be accommodated as follows.

**Example** (Multivariate regression). Let \( x_i = (1, x_{i,1}, x_{i,2})' \) and \( \beta = (\beta_0, \beta_1, \beta_2)' \), and consider \( y_i = x_i'\beta + \varepsilon_i \). The OLS estimator for \( \beta \) is

\[
\hat{\beta} = \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i y_i = \beta + \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i \varepsilon_i.
\]

Therefore,

\[
\sqrt{n}(\hat{\beta}_1 - \beta_1) = c_1 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i + c_2 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{i,1} \varepsilon_i + c_3 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_{i,2} \varepsilon_i + o_P(1)
\]

\(^8\)The assumption on convergence must be extended (from bivariate) to joint over all processes involved.
for some constants \( c_1, c_2, \) and \( c_3 \). Thus, one can reduce the weak convergence of the vector \( \sqrt{n}(\hat{\beta} - \beta) \) to the joint convergence of univariate empirical quantiles of \( \varepsilon_i, x_{i,1}\varepsilon_i, \) and \( x_{i,2}\varepsilon_i \).

We note some possibilities for generalizing our results to other cases that are not considered in this chapter. For instance, we do not explicitly consider cases where data are dependent (Dehling et al., 2002, 2014), where smoothed or estimated cdfs are substituted for empirical cdfs (Hall et al., 1999; Berg and Politis, 2009), or where a non-conventional convergence device such as extreme value theory is used for the stochastic processes (Einmahl, 1992; Rootzén, 2009; Drees and Rootzén, 2010, 2016). The part that requires additional work is the proof that \( F_n \) and \( K_n \) in each case converge weakly in our space \( L_\mu \). Fortunately, since we have completely characterized weak convergence in \( L_\mu \) in Theorem 1.1 without relying particularly on the central limit theorem structure (see Section 1.6.1 and Appendix 1.D), half of such work is already taken care of. Once convergence of \( F_n \) and \( K_n \) is established, weak convergence of their transformations follows immediately by the Hadamard differentiability of the maps proved herein.

This concludes the overview of the key ideas and main results of the chapter. Interested readers may consult the Appendices for general statements and proofs.

1.4 Application to Outlier Robustness Analysis

1.4.1 Test of Robustness to Outliers

We apply the results developed in Section 1.3 to the problem described in Section 1.2.2 and construct a statistical test of outlier robustness analysis. We briefly recall our setup from Section 1.2.2. Let \( \beta_1 \) be the parameter of interest and \( \hat{\beta}_1 \) its estimator. Denote by \( \hat{\beta}_2 \) the estimator that is computed with outlier-adjusted sample, i.e., the sample that excludes or winsorizes outliers. Since outlier removal or winsorization can change the true parameter in the population, we let \( \beta_2 \) be the true parameter.
from the outlier-adjusted population. The null hypothesis we want to test is given by

\[ H_0 : \|\beta_1 - \beta_2\| \leq h \]

for a fixed \( h \geq 0 \).

We assume that \( h \) is a scalar while \( \beta \) can be a vector, and we take the norm \( \| \cdot \| \) to be the Mahalanobis distance between \( \beta_1 \) and \( \beta_2 \), that is, \( \left((\hat{\beta}_1 - \hat{\beta}_2)'\Sigma^{-1}(\hat{\beta}_1 - \hat{\beta}_2)\right)^{1/2} \) where \( \Sigma \) is either an identity, the covariance matrix of \( \hat{\beta}_1 - \hat{\beta}_2 \), or some other positive definite symmetric matrix. The natural test statistic to use is \( \|\hat{\beta}_1 - \hat{\beta}_2\| \).

Let \( \alpha \in (0, 1) \) be the size of the test. According to the main result, the variance \( \Sigma \) of the difference \( \hat{\beta}_1 - \hat{\beta}_2 \) can be estimated either by the analytic formula or by the bootstrap. Note that if \( h > 0 \), the null hypothesis is composite; hence the definition of critical values includes taking supremum over the set of point null hypotheses. In particular, the critical value \( c_\alpha \) in a general case must satisfy

\[ \sup_{\|v\| \leq 1} \Pr(\|hv + \xi\|^2 > c_\alpha) \leq \alpha, \]

where \( \xi \sim N(0, \Sigma) \). If \( \beta \) is a scalar, it reduces to finding \( c_\alpha \) such that

\[ \Pr((h + \xi)^2 > c_\alpha) = \alpha \]

for \( \xi \sim N(0, \sigma^2) \) where \( \sigma^2 \) is the variance of \( \hat{\beta}_1 - \hat{\beta}_2 \).

### 1.4.2 Empirical Application to the Effect of Democracy on Growth

Now we apply this test to reinvestigate the outlier robustness analysis in [Acemoglu et al. (2017)](#). The aim of their paper is to answer the long-standing question of whether democracy affects economic growth in a negative or positive way. To address difficulties arising from the effect of DGP dynamics and endogenous selection into democracy, [Acemoglu et al. (2017)](#) conduct three analyses that guard against
different possibilities and find very similar results: after 25 years from permanent democratization, GDP per capita is about 20% higher than it would be otherwise. The three analyses in Acemoglu et al. (2017) consist of fixed effects regression on a dynamic panel that models GDP dynamics, treatment effects analysis that does not impose parametric assumptions on the GDP process, and IV fixed effects regression on the same dynamic panel instrumenting a wave of democratization. Acemoglu et al. (2017) then check robustness of their results to outliers for the two panel regressions. In this section, we estimate the joint distribution of the baseline and outlier-removed estimates in Acemoglu et al. (2017) and conduct a test of outlier robustness as developed above.

The first regression equation is given by:

$$\log GDP_{i,t} = \beta_0 Democra cy_{i,t} + \sum_{s=1}^{4} \beta_s \log GDP_{i,t-s} + \alpha_i + \delta_t + \varepsilon_{i,t},$$

where \( i \) represents a country, \( t \) a year, and \( Democra cy_{i,t} \) the indicator of democracy at country \( i \) in year \( t \). Here, Acemoglu et al. (2017) assume sequential exogeneity, which means the error term is mean independent with all contemporary and past variables, namely democracy, the GDP, and fixed effects:

$$E[\varepsilon_{i,t} | \log GDP_{i,t}, Democra cy_{i,t-u}, \alpha_i, \delta_t : s = 1, \ldots, t, u = 0, \ldots, t] = 0$$

for every \( i \) and \( t \). The data consist of 6,336 observations. The original paper examines two more specifications, but we omit them as the results of reexamination are similar.

In the third analysis, Acemoglu et al. (2017) use the regional wave of democratization as an instrument. The first-stage equation is now

$$Democra cy_{i,t} = \sum_{s=1}^{4} \pi_s WaveOfDemocra cy_{i,t-s} + \sum_{s=1}^{4} \phi_s \log GDP_{i,t-s} + \theta_i + \eta_t + \nu_{i,t},$$

where \( WaveOfDemocra cy_{i,t} \) is the instrument that is constructed by indicators of democracy of nearby countries that share similar political history as country \( i \). The
assumption needed for this IV model is the exclusion restriction:

$$\mathbb{E}[\varepsilon_{i,t} \mid \log GDP_{i,t-s}, \text{WaveOfDemocracy}_{i,t-s}, \alpha_i, \delta_t : s = 1, \ldots, t] = 0$$

for every $i$ and $t$. Since the panel data is unbalanced, each country has a varied number of observations. Let $t_i$ be the year of a country $i$’s first appearance in the sample and $T_i$ be the number of observations country $i$ has. Then, $i$’s array of time observations consists of $(i, t_i), (i, t_i + 1), \ldots, (i, t_i + T_i - 1)$.

Aside from the regression coefficients, Acemoglu et al. (2017) report three more parameters. The first is the long-run effect of democracy defined as $\beta_5 := \beta_0 / (1 - \beta_1 - \beta_2 - \beta_3 - \beta_4)$, which represents the impact on $\log GDP_{i,\infty}$ of the transition from non-democracy $D_{i,t-1} = 0$ to permanent democracy $D_{i,t+s} = 1$ for every $s \geq 0$. The second parameter is the effect of transition to democracy after 25 years given by $\beta_6 := e_{25}$, where $e_j = \beta_0 + \beta_1 e_{j-1} + \beta_2 e_{j-2} + \beta_3 e_{j-3} + \beta_4 e_{j-4}$ and $e_0 = e_{-1} = e_{-2} = e_{-3} = 0$, which represents the impact on $\log GDP_{i,25}$ of the same transition. The third parameter is persistence of the GDP process defined to be $\beta_7 := \beta_1 + \beta_2 + \beta_3 + \beta_4$, which represents how persistently a unit change in $\log GDP$ would remain.

To check robustness of their results to outliers, Acemoglu et al. (2017) carry out the same regression but exclude some observations that have large residuals. For notational convenience, let

$$x_{i,t} := \begin{bmatrix}
\text{Democracy}_{i,t} \\
\log GDP_{i,t-1} \\
\vdots \\
\log GDP_{i,t-4} \\
1_{t=1} \\
\vdots \\
1_{t=N} \\
1_{t=0} \\
\vdots \\
1_{t=T}
\end{bmatrix}, \quad \beta := \begin{bmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_4 \\
\alpha_1 \\
\vdots \\
\alpha_N \\
\delta_1 \\
\vdots \\
\delta_T
\end{bmatrix}, \quad z_{i,t} := \begin{bmatrix}
\text{WaveOfDemocracy}_{i,t-1} \\
\vdots \\
\text{WaveOfDemocracy}_{i,t-4} \\
\log GDP_{i,t-1} \\
1_{t=1} \\
\vdots \\
1_{t=N} \\
1_{t=0} \\
\vdots \\
1_{t=T}
\end{bmatrix}, \quad \pi := \begin{bmatrix}
\pi_1 \\
\vdots \\
\pi_4 \\
\phi_1 \\
\vdots \\
\phi_4 \\
\theta_1 \\
\vdots \\
\theta_N \\
\eta_1 \\
\vdots \\
\eta_T
\end{bmatrix}.$$

Outliers are defined in their paper by $|\tilde{\epsilon}_{i,t}| \geq 1.96 \hat{\sigma}_\varepsilon$, where $\hat{\sigma}_\varepsilon$ is the estimate of the
homoskedastic standard error of \( \varepsilon \)

\[
\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=t_i}^{T_i-1} (y_{i,t} - x'_{i,t} \hat{\beta})^2;
\]

and, for the IV model, also by \(|\hat{\varepsilon}_{i,t}| \geq 1.96 \hat{\sigma}_v\), where

\[
\hat{\sigma}^2_v = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=t_i}^{T_i-1} (x_{i1,t} - z'_{i,t} \hat{\pi})^2.
\]

This means that they are concerned whether tail observations in the GDP might have disproportionate effects on the estimates. Defining outliers based on \( \hat{\varepsilon} \) but not on \( y \), even if they are interested in the effects of outliers of the GDP, is a reasonable choice since, under some assumptions, sample selection based on \( \hat{\varepsilon} \) does not affect the true parameters while selection based on the dependent variable \( \log GDP \) would almost certainly bias the true parameters.

Let \( F_{n,xy} \) be the vector of marginal empirical distribution functions of \( \frac{1}{T_i} \sum_t x_{i,t,y_{i,t}} \) and \( Q_{n,xy} \) the vector of marginal empirical quantile functions of \( \frac{1}{T_i} \sum_t x_{i,t,y_{i,t}} \). Note that, with \( w_{i,t,n} = 1 \{ |\hat{\varepsilon}_{i,t}| \geq 1.96 \hat{\sigma}_\varepsilon \} \), the full-sample and outlier-removed OLS estimators are written as

\[
\hat{\beta}_{1,OLS} = \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=t_i}^{T_i-1} x_{i,t} x'_{i,t} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=t_i}^{T_i-1} x_{i,t} y_{i,t}
\]

\[
= \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=t_i}^{T_i-1} x_{i,t} x'_{i,t} \right)^{-1} \int_0^1 Q_{n,xy}(u) du,
\]

\[
\hat{\beta}_{2,OLS} = \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=t_i}^{T_i-1} x_{i,t} x'_{i,t} w_{i,t,n} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=t_i}^{T_i-1} x_{i,t} y_{i,t} w_{i,t,n}
\]

\[
= \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T_i} \sum_{t=t_i}^{T_i-1} x_{i,t} x'_{i,t} w_{i,t,n} \right)^{-1} \int_0^1 Q_{n,xy}(u) dK_{n,xy}(u),
\]

where \( K_{n,xy} \) is the vector of random selection measures whose \( j \)th element assigns

\footnote{The purpose of computing the homoskedastic standard error \( \hat{\sigma} \) is normalization. Acemoglu et al. (2017) do allow for heteroskedasticity and use heteroskedasticity-robust estimators for inference.}
density $\sum_t x_{i,t,j}y_{i,t}w_{i,t,n}/\sum_t x_{i,t,j}y_{i,t}$ to
\[
u \in \left( \mathbb{F}_{n,xy,j} \left( \frac{1}{T_t} \sum_t x_{i,t,j}y_{i,t} \right) - 1/n, \mathbb{F}_{n,xy,j} \left( \frac{1}{T_t} \sum_t x_{i,t,j}y_{i,t} \right) \right].
\]
Assume that the cdfs of $\frac{1}{T_t} \sum_t x_{i,t,j}y_{i,t}$ are smooth with $(2+c)$th moments for some $c > 0$ and $\hat{\sigma}_v$ has a well-defined limit. Then, our results indicate that the joint distribution of two vectors
\[
\int_0^1 Q_{n,zy}(u)du \quad \text{and} \quad \int_0^1 Q_{n,zy}(u)dK_{n,zy}(u)
\]
converges and can be estimated by nonparametric bootstrap. Since $\hat{\beta}_{\text{OLS}}^1$ and $\hat{\beta}_{\text{OLS}}^2$ converge to fixed combinations of elements of these vectors, their joint distribution can also be estimated by nonparametric bootstrap, as we will do.

Similarly, let $\mathbb{F}_{n,zy}$ and $Q_{n,zy}$ be the vectors of marginal empirical distribution functions and marginal empirical quantile functions of $\frac{1}{T_t} \sum_t z_{i,t}y_{i,t}$. The full-sample and outlier-removed IV estimators are written as
\[
\hat{\beta}_{\text{IV}}^1 = (xz'zz^{-1}z^x')^{-1}xz'(zz')^{-1} \int_0^1 Q_{n,zy}(u)du,
\]
\[
\hat{\beta}_{\text{IV}}^2 = (xz'wzz'w^{-1}z^x')^{-1}xz'w(zz'w)^{-1} \int_0^1 Q_{n,zy}(u)dK_{n,zy}(u),
\]
where $\bar{w}_{i,n} = 1\{|\hat{\epsilon}| \geq 1.96\hat{\sigma}_v$ and $|\hat{\nu}| \geq 1.96\hat{\sigma}_v\},$
\[
(zx')' = \bar{xz'} = \frac{1}{n} \sum_{i=1}^n \frac{1}{T_t} \sum_{t=t_i}^{t_i+T_t-1} x_{i,t}z'_{i,t}, \quad \overline{zz'} = \frac{1}{n} \sum_{i=1}^n \frac{1}{T_t} \sum_{t=t_i}^{t_i+T_t-1} z_{i,t}z'_{i,t},
\]
\[
(zx'w)' = \bar{xz'w} = \frac{1}{n} \sum_{i=1}^n \frac{1}{T_t} \sum_{t=t_i}^{t_i+T_t-1} x_{i,t}z'_{i,t} \bar{w}_{i,t,n}, \quad \overline{zz'w} = \frac{1}{n} \sum_{i=1}^n \frac{1}{T_t} \sum_{t=t_i}^{t_i+T_t-1} z_{i,t}z'_{i,t} \bar{w}_{i,t,n},
\]
and $K_{n,zy}$ is the vector of random selection measures whose $j$th element assigns density $\sum_t z_{i,t,j}y_{i,t}w_{i,t,n}/\sum_t z_{i,t,j}y_{i,t}$ to $u \in (\mathbb{F}_{n,zy,j} \left( \frac{1}{T_t} \sum_t z_{i,t,j}y_{i,t} \right) - 1/n, \mathbb{F}_{n,zy,j} \left( \frac{1}{T_t} \sum_t z_{i,t,j}y_{i,t} \right)]$. Again, if $\frac{1}{T_t} \sum_t z_{i,t,j}y_{i,t}$ has smooth cdfs with $(2+c)$th moments and $\hat{\sigma}_v$ has a well-defined limit, our results imply that the joint distribution of $\hat{\beta}_{\text{IV}}^1$ and $\hat{\beta}_{\text{IV}}^2$ can be derived by
nonparametric bootstrap.

In a simple case where $\varepsilon$ (and $v$) is independent of the covariates, outlier removal according to this criterion will not change the true values of the coefficients at least asymptotically; therefore, it is sensible to set the allowed bias $h$ to the most conservative choice, zero. Letting $\beta^1_j$ and $\beta^2_j$ be the full-sample and outlier-removed true coefficients respectively, we are testing the null hypothesis in which $\beta^1_j$ is postulated to be identical to $\beta^2_j$, i.e., $H_0: \beta^1_j = \beta^2_j$.

The quantile functions and sample selection functions that appear in the $L$-statistics formula are visualized in Figures 1.A.1a and 1.A.1b. In particular, Figures 1.A.1a and 1.A.1b show selected elements of empirical quantile functions $Q_{n,xy}$ and sample selection functions $K_{n,xy}$ used in OLS estimators. The blue line in Figure 1.A.1a is the empirical quantile function of the time average of $Democracy_{i,t} \times \log GDP_{i,t}$, which is the first element of $Q_{n,xy}$; the solid orange line is the sample selection function for outlier removals, which is the first element of $K_{n,xy}$; the dashed orange line is the identity function (the sample selection function for the baseline estimator). Similarly, Figure 1.A.1b shows the empirical quantile and sample selection functions for the time average of $\log GDP_{i,t-1} \times \log GDP_{i,t}$. Figures 1.A.1c and 1.A.1d depict selected elements of empirical quantile functions $Q_{n,zy}$ and sample selection functions $K_{n,zy}$ used in IV estimators. Now, the first element of $Q_{n,zy}$ is the empirical quantile function of the time average of $WaveOfDemocracy_{i,t} \times \log GDP_{i,t}$, which we represent with the blue line in Figure 1.A.1c. The sample selection function for this time average is the solid orange line, which is less steep than that in Figure 1.A.1a this is due to the additional removal of observations for large first-stage errors. Figure 1.A.1b shows the empirical quantile and sample selection functions for the time average of products of log GDP and its lag for IV estimators.

Outlier selection criteria are visualized in Figures 1.A.2a to 1.A.2d; they indicate that there is no “crazy” observations that can drastically change the analysis but instead error distributions are as smoothly distributed as normal distributions. Figure 1.A.2a gives the histogram of estimated errors $\hat{\varepsilon}_{i,t}$ of the OLS regression. The dotted line indicates the threshold of outliers, $1.96\hat{\sigma}_{\varepsilon}$ and $-1.96\hat{\sigma}_{\varepsilon}$. The blue
observations are included in the outlier-adjusted sample and the red are excluded. Figure 1.A.2b gives the two-dimensional histogram of estimated errors \((\hat{v}_{i,t}, \hat{\varepsilon}_{i,t})\); the blue observations in the rectangle are included in the outlier-removed sample while the red outside the rectangle are not. Figure 1.A.2c and Figure 1.A.2d show the marginal distributions of \(\hat{v}_{i,t}\) and \(\hat{\varepsilon}_{i,t}\); some observations in the blue bars are excluded because of the other error falling outside the cutoff.

We carry out nonparametric bootstrap by randomly drawing countries \(i\). All fixed effects are replaced by their corresponding dummy variables. Each draw of country \(i\) adds a \(T_i\) number observations to the bootstrap sample; equivalently, we treat each sum over time, in particular, \(\frac{1}{T_i} \sum_t x_{i,t}y_{i,t}, \frac{1}{T_i} \sum_t y_{i,t}y_{i,t-s},\) and \(\frac{1}{T_i} \sum_t z_{i,t}y_{i,t}\), as an observation in the bootstrap in order to exploit the i.i.d. structure needed for our theory. Here, the bootstrap consists of 10,000 iterations. In each iteration for OLS regression, we draw 175 random countries with replacement; for IV regression, 174 random countries with replacement.

Our reexamination of Acemoglu et al. (2017) mostly reconfirms robustness to outliers of the results found in Acemoglu et al. (2017) with the most stringent choice of a hypothesis \((h=0)\). However, there is one coefficient, persistence of the GDP process, for which the hypothesis of outlier robustness is rejected. Table 1.A.1 lists the estimates and \(p\)-values for the hypotheses that outliers have no effect on the parameters. Column 1 shows the baseline OLS estimates of key parameters that use the full sample. The figures in Column 2 are the outlier-removed OLS estimates that remove observations with \(|\hat{\varepsilon}_{i,t}| \geq 1.96\hat{\sigma}_e\). Column 3 provides the baseline IV estimates, and column 4 the outlier-removed IV estimates, removing observations with \(|\hat{\varepsilon}_{i,t}| \geq 1.96\hat{\sigma}_e\) or \(|\hat{v}_{i,t}| \geq 1.96\hat{\sigma}_v\). Columns 5 to 8 illustrate the utility of our results in formal tests of outlier robustness analysis. Column 5 gives the \(p\)-values of the hypotheses that the two parameters estimated by columns 1 and 2 are identical, \(H_0: \beta_{1j}^1 = \beta_{1j}^2\), using the standard error of the difference of two estimators estimated by bootstrap. Column 6 gives the “\(p\)-values” of the same hypotheses, but uses the standard error of the marginal distribution of the baseline OLS estimates. These results can be considered as \(p\)-values of the “heuristic arguments” explained in the
introduction. Column 7 shows the p-values of the same hypotheses calculated with IV estimates, using the standard error of the difference. Column 8 lists the “p-values” using the marginal standard error of the baseline IV estimates. We see that the identity of persistence of the GDP process is rejected in formal tests while accepted in heuristic tests at the 5% level. We note that the magnitudes of persistence are very close in both regressions (0.96 and 0.97), so if we allow bias $h$ of, say, 0.01, the hypothesis will not be rejected. The point of this chapter is that, even when we end up accepting the robustness hypothesis, such results should be rooted in correct statistical reasoning.

Positive correlation of baseline and outlier-adjusted estimators can be visualized by our bootstrap results. Figures 1.A.3a and 1.A.3b illustrate the joint distributions of baseline and outlier-removed OLS estimators, $(\hat{\beta}_1^0, \hat{\beta}_2^0)$ and $(\hat{\beta}_1^7, \hat{\beta}_2^7)$. Figures 1.A.3c and 1.A.3d show the joint distributions of baseline and outlier-removed IV estimators, $(\hat{\beta}_1^0, \hat{\beta}_2^0)$ and $(\hat{\beta}_1^7, \hat{\beta}_2^7)$. For the contour plots, we use the kernel density estimators for ease of visualization (instead of scatter-plotting the bootstrap points). We see that the estimators are positively correlated, which is anticipated by the fact that they are based on similar sets of samples. The figures illustrate why we need the joint distributions to statistically test our null hypotheses. Graphically, the tests examine if each red star in the figures is close enough to the 45 degree line shown as the black dotted line.

To see whether there were countries that were consistently labeled as outliers, we present the histograms of numbers of removal in Figures 1.A.4a and 1.A.4b. If any observation with index $i$ is removed in an iteration, we increment the “number of removal” for country $i$. There is the largest spike at 6,000–6,200, with the second largest one at 0 in each figure. This means that more than half the countries experience about 6,000 removals throughout the bootstrap, while a little fewer than half do not undergo any removal. In other words, there is no small portion of countries that is consistently marked as an outlier, while there is a slight tendency to remove a certain subset of countries. We interpret this as follows: there are likely to be no “outliers” in the sample in the sense that they potentially come from a different data-generating
process, while there are observations that happen to be relatively more extreme than the rest, which is a natural consequence of random observations.

1.5 Applications to Other Econometric Problems

Our results on $L$-statistics are new and of independent interest. As $L$-statistics appear in many places in economics, the results can be applied to other problems aside from the outlier robustness analysis. We discuss two applications and briefly describe two more applications from Chapter 3 and Kaji (2017).

1.5.1 Multiple Testing with Dependence

Economists often contend with tens or hundreds of statistical inference problems in a single research project (Banerjee et al., 2015b; Casey et al., 2012; Anderson, 2008). As a consequence, economists devote increasing attention to the simultaneous inference problem. The simultaneous inference problem refers to the issue that statistical discoveries often arise purely by chance when many hypotheses are tested simultaneously and individually. For instance, if one tests a hundred hypotheses at size 5% each, then even when all of the null hypotheses are true, we expect that about five of them come out rejected (if the hypotheses were jointly independent).

If we value even a single statistical discovery out of a large number of hypotheses, then procedures that control the probability of obtaining even one false positive, the familywise error rate (FWER), turn out to be too conservative for practical use in many contexts. Therefore, statisticians have proposed alternative forms of error control (Lehmann and Romano 2005; Romano et al. 2010). Among them, the false discovery rate (FDR) is an increasingly popular concept (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001; Yekutieli 2007; Romano et al. 2008).

To illustrate the utility of our results in this setting, suppose we have many hypotheses to test, and some of the test statistics are based on different subgroups. Consider, for example, the effects of productivity shocks on rice yields among subgroups classified by quartiles of land ownership (Demont 2013); relationship between
the wage and crop yield instrumented by the indicator of rainfall being above and below certain percentiles (Jayachandran, 2006); or the effect of access to microcredit on business revenue with and without individuals who are above the 99th percentile in business revenue (Augsburg et al., 2015). Assuming that these statistics are asymptotically linear, we are interested in $d$ statistics $\hat{\beta}_1, \ldots, \hat{\beta}_d$ of the form

$$\hat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} m_j(X_i)w_{i,j} + o_P\left(\frac{1}{\sqrt{n}}\right) = \int_0^1 m_j(Q_n)d\mathbb{K}_n + o_P\left(\frac{1}{\sqrt{n}}\right),$$

where $w_{i,j}$ is an indicator of subgroup $j$, $Q_n$ the empirical quantile function of $X_i$, and $\mathbb{K}_n$ the random measure that assigns density $w_{i,j}$ to $(\mathbb{F}_n(x_i) - 1/n, \mathbb{F}_n(x_i)]$ for the empirical distribution function $\mathbb{F}_n$ of $X_i$.[10]

Many early applications of multiple testing procedures in economics overlooked the issue of dependence among such test statistics and relied on procedures that assumed independence. But if we can estimate the joint distribution of these statistics, then we can safely rely on the multiple testing procedures that exploit the knowledge of the dependence structure, such as Yekutieli (2007) and Romano et al. (2008).

### 1.5.2 Testing Higher Degree Lorenz Dominance

For an income or wealth variable $X$ with quantile function $Q$, the Lorenz curve is the function

$$L_Q(\tau) := \frac{1}{\mathbb{E}[X]} \int_0^\tau Q(u)du.$$  

It is customary to interpret $L_Q(\tau)$ as the fraction of total income or wealth held by the lowest $\tau$-fraction (Lorenz, 1905; Csörgő, 1983). The value of the Lorenz curve at a specific point $\tau$ is called the Lorenz share at $\tau$ (Bhattacharya, 2005). The Gini coefficient is then defined as

$$G_Q := 1 - 2 \int_0^1 L_Q(\tau)d\tau = \frac{1}{\mathbb{E}[X]} \int_0^1 (2\tau - 1)Q(\tau)d\tau.$$  

[10] The (asymptotic) influence function of $X_i$ corresponds to the function $m_j(\cdot) + \beta_j$. Note that each function $m_j$ may differ as sample selection may change the influence function.
The Gini coefficient is one of the most popular inequality indices used in economics since its introduction by Gini (1912). Both the Lorenz curve and the Gini coefficient can be estimated by replacing $\mathbb{E}[X]$ and $Q$ with their sample analogues $\mathbb{E}_n[X]$ and $Q_n$. Such estimators are $L$-statistics.

The Lorenz curve is a continuous visualization of inequality over the income distribution. As such, comparing Lorenz curves across within-country, cross-country, or counterfactual income distributions has become a way to “uniformly” assess differences or changes in economic inequalities (Bishop et al., 1991, 1993; Morelli et al., 2015; Fellman, 2002). This led an important inequality comparison concept, *Lorenz dominance* (Dasgupta et al., 1973; Lambert, 2001). Namely, a Lorenz curve $L_1$ is said to *Lorenz dominate* another Lorenz curve $L_2$ if $L_1(\tau) \geq L_2(\tau)$ for every $u \in (0,1)$. If this is the case, the society with income distribution $L_1$ is considered to be “more equal” than that with $L_2$.

While conceptually simple and appealing, these concepts are criticized for being too restrictive; Lorenz curves often cross in data. Thus, in order to obtain a finer (partial) ordering of distributions, generalized versions of Lorenz dominance are proposed (Aaberge, 2009). The $k$th degree downward Lorenz curve puts more emphasis on income transfers to the poor and less emphasis on transfers to the rich; it is defined for some $k \geq 2$ by

$$L^k_Q(\tau) := \int_\tau^1 L^{k-1}_Q(u)du = \frac{1}{(k-1)!\mathbb{E}[X]} \int_\tau^1 (u - \tau)^{k-1}Q(u)du,$$

where $L^1_Q := 1 - L_Q$. The higher the value of $k$, the larger the emphasis put on the poor. We say that a Lorenz curve $L_1$ *$k$th degree downward Lorenz dominates* another $L_2$ if the corresponding $k$th degree downward Lorenz curve $L^k_1$ dominates $L^k_2$; intuitively, the society with income distribution $L_1$ is more equal than that with $L_2$ when additional emphasis is put on the poorer population. Likewise, the *$k$th degree*
upward Lorenz curve is defined by

$$\hat{L}_Q^k(\tau) := \int_0^\tau \hat{L}_Q^{k-1}(u)du = \frac{1}{(k-1)!}E[X] \int_0^\tau (\tau - u)^{k-1}Q(u)du,$$

where $\hat{L}_Q^1 := L_Q$, and puts more emphasis on the rich. The Lorenz curve $L_1$ kth degree upward Lorenz dominates another $L_2$ if $\hat{L}_1^k \geq \hat{L}_2^k$ uniformly. Again, the natural sample analogue estimators of higher degree Lorenz curves are $L$-statistics.

Comparison of Lorenz curves in applied work has relied mostly on visual inspection. To formally test dominance calls for uniform inference on the Lorenz curves; in particular, we require the joint distribution of every Lorenz share indexed by the income quantile $\tau \in (0, 1)$. The nonparametric uniform test of the standard (first degree) Lorenz dominance is recently established in the literature [Barrett et al., 2014]. The present chapter allows the extension of the uniform test to arbitrary degree Lorenz dominance.

The testing procedure is as follows. Since $(u - \tau)^{k-1} \leq 1$ for $\tau \leq u \leq 1$, we have

$$\int_\tau^1 |(u - \tau)^{k-1}Q|du \leq \int_0^1 |Q|du.$$

Therefore, the uniform convergence of the estimated $k$th degree downward Lorenz curve (uniformly over $\tau$) follows by the $L_1$ convergence of the empirical quantile $Q_n$. Suppose one has estimates of two $k$th degree downward Lorenz curves $\hat{L}_1^k$ and $\hat{L}_2^k$.

$$\hat{L}_j^k(\tau) := \frac{1}{(k-1)!}E_n[X_j] \int_\tau^1 (u - \tau)^{k-1}Q_{n,j}(u)du.$$

The assumptions posited here are that both income distributions $X_j$, $j = 1, 2$, have a $(2 + \varepsilon)$th moment for some $\varepsilon > 0$, and that both empirical quantile processes $\sqrt{n}(Q_{n,j} - Q_j)$, $j = 1, 2$, converge weakly in a joint manner. As in Barrett et al.

---

12 Bishop et al. (1988) and Arora and Jain (2006) present tests of (generalized) Lorenz dominance on finitely many points.
[2014], we consider the null hypothesis that $L_1$ dominates $L_2$:

$$H_0 : L_1^k(\tau) \geq L_2^k(\tau), \quad \forall \tau \in (0, 1).$$

Rewrite this hypothesis as

$$H_0 : \sup_{\tau \in (0,1)} (L_2^k - L_1^k) \leq 0.$$ 

Thus, the test statistic is $\sup_{\tau} (\hat{L}_2^k - \hat{L}_1^k)$, and the critical value $c_\alpha$ must satisfy

$$\Pr\left( \sup_{\tau \in (0,1)} (\hat{L}_2^k - \hat{L}_1^k) > c_\alpha \right) \leq \alpha$$

under the null hypothesis. Although involvement of a supremum makes analytic calculation of the asymptotic distribution difficult, we may rely on the nonparametric bootstrap to obtain the critical value.

Let $(X_{1,1}, X_{1,2}), \ldots, (X_{n,1}, X_{n,2})$ be the random variables representing households’ incomes in two different economic states (be it within-country, cross-country, or counterfactual). In particular, $X_{i,1}$ is the income of a household in one economy, and $X_{i,2}$ is that in another. We denote them as a pair $X_i = (X_{i,1}, X_{i,2})$ since they can be dependent (as in within-country or counterfactual comparison), but they can also be treated separately if they are clearly independent (as in cross-country comparison); they can even have different sample sizes. Let $X_1^*, \ldots, X_n^*$ denote the bootstrapped sample. Compute the bootstrap Lorenz curves by

$$\hat{L}_j^k(\tau) := \frac{1}{(k-1)!E_n[X_{i,j}]} \int_\tau^1 (u - \tau)^{k-1} Q_{n,j}^*(u)du.$$ 

Then, compute the quantity $\sup_{\tau} (\hat{L}_2^k - \hat{L}_1^k)$. We can use the bootstrap $(1 - \alpha)$-quantile of this quantity as the critical value for $\sup_{\tau} (\hat{L}_2^k - \hat{L}_1^k)$ to test for $k$th order Lorenz dominance.
1.5.3 Controlling Tail Risk with Estimation Error

In the context of risk measurement in finance, Chapter 3 develops a method to incorporate the estimation error into the risk to be estimated. For example, the expected shortfall is becoming popular in financial trading and banking regulation, which is defined as the expected return in the worst event of probability \( \alpha \), typically 5\%. Letting \( X \) be the return of a portfolio, the expected shortfall \( ES_\alpha \) is defined by

\[
\sup_{E \in \mathcal{F}} \left\{ \Pr(E) : \mathbb{E}[-X | E] \geq ES_\alpha \right\} \leq \alpha,
\]

where \( \mathcal{F} \) is the set of events. Algebra reveals that

\[
ES_\alpha = -\int_0^\alpha Q(u)du,
\]

where \( Q \) is the population quantile function of \( X \). The true expected shortfall cannot be observed, so in practice an estimated quantity is used. Suppose for simplicity that observations of returns are i.i.d.; then a natural estimator is (the negative of) the sample mean of observations of \( X \) below the \( \alpha \)-quantile, which is an \( L \)-statistic. The estimated expected shortfall, however, does not satisfy the above equation because of the estimation error. Instead, consider the \((1 - \alpha)\)-confidence set of the estimator and let \( ES_\alpha^\ast \) be its upper bound, that is,

\[
\Pr(ES_\alpha^\ast \geq ES_\alpha) \geq 1 - \alpha.
\]

Then, by the Bonferroni inequality,

\[
\sup_E \Pr(E \land \mathbb{E}[-X | E] \geq ES_\alpha^\ast) \leq \sup_E \left\{ \Pr(E) : \mathbb{E}[-X | E] \geq ES_\alpha \right\} + \alpha \leq 2\alpha.
\]

This enables us to control the risk (the probability of “bad” events) by an observable quantity \( ES_\alpha^\ast \). Chapter 3 generalizes this idea and defines a class of risk measures called the tail risk measure to which we can apply this bound. Many tail risk measures admit representations as \( L \)-statistics, and thus are susceptible to the use of our theory.
1.5.4 Assessing Outcome-Dependent Heterogeneity in Treatment Effects

Kaji (2017) proposes a method to assess outcome-dependent heterogeneity in treatment effects. Let $Y_0$ be the outcome of an individual in the control group and $Y_1$ be that of an individual in the treatment group. Letting $X$ represent the characteristics to control for, the conditional average treatment effect $E[Y_1 - Y_0 \mid X]$ is frequently used in empirical research, especially in ones with randomized controlled trials. Meanwhile, the average treatment effects conditional on outcome variables such as $E[Y_1 - Y_0 \mid Y_0 \in A]$ cannot be estimated since the joint distribution of $Y_1$ and $Y_0$ is not identified (Heckman et al., 1997). However, it often happens that treatment effects conditional outcomes are of interest. Taking the microcredit example from Section 1.2.2, if the treatment effect for households with originally high business profits ($Y_0$ is large) is positive and that for households with low profits ($Y_0$ is small) is negative, then even if the average treatment effect is positive, one may not wish to implement the microcredit. The common practice to assess such heterogeneity is the quantile treatment effect (Banerjee et al., 2015a; Augsburg et al., 2015; Tarozzi et al., 2015). However, quantile treatment effects cannot in general be interpreted as individual treatment effects. Kaji (2017) interprets quantile treatment effects as a distribution of individual treatment effects that attains the minimal sum of absolute individual treatment effects, and proposes the integral function of quantile treatment effects as an alternative measure of outcome-dependent subgroup treatment effects that has better asymptotic properties. These asymptotic results are derived using results of the present chapter.

1.6 Literature Review

We review three strands of the literature related to this chapter: (1) empirical and quantile processes in statistics/econometrics, (2) $L$-statistics in statistics and risk measures in finance, and (3) outlier detection and robust estimation in statistics.
1.6.1 Empirical and Quantile Processes

Central limit theorems in Banach spaces such as $L_p$ are a classical topic of which Ledoux and Talagrand (1991) provide an excellent exposition. It is known that tightness and the limiting properties of the Banach-valued random variables are closely tied to the structure and geometry of the Banach spaces. Among Banach spaces, however, the uniform space $L_\infty$ attracts independent attention, not only because of its own statistical importance, but for its mathematical complication epitomized by the fact that pre-Gaussianity alone does not immediately imply central limit theorems. Such difficulty called for direct characterization of asymptotic tightness and developed into rich literature—in an effort to show tightness—including tail bounds, entropy theory, and the Vapnik-Červonenkis theory (Van der Vaart and Wellner, 1996; Dudley, 2014). This chapter contributes to the literature by directly characterizing asymptotic tightness in $L_1$ in combination with $L_\infty$, enabling the establishment of weak convergence of processes that are not necessarily sample averages of i.i.d. Banach-valued random variables (thereby preventing the use of central limit theorems). In the particular context of this chapter, notwithstanding the i.i.d. assumption, we needed such characterization to show weak convergence of $K_n$.

The study of quantile processes is as old as that of empirical processes (Csörgő, 1983), but to the best of our knowledge the study is limited to quantile processes of bounded random variables and weighted (standardized) quantile processes (Shorack and Wellner, 1986; Csörgő and Horváth, 1988, 1990, 1993; Csörgő et al., 1986, 1993; Koul, 2002). This chapter is the first to show the weak convergence of raw quantile processes of unbounded random variables directly in $L_1$. It is also novel that the functional delta method is proved for the inverse map with norms replaced by more appropriate ones. In this respect, this work is related to Beutner and Zähle (2010); they consider the weighted sup norm on the space of distribution functions and establish the functional delta method for risk functionals. Their paper and ours share the similar idea that the use of a new norm gives a new functional delta method, but their work is closer in spirit to the literature on weighted suprema of empirical pro-
cesses. Although less clearly related, we note that a non-uniform norm for empirical processes has been occasionally considered in probability theory as well; e.g., Dudley (1997).

Some readers may wish to associate the results with the quantile regression literature popularized in economics (Koenker 2005). This literature, initiated by Koenker and Bassett (1978), reinvigorated the old work by Laplace (1812), and yielded many important results, including Koenker and Xiao (2002), Chernozhukov (2005), Chernozhukov and Hansen (2005), Angrist et al. (2006), Firpo et al. (2009), Chernozhukov et al. (2010), and Belloni and Chernozhukov (2011). There is an obvious relationship between the conditional mean and conditional quantiles when some simplifying assumptions hold. Let $y_i = x_i' \beta + \varepsilon_i$ and $E[\varepsilon_i \mid x_i] = 0$ for every $x_i$. Then, by the change of variables,

$$x_i' \beta = E[y_i \mid x_i] = \int_0^1 Q_{y_i}(u \mid x_i) du = \int_0^1 x_i' \beta(u) du.$$  

So, if $x_i$ is sufficiently rich, it must be that $\beta = \int_0^1 \beta(u) du$. Although this type of relationship is known and used in survival analysis (Cox et al., 2013) (and a remotely related one used in economics; Chernozhukov et al., 2013), it does not offer much to our purpose, as there is no guarantee that the relationship continues to hold in their sample analogues. It is, however, an important direction of future research to examine if weak convergence in $L_1$ takes place for quantile regression estimators on the whole of $(0, 1)$.

Also, this chapter does not explicitly consider weakly dependent samples, although as noted at the end of Section 1.3.3 it would be straightforward to extend the results to subsume such cases, e.g., by incorporating results of Dehling et al. (2002, 2014). Other extensions potentially useful for measuring financial risk are the application of extreme value theory for the tail empirical processes (Einmahl, 1992; Rootzén, 2009; Drees and Rootzén, 2010, 2016) or the use of smoothed or other explicitly estimated empirical distributions (Hall et al., 1999, Berg and Politis, 2009).

\footnote{One notable difference from our setup is that they involve optimization over a class of functions to obtain a process.}
1.6.2 \textit{L}-Statistics

\textit{L}-statistics are an old topic in statistics, especially in the study of location estimation \cite{van1998}, Chapter 22). There are two major ways to prove the asymptotic normality of \textit{L}-statistics: the Hájek projection and the functional delta method. The difficulty of showing asymptotic normality lies in that the summands are intercorrelated with each other in a complicated way. The Hájek projection projects the summands onto the space of independent observations, thereby pulling the situation back to ones of classical central limit theorems. This requires, however, an effort to find the projection and to show that the residual of the projection goes away. This can be a hard task when the statistic of interest involves complicated estimation procedures or comes from nontrivial structural models.

The functional delta method, on the other hand, directly deals with the complicated intercorrelation in the raw form of an empirical quantile function. Therefore, it is more general than the Hájek projection, yields simple representation of the asymptotic distribution, and proves the validity of bootstrap at much less or no cost. The cons of this method are that the empirical processes literature usually requires uniform convergence, which unavoidably entails boundedness of the processes. All this has led to standardization methods using bounded quantile processes, as seen in the Chibisov-O’Reilly theorem. This chapter, in combination with giving $L_1$ convergence of quantile processes, tackles this thorny issue by extending the functional delta method to $L_1$ processes and \textit{L}-statistics.

1.6.3 Outlier Detection and Robust Estimation

In this chapter, we refer to observations lying in either tail as “outliers.” The classical subfield of statistics, \textit{outlier detection}, defines outliers instead as observations arising from a different unknown data-generating process and thereby subject to elimination \cite{hawkins1980, aggarwal2013}. Starting from a null hypothesis about the true data-generating process, the literature develops a way to detect observations that fall outside of the behavior permissible under the null. Distributional assumptions in the
null hypothesis may vary from a complete description of the data generating process to only the tail behaviors to the proximity or temporal models. Outlier detection in a regression framework is also studied by Chatterjee and Hadi (1988) and Gujarati and Porter (2009).

Despite the concern for outliers and their removal, this chapter has not much to share with this literature. This chapter does not make assumptions on distributions but on existence of moments, and the null we aim to reject is about specific parameters rather than outliers themselves. Nevertheless, if one dares to draw a connection, one can say that this chapter provides a new way to formulate the null hypothesis in outlier detection. If one has a particular parameter in mind that should not be largely affected by any few observations, then by conducting the outlier robustness test for that parameter one can detect the outliers when the test is rejected.

The final literature we review is on robust estimation in statistics (Hampel et al., 1986; Koul, 2002; Maronna et al., 2006; Huber and Ronchetti, 2009). These works concern estimators that are robust against deviations from an ideal model, especially when highly influential (erroneous) outliers are introduced. Similarly to this chapter, they mostly consider estimators whose deviation from the true parameters is represented by the sum of influence functions of observations. To estimate asymptotic distributions of robust estimators, they often rely on empirical process theory (Koul, 2002, Chapter 4); in this sense, this chapter can be considered an extension of their asymptotic theory. Their motivation goes further in finding the best influence function under some criteria and construct the best estimator for the purpose of robust estimation. Although such robust estimation has been considered in the economics discipline (Krasker, 1980), it is rather a minor subject—possibly due to the resulting estimators’ non-straightforward interpretability.

1.7 Conclusion

Motivated by a need for formal tests of outlier robustness, this chapter develops substantially generalized asymptotic theory of $L$-statistics. In particular, observing that
essential for convergence of $L$-statistics is not the uniform convergence of empirical quantile processes but $L_1$ convergence, we establish the theory of $L$-statistics through the development of the theory of integrable empirical processes. The highlights of this theoretical development are the new norms introduced to the spaces of functions.

First, we consider distribution functions in the space of bounded integrable functions. Distribution functions need to converge uniformly in order for their inverses to be well-defined, and they also need to converge in $L_1$ in order for their inverses (quantile functions) to be integrable. We characterize weak convergence in this space by asymptotic uniform equicontinuity and equiintegrability in probability. Uniform equicontinuity is needed for uniform convergence, and equiintegrability for $L_1$ convergence. Using this, we show that empirical processes converge in this norm if the underlying distribution $F$ has a $(2 + \varepsilon)$th moment for some small $\varepsilon > 0$.

Second, we consider quantile functions in the space of integrable functions, and derive weak convergence using the functional delta method. The key to the proof is the compatibility of the $L_1$ norm with Fubini’s theorem. This is in contrast to classical results such as the Chibisov-O’Reilly theorem that use the $L_2$ norm.

Then, we consider sample selection functions in the set of Lipschitz functions in the space of bounded integrable functions. We need the Lipschitz property to make $\int Q_n dK_n$ converge whenever $\int Q_n du$ does, boundedness to ensure that the Lebesgue-Stieltjes integral with respect to $K$ is well-defined, and integrability to ensure convergence of the integral itself. We derive weak convergence of sample selection functions by another application of our earlier results.

Finally, we derive weak convergence of $L$-statistics using the functional delta method on the map from quantile functions and sample selection functions to $L$-statistics. This can be seen as a generalization of the results on Wilcoxon statistics to subsume unbounded functions. As a byproduct of our functional delta method approach, we derive validity of nonparametric bootstrap.

Using our results, we construct a formal test of outlier robustness analysis. We apply our test to Acemoglu et al. (2017) and contrast heuristic arguments to formal tests. For one of the parameters, we “discover” sensitivity to outliers that could not
have been discovered by heuristics.

Our theory of $L$-statistics is itself new and of independent interest. As applications other than outlier robustness analysis, we explained multiple testing problems, tests of higher-order Lorenz dominance, risk control by tail risk measures in Chapter 3 and estimation of bounds on outcome-dependent treatment effects by Kaji (2017).
Appendices

1.A  Figures and Tables
(a) Empirical quantile function and sample selection functions for OLS estimators. $\frac{1}{T} \sum_t \text{Democracy}_{i,t} \times \log GDP_{i,t}$ [in 100 units].

(b) Empirical quantile function and sample selection functions for OLS estimators. $\frac{1}{T} \sum_t \log GDP_{i,t-1} \times \log GDP_{i,t}$ [in 100,000 units].

(c) Empirical quantile function and sample selection functions for IV estimators. $\frac{1}{T} \sum_t \text{WaveOfDemocracy}_{i,t} \times \log GDP_{i,t}$ [in 100 units].

(d) Empirical quantile function and sample selection functions for IV estimators. $\frac{1}{T} \sum_t \log GDP_{i,t-1} \times \log GDP_{i,t}$ [in 100,000 units].

Figure 1.A.1: Empirical quantile functions and sample selection functions for OLS and IV estimators for Acemoglu et al. (2017).
Figure 1.A.2: Distributions of residuals of OLS and IV regressions in \cite{Acemoglu2017}. Clusters at the boundaries indicate how many observations fall outside of the range.
Table 1.A.1: Estimates and comparison of formal and heuristic $p$-values for the outlier robustness test in "Acemoglu et al. (2017)."

<table>
<thead>
<tr>
<th>Notation</th>
<th>OLS Estimate</th>
<th>IV Estimate</th>
<th>$p$-value for $H_0: \beta_1^* = \beta_2^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Democracy</td>
<td>0.79 (0.23)</td>
<td>0.56 (0.20)</td>
<td>1.15 (0.59)</td>
</tr>
<tr>
<td>log GDP first lag</td>
<td>1.24 (0.04)</td>
<td>1.23 (0.02)</td>
<td>1.24 (0.04)</td>
</tr>
<tr>
<td>log GDP second lag</td>
<td>-0.21 (0.05)</td>
<td>-0.20 (0.03)</td>
<td>-0.21 (0.05)</td>
</tr>
<tr>
<td>log GDP third lag</td>
<td>-0.03 (0.03)</td>
<td>-0.03 (0.02)</td>
<td>-0.03 (0.03)</td>
</tr>
<tr>
<td>log GDP fourth lag</td>
<td>-0.04 (0.02)</td>
<td>-0.03 (0.02)</td>
<td>-0.04 (0.02)</td>
</tr>
<tr>
<td>Long-run effect of democracy</td>
<td>21.24 (7.32)</td>
<td>19.32 (8.54)</td>
<td>31.52 (18.49)</td>
</tr>
<tr>
<td>Effect of democracy after 25 years</td>
<td>16.90 (5.32)</td>
<td>13.00 (5.02)</td>
<td>24.87 (13.53)</td>
</tr>
<tr>
<td>Persistence of GDP process</td>
<td>0.96 (0.01)</td>
<td>0.97 (0.005)</td>
<td>0.96 (0.01)</td>
</tr>
<tr>
<td>Number of observations</td>
<td>6,336</td>
<td>6,044</td>
<td>6,309</td>
</tr>
<tr>
<td>Number of countries</td>
<td>175</td>
<td>175</td>
<td>174</td>
</tr>
<tr>
<td>Minimum number of years</td>
<td>6</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Maximum number of years</td>
<td>47</td>
<td>47</td>
<td>47</td>
</tr>
<tr>
<td>Average number of years</td>
<td>36.2</td>
<td>34.5</td>
<td>36.3</td>
</tr>
<tr>
<td>Median number of years</td>
<td>42</td>
<td>42</td>
<td>42</td>
</tr>
<tr>
<td>Number of bootstrap iterations</td>
<td>10,000</td>
<td>10,000</td>
<td>10,000</td>
</tr>
</tbody>
</table>

* (1) and (3) Baseline estimates; (2) Estimates on sample $|\hat{\epsilon}_{i,t}| < 1.96 \hat{\sigma}_\epsilon$; (4) Estimates on sample $|\hat{\epsilon}_{i,t}| < 1.96 \hat{\sigma}_\epsilon$ and $|\hat{v}_{i,t}| < 1.96 \hat{\sigma}_v$; (5) and (7) $p$-values of the formal tests that use the standard errors of $\hat{\beta}_1^* - \hat{\beta}_2^*$; (6) and (8) "$p$-values" of the heuristic tests that use the marginal standard errors of $\hat{\beta}_1^*$. Some numbers in Columns (1) and (2) differ slightly from "Acemoglu et al. (2017)" since our own bootstrap results are used for standard errors.
(a) Distribution of full-sample and outlier-removed OLS estimators for the effect of democracy $\beta_0$. $p = 0.15$.

(b) Distribution of full-sample and outlier-removed OLS estimators for persistence of GDP $\beta_7$. $p = 0.0002$.

(c) Distribution of full-sample and outlier-removed IV estimators for the effect of democracy $\beta_0$. $p = 0.20$.

(d) Distribution of full-sample and outlier-removed IV estimators for persistence of GDP $\beta_7$. $p = 0.004$.

Figure 1.A.3: Joint distributions of full-sample and outlier-removed OLS and IV estimators for Acemoglu et al. [2017]. Outliers are defined by $|\hat{\varepsilon}_{i,t}| \geq 1.96 \hat{\sigma}_e$ or $|\hat{v}_{i,t}| \geq 1.96 \hat{\sigma}_v$. The black dotted lines indicate the 45-degree line. Nonparametric bootstrap is repeated for 10,000 times, randomly sampling across $i$. The contours drawn are of kernel density estimators.
Figure 1.A.4: Histogram of numbers of removal that countries had in 10,000 iterations of OLS bootstrap.

(a) Histogram of numbers of removal that countries had in 10,000 iterations of OLS bootstrap.

(b) Histogram of numbers of removal that countries had in 10,000 iterations of IV bootstrap.

Figure 1.A.4: Histogram of numbers of removal that countries had in 10,000 iterations of bootstrap for Acemoglu et al. (2017).
1.B Proofs of Results Stated in the Main Text

Proof of Theorem 1.1. If \((a, b)\) is totally bounded in \(\rho_1\) and \(\rho_2\), then so it is in \(\rho := \rho_1 \lor \rho_2\). Then the theorem follows in combination of Theorems 1.D.4 and 1.D.5. ■

Proof of Proposition 1.2. This proposition is proved as Proposition 1.D.1. ■

Proof of Theorem 1.3. This theorem is proved as Lemma 1.E.1 and its extension in Theorem 1.E.3. ■

Proof of Proposition 1.4. This proposition is proved as Proposition 1.E.4. ■

Proof of Theorem 1.5. This theorem is proved as Theorem 1.F.1. ■

Proof of Proposition 1.6. This proposition is proved as Proposition 1.F.3. ■

Proof of Proposition 1.7. This proposition is proved as Proposition 1.F.5. ■

Proof of Proposition 1.8. This proposition is proved as Proposition 1.G.5. ■

1.C Mathematical Preliminaries

The inverse function \(f^{-1} : \mathbb{R} \to \mathbb{R}\) of a function \(f : \mathbb{R} \to \mathbb{R}\) is defined by the left-continuous generalized inverse, i.e.,

\[
f^{-1}(y) := \inf\{x \in \mathbb{R} : f(x) \geq y\}.
\]

This inverse is often denoted by \(f^{\leftarrow}\) or \(f^{-}\) in the literature (Dudley 1997; Embrechts and Hofert 2013). While we keep the notation \(f^{-1}\) for this, when we refer to the right-continuous generalized inverse, we use the notation \(f^{\rightarrow}\), that is,

\[
f^{\rightarrow}(y) := \sup\{x \in \mathbb{R} : f(x) \leq y\}.
\]

For properties of generalized inverses, see Feng et al. (2012) and Embrechts and Hofert (2013).
Let $-\infty \leq a < b \leq +\infty$. The Lebesgue-Stieltjes measure $\mu$ on $(a, b)$ associated with an increasing function $m : (a, b) \to \mathbb{R}$ assigns to an open interval $(c, d)$ the measure $m(d-) - m(c+)$, where $m(-)$ is the left limit and $m(\cdot +)$ the right limit. Conversely, a function $m : (a, b) \to \mathbb{R}$ associated with the Lebesgue-Stieltjes measure $\mu$ on $(a, b)$ is any function such that $\mu((c, d]) = m(d) - m(c)$ for every $c < d$\footnote{Note that the function constructed from a Lebesgue-Stieltjes measure is cadlag, while the Lebesgue-Stieltjes measure can be given to any increasing but not necessarily cadlag function. This asymmetry introduces minor adjustments to the change of variables for Lebesgue-Stieltjes integrals. See \cite{Falkner and Teschl 2012}.}. Because of this relationship, we often denote both the function and the measure by the same letter.

The following lemma is used throughout Appendices.

\textbf{Lemma 1.C.1.} Let $F$ be a probability distribution on $\mathbb{R}$, $\tilde{F}(x) := F(x) - 1\{x \geq 0\}$, and $Q := F^{-1}$ the quantile function. For $p > 0$ we have (i) $\iff$ (ii) $\iff$ (iii) $\implies$ (iv) $\iff$ (v), where

1. $F$ has a $p$th moment;
2. $Q$ is in $L_p(0, 1)$;
3. $|x|^{p-1} \tilde{F}$ is integrable;
4. $|x|^p \tilde{F}$ converges to 0 as $x \to \pm \infty$;
5. $u^{1/p}(1-u)^{1/p}Q$ converges to 0 as $u \to \{0, 1\}$.

\textit{Proof.} We prove the following directions in order: (i) $\implies$ (iv), (iii) $\implies$ (iv), (i) $\iff$ (iii), (i) $\iff$ (ii), and (iv) $\iff$ (v). The second claim seems unnecessary, but will be used in proving the third claim.

(i) $\implies$ (iv). For $M > 0$, note that

$$\int_{\mathbb{R}} |x|^p dF \geq \int_{[-M, M]} |x|^p dF + M^p |\tilde{F}(-M)| + M^p |\tilde{F}(M)|.$$
Since the left-hand side (LHS) is finite, one may take $M$ large enough that
\[
\int_{\mathbb{R}} |x|^p F - \int_{[-M,M]} |x|^p dF
\]
is smaller than an arbitrarily small positive number, which then bounds the two nonnegative terms. Hence $|x|^p \bar{F}(x) \to 0$ as $x \to \pm \infty$.

(iii) ⇒ (iv). Suppose that $|x|^{p-1} |\bar{F}|$ is integrable but $|x|^p F$ does not vanish as $x \to -\infty$, that is, there exist a constant $c > 0$ and a sequence $0 > x_1 > x_2 > \cdots \to -\infty$ such that $|x_i|^p F(x_i) \geq c$. Since $F \to 0$, one may take a subsequence such that
\[
|x_i|^p F(x_{i+1}) \leq 2^{-i}.
\]
By monotonicity of $F$,
\[
p \int_{-\infty}^0 |x|^{p-1} F(x) dx \geq F(x_1) \int_{x_1}^0 p |x|^{p-1} dx + F(x_2) \int_{x_2}^{x_1} p |x|^{p-1} dx + \cdots
\]
\[
= |x_1|^p F(x_1) + (|x_2|^p - |x_1|^p) F(x_2) + (|x_3|^p - |x_2|^p) F(x_3) + \cdots
\]
\[
\geq c + \sum_{i=1}^{\infty} (c - 2^{-i}) = \infty,
\]
which is a contradiction. Hence $|x|^p F$ must vanish. Deduce similarly that $|x|^p (1 - \bar{F}) \to 0$ as $x \to +\infty$.

(i) ⇔ (iii). Note that $dF = d\bar{F}$ for $x \neq 0$. Integration by parts yields
\[
\int_{\mathbb{R}} |x|^p dF = \int_{\mathbb{R}} |x|^p d\bar{F} = \left[ |x|^p \bar{F} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} |x|^{p-1} |\bar{F}| dx.
\]
If the LHS is finite (i), then the first term in the right-hand side (RHS) is 0 (iv) and hence the second term is finite (iii). Conversely, if the second term is finite (iii), then the first term is 0 (iv) and hence the LHS is finite (i).

(ii) ⇔ (iv). By the change of variables,
\[
\int_{\mathbb{R}} |x|^p dF = \int_0^1 |Q|^p du.
\]
Hence the LHS is finite if and only if the RHS is.

\[(\text{iv}) \iff (\text{v}).\] Let \(u = F(x)\). Then, \(\lim_{x \to -\infty} |x|^p \tilde{F} = \lim_{u \to 0} (u^{1/p} Q)^p = 0\). Convergence of the other tail can be shown analogously.

\[\Box\]

**Remark.** One-sided implication in the lemma is strict. One can construct \(\tilde{F}\) such that \(\text{(iv)}\) holds but \(\text{(iii)}\) does not. Let \(\tilde{F}\) satisfy \(|x|^{p-1} \tilde{F} \approx 1/p|x|\) where \(p_n\) denotes the \(n\)th prime number. Then, it is not integrable since the sum of the reciprocals of the primes diverges, but \(|x|^p \tilde{F} \to 0\) since the primes only increase at a logarithmic speed.

**Remark.** Similar properties in norms play an important role in characterizing the asymptotic behaviors in general Banach spaces (Ledoux and Talagrand, 1991).

### 1.D Tightness of Bounded Integrable Processes

First, we consider the stochastic processes that are integrable with respect to a general measure. The main objective of this section is to develop the conditions for the sequence of integrable processes to converge weakly in the corresponding \(L_1\) space. Most exposition of this section parallels the flow of arguments of Van der Vaart and Wellner (1996, Chapter 1.5).

**Definition.** Let \((T, \mathcal{T}, \mu)\) be a measure space where \(T\) is an arbitrary set, \(\mathcal{T}\) a \(\sigma\)-field on \(T\), and \(\mu\) a \(\sigma\)-finite signed measure on \(\mathcal{T}\). Let \(\mathbb{L}_\mu\) be the Banach space of bounded and \(\mu\)-integrable functions \(z : T \to \mathbb{R}\), that is,

\[
\|z\|_{\mathbb{L}_\mu} := \|z\|_T \vee \|z\|_\mu := \left(\sup_{t \in T} |z(t)|\right) \vee \left(\int_T |z| \, |d\mu|\right) < \infty,
\]

where \(|d\mu|\) represents integration with respect to the total variation measure of \(\mu\).

In the main text, special cases of this are used for the distribution functions and the sample selection measures. General construction allows us to accommodate many other cases, including the following.

\[15\text{Integration with respect to the total variation measure is often denoted with } d|\mu|. \text{ However, because we sometimes mix Lebesgue-Stieltjes integrals, we denote the total variation integration by } |d\mu| \text{ so it not be confused with the Lebesgue-Stieltjes integration with respect to the “function” } |\mu|.\]
Example \((p\text{th moment})\). Let \(F_n\) be the empirical distribution of a real-valued random variable. By integration by parts, the sample \(p\text{th moment}\) is given by

\[
\int x^p dF_n = -\int \tilde{F}_n dx^p \quad \text{where} \quad \tilde{F}_n(x) = \begin{cases} 
F_n(x) & x < 0, \\
F_n(x) - 1 & x \geq 0.
\end{cases}
\]

Then it is natural to consider \(\tilde{F}_n\) as a stochastic process integrable with respect to the \(\sigma\)-finite signed measure \(\mu((a, b]) := b^p - a^p\) on \(\mathbb{R}\). \(\Box\)

For processes represented as the sum of i.i.d. random variables, such as the empirical process \(\sqrt{n}(F_n - F)\) itself, one can easily prove weak convergence in this space by the combination of classical central limit theorems (CLT) \((\text{Van der Vaart and Wellner 1996, Dudley 2014, Ledoux and Talagrand 1991})\), as shown in the next proposition. However, for other types of processes that are not an average of i.i.d. variables, notably the random measure process \(\sqrt{n}(\mathbb{K}_n - K)\) in this chapter, we are unable to resort to CLT-type results. Therefore a more general, direct way of showing weak convergence will be developed subsequently.

**Proposition 1.D.1.** Let \(m : \mathbb{R} \to \mathbb{R}\) be a function of locally bounded variation and \(\mu\) the Lebesgue-Stieltjes measure associated with \(m\). For a probability distribution \(F\) on \(\mathbb{R}\) such that \(m(X)\) has a \((2 + c)\)th moment for \(X \sim F\) and some \(c > 0\), the empirical process \(\sqrt{n}(F_n - F)\) converges weakly in \(L_\mu\) to a Gaussian process with mean zero and covariance function \(\text{Cov}(x, y) = F(x \wedge y) - F(x)F(y)\).

**Proof.** The marginal convergence is trivial. According to \(\text{Van der Vaart and Wellner 1996, Example 2.5.4}\), the empirical process \(\sqrt{n}(F_n - F)\) converges weakly in \(L_\infty\). In light of \(\text{Van der Vaart and Wellner 1996, Proposition 2.1.11}\), it suffices to show that for \(X_i \sim F\) and \(Z_i(x) := 1\{X_i \leq x\} - F(x)\),

\[
\Pr(\|Z_i\|_\mu > t) = o(t^{-2}) \quad \text{as} \quad t \to \infty
\]

and

\[
\int_{\mathbb{R}} (\mathbb{E}[Z_i(x)^2])^{1/2} d\mu = \int_{\mathbb{R}} (F(x)[1 - F(x)])^{1/2} d\mu < \infty.
\]

73
Since a function of locally bounded variation can be written as the difference of two increasing functions, $m$ can be assumed without loss of generality increasing, that is, $\mu$ be a positive measure. Observing $Z_i(x) = (\mathbb{1}\{X_i \leq x\} - \mathbb{1}\{0 \leq x\}) - (F(x) - \mathbb{1}\{0 \leq x\})$, find
\[
\|Z_i\|_\mu \leq |m(X_i+) - m(0+)| + \int_{\mathbb{R}} |\tilde{F} \circ m^{-1}| \, dx.
\]
Therefore, the first condition is satisfied if $\tilde{F} \circ m^{-1}(t) = o(t^{-2})$, which is the case if $m(X_i)$ has a variance by Lemma 1.C.1. Secondly, if $m(X_i)$ has a $(2+c)$th moment, then by Lemma 1.C.1 again, $|x|^{1+c} \tilde{F} \circ m^{-1}$ is integrable and $|x|^{2+c} \tilde{F} \circ m^{-1} \to 0$ as $x \to \pm \infty$. Therefore, $\tilde{F} \circ m^{-1}(x) = o(1/|x|^{2+c})$ and hence $[(F \circ m^{-1})(1 - F \circ m^{-1})]^{1/2}$ is integrable, which means that $[F(1 - F)]^{1/2}$ is integrable with respect to $\mu$. Thus, the empirical process $\sqrt{n}(F_n - F)$ converges weakly in $L_1(\mu)$ as well, as desired. ■

As in the classical empirical process literature, we first characterize weak convergence in $L_\mu$ by asymptotic tightness plus weak convergence of marginals. For this purpose, we consider a sequence of random elements taking values in $L_\mu$, that is, $X_n : \Omega \to L_\mu$. Following Van der Vaart and Wellner (1996), there are some generalizations we allow in our setup. We consider the generalized version of a sequence, a net $X_\alpha$ indexed by an arbitrary directed set, rather than a sequence $X_n$ indexed by natural numbers. Note that a sequence is a special case of a net. Moreover, we allow the sample space $\Omega$ to be different for each element in a net, that is, we consider $X_\alpha : \Omega_\alpha \to L_\mu$. Finally, we allow that each element in the net is not necessarily measurable, so, when we seek rigor, we shall call $X_\alpha$ a net of “arbitrary maps from $\Omega_\alpha$ to $L_\mu$” in lieu of “random elements taking values in $L_\mu$.” There is also a caveat on the notation: when we write $X(t)$ for a map $X : \Omega \to L_\mu$, $t$ is understood to be an element of $T$ and we regard $X(t)$ as a map from $\Omega$ to $\mathbb{R}$ indexed by $T$; when we explicitly use $\omega \in \Omega$ in the discussion, we write $X(t, \omega)$.

The following lemmas and theorem characterize weak convergence in our space.

**Lemma 1.D.2.** Let $X_\alpha : \Omega_\alpha \to L_\mu$ be asymptotically tight. Then, it is asymptotically measurable if and only if $X_\alpha(t)$ is asymptotically measurable for every $t \in T$.

\(^{16}\) See Falkner and Teschl (2012) for the change of variables formula.
Lemma 1.D.3. Let $X$ and $Y$ be tight Borel measurable maps into $\mathbb{L}_\mu$. Then, $X$ and $Y$ are equal in Borel law if and only if all corresponding marginals of $X$ and $Y$ are equal in law.

Theorem 1.D.4. Let $X_\alpha : \Omega_\alpha \to \mathbb{L}_\mu$ be arbitrary. Then, $X_\alpha$ converges weakly to a tight limit if and only if $X_\alpha$ is asymptotically tight and the marginals $(X_\alpha(t_1), \ldots, X_\alpha(t_k))$ converge weakly to a limit for every finite subset $t_1, \ldots, t_k$ of $T$. If $X_\alpha$ is asymptotically tight and its marginals converge weakly to the marginals $(X(t_1), \ldots, X(t_k))$ of a stochastic process $X$, then there is a version of $X$ with sample paths in $\mathbb{L}_\mu$ and $X_\alpha \Rightarrow X$.

Proofs. Since our norm is stronger than the uniform norm, Lemmas 1.D.2 and 1.D.3 follow as corollaries to Van der Vaart and Wellner (1996, Lemmas 1.5.2 and 1.5.3). Now we prove the theorem.

Necessity is immediate. We prove sufficiency. If $X_\alpha$ is asymptotically tight and its marginals converge weakly, then $X_\alpha$ is asymptotically measurable by Lemma 1.D.2. By Prohorov’s theorem (Van der Vaart and Wellner, 1996, Theorem 1.3.9), $X_\alpha$ is relatively compact. Take any subnet in $X_\alpha$ that is convergent. Its limit point is unique by Lemma 1.D.3 and the assumption that every marginal converges weakly. Thus, $X_\alpha$ converges weakly. The last statement is another consequence of Prohorov’s theorem.

Although we consider a different norm, our space contains the same elements as the classical literature (e.g., empirical processes). Therefore, weak convergence of marginals can easily be established by the classical results such as the multivariate central limit theorems. Hence the question that remains is how to establish asymptotic tightness.

The space of interest $\mathbb{L}_\mu$ is the intersection of the uniform space (with respect to the norm $\| \cdot \|_U$) and the $L_1$ space (with respect to the norm $\| \cdot \|_\mu$). As such, asymptotic tightness in $\mathbb{L}_\mu$ is equivalent to joint satisfaction of asymptotic tightness in each space. Again, tightness in the uniform space can be established by classical results. Following Van der Vaart and Wellner (1996), we characterize tightness in $L_1$ in two ways:
through the finite approximation and by the Dunford-Pettis theorem. The second characterization connects asymptotic tightness in $L_1$ to asymptotic equiintegrability of the sample paths. In light of this, define the following.

**Definition.** For a $\mu$-measurable semimetric $\rho$ on $T$\footnote{We define a semimetric to be $\mu$-measurable if every open set thereby induced is measurable with respect to $\mu$.} a function $f : T \to \mathbb{R}$ is uniformly $\rho$-continuous and $(\rho, \mu)$-integrable if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every $t \in T$

$$\left( \sup_{\rho(s, t) < \delta} |f(s) - f(t)| \right) \vee \left( \int_{0 < \rho(s, t) < \delta} |f(s)||d\mu(s)| \right) < \varepsilon.$$

**Definition.** For a $\mu$-measurable semimetric $\rho$ on $T$, the net $X_\alpha : \Omega_\alpha \to \mathbb{L}_\mu$ is asymptotically uniformly $\rho$-equicontinuous and $(\rho, \mu)$-equiintegrable in probability if for every $\varepsilon, \eta > 0$ there exists $\delta > 0$ such that

$$\limsup_{\alpha} P^\ast \left( \sup_{t \in T} \left[ \left( \sup_{\rho(s, t) < \delta} |X_\alpha(s) - X_\alpha(t)| \right) \vee \left( \int_{0 < \rho(s, t) < \delta} |X_\alpha(s)||d\mu(s)| \right) \right] > \varepsilon \right) < \eta.$$

**Remark.** The nomenclature equiintegrable is based on the fact that the standard definition of $\mu$-equiintegrability (or “uniform” $\mu$-integrability) in functional analysis roughly coincides with $(\mu, \mu)$-equiintegrability defined herein, albeit $\mu$ is not a metric. Here we prefer the prefix equi- over the (arguably more popular) qualifier uniformly for the meaning “equally among the class of functions” in order to maintain coherence with uniformly equicontinuous.

**Theorem 1.D.5.** The following are equivalent.

(i) A net $X_\alpha : \Omega_\alpha \to \mathbb{L}_\mu$ is asymptotically tight.

(ii) $X_\alpha(t)$ is asymptotically tight in $\mathbb{R}$ for every $t \in T$, $\|X_\alpha\|_\mu$ is asymptotically tight in $\mathbb{R}$, and for every $\varepsilon, \eta > 0$ there exists a finite $\mu$-measurable partition...
\[ T = \bigcup_{i=1}^{k} T_i \text{ such that} \]

\[
\limsup_{\alpha} P^* \left( \sup_{1 \leq i \leq k} \sup_{s, t \in T_i} |X_\alpha(s) - X_\alpha(t)| \right) \vee \sum_{i=1}^{k} \inf_{x \in \mathbb{R}} \int_{T_i} |X_\alpha - x| \, d\mu > \varepsilon \right) < \eta.
\] (1.4)

(iii) \( X_\alpha(t) \) is asymptotically tight in \( \mathbb{R} \) for every \( t \in T \) and there exists a \( \mu \)-measurable semimetric \( \rho \) on \( T \) such that \( (T, \rho) \) is totally bounded and \( X_\alpha \) is asymptotically uniform \( \rho \)-equicontinuous and \( (\rho, \mu) \)-equiintegrable in probability.

If, moreover, \( X_\alpha \rightharpoonup X \), then almost all paths \( t \mapsto X(t, \omega) \) are uniformly \( \rho \)-continuous and \( (\rho, \mu) \)-integrable; and the semimetric \( \rho \) can without loss of generality be taken equal to any semimetric \( \rho \) for which this is true and \( (T, \rho) \) is totally bounded.

Remark. The condition on the supremum “\( 0 < \rho(s, t) \)” is to allow for the point masses in \( \mu \) and plateaus in \( X_\alpha \). In (1.4), this condition corresponds to the subtraction of “\( x \)’.”

Proof. We prove (ii) \( \Rightarrow \) (i) \( \Rightarrow \) (iii) \( \Rightarrow \) (ii), and then the addendum.

(ii) \( \Rightarrow \) (i). Fix \( \varepsilon, \eta > 0 \) and take the given partition \( T_1, \ldots, T_k \). Pick one \( t_i \) from each \( T_i \). Then, \( \|X_\alpha\|_T \leq \max_i |X_\alpha(t_i)| + \varepsilon \) with inner probability at least \( 1 - \eta \).

Since the maximum of finitely many tight nets of real variables is tight and \( \|X_\alpha\|_\mu \) is assumed to be tight, it follows that the net \( \|X_\alpha\|_{L_\mu} \) is asymptotically tight in \( \mathbb{R} \).

Fix \( \zeta > 0 \) and take \( \varepsilon_{m, \gamma} > 0 \). Let \( M \) be a constant such that \( \limsup P^*(\|X_\alpha\|_{L_\mu} > M) < \zeta \). Taking \( (\varepsilon, \eta) \) in (1.4) to be \( (\varepsilon_{m, 2^{-m-1}}, \zeta) \), we obtain for each \( m \) a measurable partition \( T = \bigcup_{i=1}^{k} T_i \) (suppressing the dependence on \( m \)). For each \( T_i \), enumerate all of the finitely many values \( 0 = a_{i,0} \leq a_{i,1} \leq \cdots \leq a_{i,p} \leq M \) such that

\[
\int_{T_i} (a_{i,j} - a_{i,j}) |d\mu| \leq \frac{\varepsilon_{m}}{k} \quad \text{for} \quad j = 1, \ldots, p \quad \text{and} \quad \int_{T_i} |a_{i,j}| |d\mu| \leq M.
\]

Since \( \mu \) is not necessarily finite on the whole \( T \), it can be that on some partition \( T_i \) the only possible choice of \( a_{i,j} \) is 0. Let \( z_1, \ldots, z_q \) be the finite exhaustion of all
functions in $\mathbb{L}_\mu$ that are constant on each $T_i$ and take values only on

$$0, \pm \varepsilon_m, \ldots, \pm \lfloor M/\varepsilon_m \rfloor \varepsilon_m, \pm a_{1,1}, \ldots, \pm a_{1,p}, \ldots, \pm a_{k,1}, \ldots, \pm a_{k,p}.$$ 

Again, it can be that on a partition with $\int_{T_i} |d\mu| = \infty$, the only value that $z_i$ can take is 0. Let $K_m$ be the union of $q$ closed balls of radius $2\varepsilon_m$ around each $z_i$. Then, since $\inf_j \int_{T_i} |X_\alpha - a_{i,j}| d\mu| \leq \varepsilon_m + \inf_x \int_{T_i} |X_\alpha - x| d\mu|$, the following three conditions

$$\|X_\alpha\|_T \leq M, \quad \sup_i \sup_{s,t \in T_i} |X_\alpha(s) - X_\alpha(t)| \leq \varepsilon_m, \quad \sum_i \inf_x \int_{T_i} |X_\alpha - x| d\mu| \leq \varepsilon_m$$

imply that $X_\alpha \in K_m$. This holds for each $m$.

Let $K = \bigcap_{m=1}^{\infty} K_m$, which is closed, totally bounded, and therefore compact. Moreover, we argue that for every $\delta > 0$ there exists $m$ with $K^\delta \supset \bigcap_{j=1}^{m} K_j$. Suppose not. Then there is a sequence $z_m$ not in $K^\delta$, but with $z_m \in \bigcap_{j=1}^{m} K_j$ for every $m$. This has a subsequence contained in only one of the closed balls constituting $K_1$, and a further subsequence contained in only one of the balls constituting $K_2$, and so on. The diagonal sequence of such subsequences would eventually be contained in a ball of radius $2\varepsilon_m$ for every $m$. Therefore, it is Cauchy and its limit should be in $K$, which is a contradiction to the supposition $d(z_m, K) \geq \delta$ for every $m$.

Thus, we conclude that if $X_\alpha$ is not in $K^\delta$, then it is not in $\bigcap_{j=1}^{m} K_j$ for some $m$. Therefore,

$$P^*(X_\alpha \notin K^\delta) \leq P^* \left( X_\alpha \notin \bigcap_{j=1}^{m} K_j \right)$$

$$\leq P^*(\|X_\alpha\|_{\mathbb{L}_\mu} > M)$$

$$+ \sum_{j=1}^{m} P^* \left( \sup_i \sup_{s,t \in T_i} |X_\alpha(s) - X_\alpha(t)| \lor \sum_i \inf_x \int_{T_i} |X_\alpha - x| d\mu| > \varepsilon_j \right)$$

$$\leq \zeta + \sum_{j=1}^{m} \zeta 2^{-j} < 2\zeta.$$ 

Hence, we obtain $\limsup_{\alpha} P^*(X_\alpha \notin K^\delta) < 2\zeta$, as asserted.
(i) ⇒ (iii). If $X_\alpha$ is asymptotically tight, then so is each coordinate projection. Therefore, $X_\alpha(t)$ is asymptotically tight in $\mathbb{R}$ for every $t \in T$.

Let $K_1 \subset K_2 \subset \cdots$ be a sequence of compact sets such that $\liminf P_*(X_\alpha \in K_\varepsilon^c) \geq 1 - 1/m$ for every $\varepsilon > 0$. Define a semimetric $d$ on $T$ induced by $z$ by

$$d(s, t; z) := |z(s) - z(t)| \vee \int_T |z| \mathbb{I}\{z(s) \wedge z(t) \leq z \leq z(s) \vee z(t)\} \mathbb{I}\{z(s) \neq z(t)\} |d\mu|.$$  

Observe that $d(s, s; z) = 0$ and that $d$ is measurable with respect to $\mu$. Now for every $m$, define a semimetric $\rho_m$ on $T$ by

$$\rho_m(s, t) := \sup_{z \in K_m} d(s, t; z).$$  

We argue that $(T, \rho_m)$ is totally bounded. For $\eta > 0$, cover $K_m$ by finitely many balls of radius $\eta$ centered at $z_1, \ldots, z_k$. Consider the partition of $\mathbb{R}^{2k}$ into cubes of edge length $\eta$. For each cube, if there exists $t \in T$ such that the following $2k$-tuple is in the cube,

$$r(t) := \left( z_1(t), \int_T z_1 \mathbb{I}\{0 \wedge z_1(t) \leq z_1 \leq 0 \vee z_1(t)\} |d\mu|, \ldots, z_k(t), \int_T z_k \mathbb{I}\{0 \wedge z_k(t) \leq z_k \leq 0 \vee z_k(t)\} |d\mu| \right),$$

then pick one such $t$. Since $\|z_j\|_{L_m}$ is finite for every $j$ (i.e., the diameter of $T$ measured by each $d(\cdot, \cdot; z_j)$ is finite), this gives finitely many points $t_1, \ldots, t_p$. Notice that the balls $\{t : \rho_m(t, t_i) < 3\eta\}$ cover $T$, that is, $t$ is in the ball around $t_i$ for which $r(t)$ and $r(t_i)$ are in the same cube; this follows because $\rho_m(t, t_i)$ can be bounded by

$$2 \sup_{z \in K_m} \inf_j \|z - z_j\|_{L_m} + \sup_j d(t, t_i; z_j) < 3\eta.$$  

The first term is the error of approximating $z(t)$ and $z(t_i)$ by $z_j(t)$ and $z_j(t_i)$; the second is the distance of $t$ and $t_i$ measured by $d(\cdot, \cdot; z_j)$.

\textsuperscript{18} $T$ is not necessarily complete with respect to $d$.  

79
Define the semimetric \( \rho \) by

\[
\rho(s, t) := \sum_{m=1}^{\infty} 2^{-m} (\rho_m(s, t) \land 1).
\]

We show that \((T, \rho)\) is still totally bounded. For \(\eta > 0\) take \(m\) such that \(2^{-m} < \eta\). Since \(T\) is totally bounded in \(\rho_m\), we may cover \(T\) with finitely many \(\rho_m\)-balls of radius \(\eta\). Denote by \(t_1, \ldots, t_p\) the centers of such a cover. Since \(K_m\) is nested, we have \(\rho_1 \leq \rho_2 \leq \cdots\). Since we also have \(\rho_m(t, t_i) < \eta\) for every \(t\) there exists \(t_i\) such that \(\rho(t, t_i) \leq \sum_{k=1}^{m} 2^{-k} \rho_k(t, t_i) + 2^{-m} < 2\eta\). Therefore, \((T, \rho)\) is totally bounded.

By definition we have \(d(s, t; z) \leq \rho_m(s, t)\) for every \(z \in K_m\) and that \(\rho_m(s, t) \land 1 \leq 2^m \rho(s, t)\). And if \(\|z_0 - z\|_{\mu_x} < \varepsilon\) for \(z \in K_m\), then \(d(s, t; z_0) < 2\varepsilon + d(s, t; z)\) for every pair \((s, t)\). Hence, we conclude that

\[
K_m^\varepsilon \subseteq \left\{ z : \sup_{\rho(s, t) < 2^{-m}\varepsilon} d(s, t; z) \leq 3\varepsilon \right\}.
\]

Therefore, for \(\delta < 2^{-m}\varepsilon\),

\[
\liminf_{\alpha} P_{\alpha} \left( \sup_{\rho(s, t) < \delta} d(s, t; X_\alpha) \leq 3\varepsilon \right) \geq \liminf_{\alpha} P_{\alpha} \left( \sup_{t \in T} \left[ \sup_{\rho(s, t) < \delta} |X_\alpha(s) - X_\alpha(t)| \lor \int_{0 < \rho(s, t) < \delta} |X_\alpha(s)||d\mu| \right] \leq 3\varepsilon \right) \geq 1 - \frac{1}{m}.
\]

\((\text{iii}) \Rightarrow (\text{ii})\). For \(\varepsilon, \eta > 0\) and correspondingly taken \(\delta > 0\), one may construct the finite partition of \(T\), denoted by \(\{T_i^\varepsilon\}\), as follows. Since \(T\) is totally bounded, it can be covered with finitely many balls of radius \(\delta\); let \(t_1, \ldots, t_K\) be their centers. Disjointify the balls to obtain \(\{T_i^\varepsilon\}\). If \(\int_{\{t_i\}} |X_\alpha||d\mu| > 0\), then further separate the partition \(T_i^\varepsilon\) into \(\{t_i\}\) and \(T_i^\varepsilon \setminus \{t_i\}\) (then they both have the same center).

There are three types of components in the partition: (a) singleton components of mass points of \(\mu\), (b) components with \(|\mu|(T_i^\varepsilon) = \infty\), and (c) components with \(|\mu|(T_i^\varepsilon) < \infty\). The size of (a) is controlled by construction, so we are to control (b)
and (c) one by one. Clearly,

\[ \sup_{s, t \in T_\varepsilon} |X_\alpha(s) - X_\alpha(t)| \leq 2 \sup_{\rho(s, t) < \delta} |X_\alpha(s) - X_\alpha(t_i)| \leq 2 \varepsilon. \] (1.5)

Denote by \( i_\infty \) the index for which \( |\mu|(T_{i_\infty}^\varepsilon) = \infty \). We argue that \( \sum_{i_\infty} \int_{T_{i_\infty}^\varepsilon} |X_\alpha|d\mu \) can be arbitrarily small (with inner probability at least \( 1 - \eta \)) for sufficiently small \( \varepsilon \). By the construction of the partition, \( \sup_{s \in T_{i_\infty}^\varepsilon} |X_\alpha(s)| \leq 2 \varepsilon \)\(^{19}\) Thus, \( \sum_{i_\infty} \int_{T_{i_\infty}^\varepsilon} |X_\alpha|d\mu \leq \int_T |X_\alpha|1\{|X_\alpha| \leq 2\varepsilon\}d\mu \). Since \( T \) is totally bounded by the given semimetric, \( \int_T |X_\alpha|d\mu \) is bounded by \( K\varepsilon \) with inner probability at least \( 1 - \eta \) (proving asymptotic tightness of \( \|X_\alpha\|_\mu \)), and hence the previous integral must be arbitrarily small for small \( \varepsilon \). Now we turn to (c). Let \( \varepsilon' \) be such that

\[ \limsup_{\alpha} P^s \left( \int_T |X_\alpha|1\{|X_\alpha| \leq 3\varepsilon'\}d\mu > \varepsilon \right) < 1 - \eta. \] (1.6)

Take the partition for this \( \varepsilon' \), namely \( T_{i_\infty}^{\varepsilon'} \), to be nested on \( T_{i_\infty}^\varepsilon \) and pick up only the components \( \{T_{i_\infty}^{\varepsilon'}\} \) on \( |\mu|(T_{i_\infty}^\varepsilon) < \infty \). Note that \( \{T_{i_\infty}^\varepsilon\} \cup \{T_{i_\infty}^{\varepsilon'}\} \) defines another finite partition of \( T \). For ease of notation, denote \( T = T_{i_\infty} \sqcup T' \). If there exists \( s \in T_{i_\infty}^{\varepsilon'} \) such that \( |X_\alpha(s)| \leq \varepsilon' \), then by the construction of the partition \( \sup_{t \in T_{i_\infty}^{\varepsilon'}} |X_\alpha(t)| \leq 3\varepsilon' \). The contrapositive of this is also true. Thus, observing

\[
\sum_j \inf_x \int_{T_{i_\infty}^{\varepsilon'}} |X_\alpha - x|d\mu
\leq \sum_j \inf_x \int_{T_{i_\infty}^{\varepsilon'}} |X_\alpha - x|1\{|X_\alpha| > \varepsilon'\}d\mu + \int_T |X_\alpha|1\{|X_\alpha| \leq 3\varepsilon'\}d\mu,
\]

we may assume \( \inf_{T'} |X_\alpha(s)| \geq \varepsilon' > 0 \) at the cost of one more \( \varepsilon \). Then, we also have \( \int_{T'} |d\mu| \leq K\varepsilon/\varepsilon' \) since \( \varepsilon' \int_{T'} |d\mu| \leq \int_T |X_\alpha||d\mu| \). For the partition \( T_{i_\infty}^{\varepsilon'} \) of \( T' \), further

\(^{19}\text{This follows because } \inf_{T_{i_\infty}^{\varepsilon'}} |X_\alpha| = 0 \text{ given that } \int_{T_{i_\infty}^{\varepsilon'}} |X_\alpha||d\mu| < \infty.\)
construct a nested finite partition $T_k^{c'/K}$. Now

$$\sum_k \inf_x \int_{T_k^{c'/K}} |X_\alpha - x| |d\mu| \leq \sum_k \int_{T_k^{c'/K}} |d\mu| \sup_{s,t \in T_k^{c'/K}} |X_\alpha(s) - X_\alpha(t)| \leq \frac{c'}{K} \int_{T_k^{c'/K}} |d\mu| \leq \varepsilon$$  \hfill (1.7)

with inner probability at least $1 - \eta$. Combine (1.5), (1.6), and (1.7) to yield the result.

Finally, we prove the addendum. Define $K_m$ as before. Then, if $X_\alpha \rightsquigarrow X$, we have $P(X \in K_m) \geq 1 - 1/m$, and hence $X$ concentrates on $\bigcup_{m=1}^{\infty} K_m$. Since elements of $K_m$ are uniformly $\rho_m$-equicontinuous and $(\rho_m, \mu)$-equiintegrable, they are also uniformly $\rho$-equicontinuous and $(\rho, \mu)$-equiintegrable. This proves the first statement. Next, note that the set of uniformly continuous and integrable functions on a totally bounded, semimetric (denote by $d$) space is complete and separable in $L_\mu$. Thus the map $X$ that takes its values in this set is tight. If, moreover, $X_\alpha \rightsquigarrow X$, then $X_\alpha$ is asymptotically tight, so the compact sets for asymptotic tightness of $X_\alpha$ can be taken to be the compact sets for tightness of $X$. If every path of $X$ is uniformly $d$-continuous and $(d, \mu)$-integrable, then these compact sets can be chosen from the space of uniformly $d$-continuous and $(d, \mu)$-integrable functions. Since a compact set is totally bounded, every one of the compact sets is necessarily uniformly $d$-equicontinuous and $(d, \mu)$-equiintegrable. This completes the proof.

Without having to resort to the classical central limit theorem for the $L_1$ spaces, Proposition 1.D.1 can also be proved using Theorem 1.D.5. This (loosely) checks consistency of our theorem with known results.

Proof of Proposition 1.D.1 through Theorem 1.D.5: Being continuously differentiable, $m$ can be represented as a difference $m_1 - m_2$ of two strictly increasing and continuously differentiable functions $m_1$ and $m_2$ such that $m_1(X)$ and $m_2(X)$ has a $(2 + \epsilon)$th moment for $X \sim F$. In other words, we assume without loss of generality that $m$ is strictly increasing. Since asymptotic uniform equicontinuity is classical, in light
of Theorem 1.D.5 it remains only to show that the process \( X_n := \sqrt{n}(F_n - F) \) is asymptotically equiintegrable in probability. By Lemma 1.C.1, \( |y|^{1+c} \tilde{F} \circ m^{-1}(y) \) is integrable. This enables us to use the semimetric

\[
\rho(s, t) := \left( \int_{m^{-1}(s)}^{m^{-1}(t)} (|y|^{1+c} \vee 1) \tilde{F} \circ m^{-1}(y) dy \right)^{1/2}
\]

as it makes \( \mathbb{R} \) totally bounded. By the Cauchy-Schwarz inequality,

\[
\left( \int_s^t |X_n| d\mu \right)^2 \leq \left( \int_s^t (|m(x)|^{1+c} \vee 1) F(x) d\mu \right) \left( \int_{-\infty}^\infty \frac{1}{|m(x)|^{1+c} \vee 1} d\mu \right).
\]

By the change of variables,

\[
\int_{-\infty}^\infty \frac{1}{|m(x)|^{1+c} \vee 1} d\mu = \int_{-\infty}^\infty \frac{1}{|y|^{1+c} \vee 1} dy < \infty.
\]

With \( \mathbb{E}[X_n^2(x)] = F(x)[1 - F(x)] \), this implies that for some constant \( C \),

\[
\mathbb{E}\left[ \left( \int_s^t |X_n| d\mu \right)^2 \right] \leq C \int_s^t (|m(x)|^{1+c} \vee 1) F(x)[1 - F(x)] d\mu \leq C \rho(s, t)^2.
\]

Therefore, by Van der Vaart and Wellner (1996 Theorem 2.2.4), for any \( \eta, \delta > 0 \),

\[
\mathbb{E}\left[ \left( \sup_{\rho(s, t) \leq \delta} \int_s^t |X_n| d\mu \right)^2 \right] \leq K \left[ \int_0^\eta \sqrt{D(\varepsilon, \rho)} d\varepsilon + \delta D(\eta, \rho) \right]^2
\]

for some constant \( K \) where \( D(\varepsilon, \rho) \) is the packing number of \( T \) with respect to \( \rho \).

With this choice of \( \rho \), the packing number satisfies \( D(\varepsilon, \rho) \approx 1/\varepsilon \). Thus, we obtain

\[
\mathbb{E}\left[ \left( \sup_{\rho(s, t) \leq \delta} \int_s^t |X_n| d\mu \right)^2 \right] \leq \tilde{K} \left( 2\sqrt{\eta} + \frac{\delta}{\eta} \right)^2
\]

83
for some $\tilde{K}$. With Markov's inequality,

$$P\left(\sup_{t \in \mathbb{R}} \int_{0 < \rho(s,t) \leq \delta} |X_n|d\mu(s) > \varepsilon\right) \leq \frac{1}{\varepsilon^2} \mathbb{E}\left[ \left( \sup_{\rho(s,t) \leq \delta} \int_s^t |X_n|d\mu\right)^2 \right] \leq \frac{\tilde{K}}{\varepsilon^2} \left(2\sqrt{\eta} + \frac{\delta}{\eta}\right)^2.$$

This can be however small for any $\varepsilon$ by the choice of $\eta$ and $\delta$.

\section{Differentiability of the Inverse Map}

We use the space $\mathbb{L}$ constructed in the previous section to show differentiability of the inverse map for distribution functions. Let $m$ be a strictly increasing continuous function and $\mu$ the Lebesgue-Stieltjes measure associated with $m$. For a real-valued random variable $X$ distributed as $F$, Lemma 1.C.1 implies that $m(X)$ has a first moment if and only if $\tilde{F}$ belongs to $\mathbb{L}_\mu$ with $T$ equal to $\mathbb{R}$ and $\mathcal{T}$ appropriately chosen. Specifically, the space we work on is as follows.

**Definition.** Let $\mathbb{L}_\mu$ be the space of $\mu$-measurable functions $z$ from $\mathbb{R}$ to $\mathbb{R}$ with limits $z(\pm \infty) := \lim_{x \to \pm \infty} z(x)$ and the norm

$$\|z\|_{\mathbb{L}_\mu} := \|z\|_\infty \lor \|z\|_\mu := \left(\sup_{x \in \mathbb{R}} |z(x)|\right) \lor \left(\int_{-\infty}^{\infty} |\tilde{z}(x)|d\mu\right),$$

where

$$\tilde{z}(x) = \begin{cases} z(x) - z(-\infty) & x < 0, \\ z(x) - z(+\infty) & x \geq 0. \end{cases}$$

Denote by $\mathbb{L}_{\mu,\phi}$ the subset of $\mathbb{L}_\mu$ of monotone cadlag functions with $z(-\infty) = 0$ and $z(+\infty) = 1$. If $\mu$ is the Lebesgue measure, we omit the subscript $\mu$ and denote them by $\mathbb{L}$ and $\mathbb{L}_{\phi}$, and $\|\cdot\|_{\mu}$ by $\|\cdot\|_1$.

Next, we define the space of quantile functions for distributions that have first moments. Note that there is no supremum component in its norm.

**Definition.** Let $\mathbb{B}$ be the space of cadlag functions $z$ from $(0,1)$ to $\mathbb{R}$ with the norm

$$\|z\|_\mathbb{B} := \int_0^1 |z(u)|du.$$
Remark. The metric on quantile functions induced by \( \| \cdot \|_B \) is known as the Wasserstein metric in probability theory and the Mallows distance in statistics.

The next lemma establishes differentiability of the inverse map for distribution functions with finite first moments (or more generally, monotone functions \( F \) whose modifications \( \tilde{F} \) are integrable).

**Lemma 1.E.1 (Inverse map).** Let \( F \in L_\phi \) be a distribution function on (an interval of) \( \mathbb{R} \) that has at most finitely many jumps and is otherwise continuously differentiable with strictly positive density \( f \). Then, the inverse map \( \phi : L_\phi \to \mathbb{B} \), \( \phi(F) := Q = F^{-1} \), is Hadamard differentiable at \( F \) tangentially to the set \( L_0 \) of all continuous functions in \( L \). The derivative is given by

\[
\phi'_F(z) = -(z \circ Q)Q' = \begin{cases} 
-(z/f) \circ Q & \text{if } Q \text{ is increasing}, \\
0 & \text{if } Q \text{ is flat}.
\end{cases}
\]

**Proof.** Take \( z_t \to z \) in \( L \) and \( F_t := F + tz_t \in L_\phi \). We want to show that

\[
\left\| \frac{\phi(F_t) - \phi(F)}{t} - \phi'_F(z) \right\|_B = \int_0^1 \left| \frac{\phi(F_t) - \phi(F)}{t} - \phi'_F(z) \right| du \to 0 \text{ as } t \to 0.
\]

Let \( j \in \mathbb{R} \) be a point of jump of \( F \). For small \( \varepsilon > 0 \), one can separate the integral as

\[
\left( \int_0^{F(j-\varepsilon)} + \int_{F(j-\varepsilon)}^{F(j+\varepsilon)} + \int_{F(j+\varepsilon)}^1 \right) \left| \frac{\phi(F_t) - \phi(F)}{t} - \phi'_F(z) \right| du.
\]

Observe that

\[
\int_{F(j-\varepsilon)}^{F(j+\varepsilon)} \left| \frac{\phi(F_t) - \phi(F)}{t} - \phi'_F(z) \right| du \leq 2\varepsilon \left\| \frac{F_t - F}{t} \right\|_\infty + \left( \int_{F(j-\varepsilon)}^{F(j+\varepsilon)} \phi'_F(z) \right) |du|.
\]

The first term equals \( 2\varepsilon \| z_t \|_\infty \) and can be arbitrarily small by the choice of \( \varepsilon \). If \( \varepsilon \) is small enough that there is no other jump in \([j - \varepsilon, j + \varepsilon]\), by Fubini’s theorem,

\[
\int_{F(j-\varepsilon)}^{F(j-\varepsilon)} |\phi'_F(z)| du = \int_{j-\varepsilon}^{j} \left| \frac{z}{f} \right| dF \leq \varepsilon \| z \|_\infty.
\]
which can also be arbitrarily small by the choice of \( \varepsilon \). Similarly, the last integral can as well be arbitrarily small. Therefore, one can ignore any finitely many jumps of \( F \); we assume hereafter that \( F \) has no jump and has positive density \( f \) everywhere.

For every \( \varepsilon > 0 \) there exists a large number \( M \) such that \( F(-M) < \varepsilon \) and \( 1 - F(M) < \varepsilon \). Write

\[
\left\| \frac{\phi(F_t) - \phi(F)}{t} - \phi_F'(z) \right\|_B \leq \int_{F(-M)+\varepsilon}^{F(M)-\varepsilon} \left| \frac{\phi(F_t) - \phi(F)}{t} - \phi_F'(z) \right| du + \left( \int_0^{2\varepsilon} + \int_{1-2\varepsilon}^1 \right) \left| \frac{\phi(F_t) - \phi(F)}{t} - \phi_F'(z) \right| du.
\]

By Van der Vaart and Wellner (1996 Theorem 3.9.23 (i)), uniform convergence of the integrand holds on \([F(-M)+\varepsilon, F(M)-\varepsilon]\). Since the first integral is bounded by

\[
\sup_{u \in [F(-M)+\varepsilon, F(M)-\varepsilon]} \left| \frac{\phi(F_t) - \phi(F)}{t} - \phi_F'(z) \right|,
\]

it vanishes as \( t \to 0 \).

We now turn to the second integral. The triangle inequality bounds the integral by

\[
\int_0^{2\varepsilon} \left| \frac{\phi(F_t) - \phi(F)}{t} \right| du + \int_0^{2\varepsilon} |\phi_F'(z)| du.
\]

Since \( F \) and \( F_t \) are nondecreasing, by Fubini’s theorem,

\[
\int_0^{2\varepsilon} \left| \frac{\phi(F_t) - \phi(F)}{t} \right| du = \frac{1}{|t|} \int_0^{2\varepsilon} |F_t^{-1} - F^{-1}| du \leq \frac{1}{|t|} \int_{-\infty}^{F^{-1}(2\varepsilon + \|tz\|_\infty)} |tz_t| dx \\
\leq \|z_t - z\|_1 + \int_{-\infty}^{F^{-1}(2\varepsilon + \|tz\|_\infty)} |z| dx.
\]

The first term goes to 0 and the second term can be arbitrarily small by the choice of \( \varepsilon \). Finally, by the change of variables,

\[
\int_0^{2\varepsilon} |\phi_F'(z)| du = \int_{-\infty}^{F^{-1}(2\varepsilon)} z |F^{-1}(z)| dF = \int_{-\infty}^{F^{-1}(2\varepsilon)} |z| dx.
\]

This quantity can be arbitrarily small. Similarly, the integral from \( 1 - 2\varepsilon \) to \( 1 \) can be...
shown to converge to 0. This completes the proof. ■

Remark. A probability distribution $F$ has a $p$th moment if and only if its quantile function $Q$ is in $L_p$ (Lemma [L.C.1]). This may spur speculations that if $F$ has a $p$th moment, then the map $F \mapsto Q$ may be differentiable for $Q \in L_p$. However, we have not been able to prove that this is the case (although the Glivenko-Cantelli type results do hold; see Addendum [1.H.2]). The success of Lemma [1.E.1] hinges upon the fact that Fubini’s theorem is compatible with the $L_1$ norm. Nevertheless, it is possible to extend differentiability of inverse maps to subsume $p$th moments, or even to more general transformations, by regarding the range space to be $L_1$ as done below.

Now we extend the result to subsume transformations. Consider, for example, the second moment. Observe that the second moment of $X$ can be thought of as the first moment of $Z := X^2$. In other words,

$$\int_{-\infty}^{\infty} x^2 dF(x) = \int_{0}^{\infty} zdF(\sqrt{z}) + \int_{-\infty}^{0} -zdF(-\sqrt{-z}) = \int_{0}^{\infty} zd(F(\sqrt{z}) + F(-\sqrt{z})).$$

Here, $F(\sqrt{\cdot}) + F(-\sqrt{\cdot})$ is the distribution function of the random variable $Z$.\footnote{After a minor fix for right-continuity.} This is in line with the informal mental exercise that if one inverts the $p$th power of an inverse function, $(F^{-1})^p$, one obtains the composition of the original function and the $1/p$th power, $F \circ (\cdot)^{1/p}$. Thus, one expects that this composition, $F \circ (\cdot)^{1/p}$, is in $L$ whenever $F$ is in $L_{|x|^p}$. Then, Hadamard differentiability of the map $F \mapsto (F^{-1})^p$ may follow by the chain rule on $F \mapsto F \circ (\cdot)^{1/p} \mapsto [F \circ (\cdot)^{1/p}]^{-1} = (F^{-1})^p$. For this, one only needs Hadamard differentiability (or anything stronger) of the first map, $F \mapsto F \circ (\cdot)^{1/p}$.

Remark. As remarked above, a subtle but important distinction is that the last element in this chain $(F^{-1})^p$ should be seen as itself belonging to $B$ (the $L_1$ space), but not as $F^{-1}$ belonging to the $L_p$ space.

More generally, for a monotone function $m$, we exploit the relationship $m(F^{-1}) = (F \circ m^{-1})^{-1}$. The requirement of monotonicity of $m$ is almost innocuous since any
function of locally bounded variation can be represented by a difference of two increasing functions, so the lemma extends naturally to more general transformation.

**Lemma 1.E.2.** Let \( m : \mathbb{R} \to \mathbb{R} \) be a strictly increasing continuous function and \( \mu \) be the associated Lebesgue-Stieltjes measure. Then, the map \( \psi : \mathbb{L}_\mu \to \mathbb{L} \), \( \psi(F) := F \circ m^{-1} \), is uniformly Fréchet differentiable with rate function \( q \equiv 0 \). The derivative is given by \( \psi'_F(z) := z \circ m^{-1} \).

**Proof.** Observe that \( \psi(F+z) - \psi(F) = (F+z)(m^{-1}) - F(m^{-1}) = z(m^{-1}) \). Therefore, \( \psi(F+z) - \psi(F) - \phi'_F(z) = 0 \). ■

This lemma is obvious since the map \( F \mapsto F \circ m^{-1} \) is by itself linear. Now differentiability of the inverse map for general transformations follows by the chain rule.

**Theorem 1.E.3** (Transformed inverse map). Let \( m : \mathbb{R} \to \mathbb{R} \) be a continuously differentiable function and \( \mu \) be the associated Lebesgue-Stieltjes measure. Let \( F \in \mathbb{L}_{\mu,\phi} \) be a distribution function on (an interval of) \( \mathbb{R} \) that has at most finitely many jumps and is otherwise continuously differentiable with strictly positive density \( f \). Then, the map \( \phi \circ \psi : \mathbb{L}_{\mu,\phi} \to \mathbb{B} \), \( \phi \circ \psi(F) := m(Q) \), is Hadamard differentiable at \( F \) tangentially to the set \( \mathbb{L}_{\mu,0} \) of all continuous functions in \( \mathbb{L}_\mu \). The derivative is given by

\[
(\phi \circ \psi)'_F(z) := -(m'z/f) \circ Q.
\]

**Proof.** Since \( m \) is continuously differentiable and hence of locally bounded variation, one can set \( m(x) = m_1(x) - m_2(x) \), where \( m_1 \) and \( m_2 \) are both increasing and continuously differentiable functions. Moreover, \( m_1 \) and \( m_2 \) can be taken such that they are strictly increasing, and for their corresponding Lebesgue-Stieltjes measures \( \mu_1 \) and \( \mu_2 \), \( F \) belongs to both \( \mathbb{L}_{\mu_1,\phi} \) and \( \mathbb{L}_{\mu_2,\phi} \). Since the derivative formula is linear in \( m' \), it suffices to show that the claim holds for each of \( m_1 \) and \( m_2 \) separately. In other words, we can assume without loss of generality that \( m \) is strictly increasing.

---

\[21\] A map \( \psi : \mathbb{L} \to \mathbb{B} \) is uniformly Fréchet differentiable with rate function \( q \) if there exists a continuous linear map \( \psi'_F : \mathbb{L} \to \mathbb{B} \) such that \( \|\psi(F+z) - \psi(F) - \psi'_F(z)\|_B = O(q(\|z\|_L)) \) uniformly over \( F \in \mathbb{L} \) as \( z \to 0 \) and \( q \) is monotone with \( q(t) = o(t) \).

\[22\] For example, take \( m_1 \) and \( m_2 \) to be the least steep functions and add a normal cdf.
Now observe that $z$ is in $\mathbb{L}_\mu$ (or $\mathbb{L}_{\mu,0}$) if and only if $z \circ m^{-1}$ is in $\mathbb{L}$ (or $\mathbb{L}_0$). The assertion then follows by the chain rule (Van der Vaart and Wellner [1996] Lemma 3.9.3) applied to Lemmas 1.E.1 and 1.E.2.

The main conclusion of this section is summarized as follows.

**Proposition 1.E.4.** Let $m : \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function. For a distribution function $F$ on (an interval of) $\mathbb{R}$ that has at most finitely many jumps and is otherwise continuously differentiable with strictly positive density $f$ such that $m(X)$ has a $(2 + c)$th moment for $X \sim F$ and some $c > 0$, the process $\sqrt{n}(m(Q_n) - m(Q))$ converges weakly in $\mathbb{B}$ to a Gaussian process with mean zero and covariance function

$$\text{Cov}(s, t) = m'(Q(s))Q'(s)m'(Q(t))Q'(t)(s \wedge t - st).$$

**Proof.** This follows in combination of Proposition 1.D.1 and Theorem 1.E.3.

---

### 1.F Differentiability of $L$-Statistics

Given the weak convergence of the integral of empirical quantiles, we investigate whether this is the case for the integral with respect to a random measure. Or precisely, we seek conditions under which the integral of a stochastic process with respect to another stochastic process converges weakly. This is the extension of the results on Wilcoxon statistics (Van der Vaart and Wellner [1996] Section 3.9.4.1) to allow for an unbounded integrand. Then, the general $L$-statistics result follows with Proposition 1.E.4.

The spaces we work on in this section are given as follows.

**Definition.** Let $Q \in \mathbb{B}$ and define the space $\mathbb{L}_Q$ of functions $\kappa : (0, 1) \to \mathbb{R}$ with the norm

$$\|\kappa\|_{\mathbb{L}_Q} := \|\kappa\|_{\mathbb{L}_Q, \infty} \vee \|\kappa\|_{Q} := \left(\sup_{u \in (0, 1)} |(Q \vee 1)(u)\kappa(u)|\right) \vee \left(\int_{0}^{1} |\kappa(u)||dQ(u)|\right).$$

Let $\mathbb{L}_{Q, M}$ be the subset of $\mathbb{L}_Q$ of Lipschitz functions.

First, we give a generalization of differentiability of Wilcoxon statistics.
**Theorem 1.F.1 (Wilcoxon statistic).** For each fixed $M$, the maps $\lambda : \mathbb{B} \times \mathbb{L}_{Q,M} \to \mathbb{R}$ and $\tilde{\lambda} : \mathbb{B} \times \mathbb{L}_{Q,M} \to L_{\infty}(0,1)^2$,

$$
\lambda(Q,K) := \int_0^1 QdK \quad \text{and} \quad \tilde{\lambda}(Q,K) := \int_s^t QdK,
$$

are Hadamard differentiable at every $(Q,K) \in \mathbb{B} \times \mathbb{L}_{Q,M}$ uniformly over $\mathbb{L}_{Q,M}$. The derivative maps are given by

$$
\lambda_{Q,K}^t(z,\kappa) := \int_0^1 Qd\kappa + \int_0^1 zdK,
$$

where $\int Qd\kappa$ is defined via integration by parts if $\kappa$ is of unbounded variation.

**Proof.** The derivative map is linear by construction; it is also continuous since

$$
|\lambda_{Q,K}^t(z_1,\kappa_1) - \lambda_{Q,K}^t(z_2,\kappa_2)| = \left| \int_0^1 Q(d(\kappa_1 - \kappa_2) + \int_0^1 (z_1 - z_2)dK) \right|
\leq \|\kappa_1 - \kappa_2\|_{Q,\infty} + \|\kappa_1 - \kappa_2\|_Q + M\|z_1 - z_2\|_B,
$$

which vanishes as $\|z_1 - z_2\|_B \to 0$ and $\|\kappa_1 - \kappa_2\|_{\mathbb{L}_Q} \to 0$. Let $z_t \to z$ and $\kappa_t \to \kappa$ such that $Q_t := Q + tz_t$ is in $\mathbb{B}$ and $K_t := K + t\kappa_t$ is in $\mathbb{L}_{Q,M}$. Observe

$$
\frac{\lambda(Q_t, K_t) - \lambda(Q, K)}{t} - \lambda_{Q,K}^t(z_t,\kappa_t) = \int (z_t - z)d(K_t - K) + \int zd(K_t - K).
$$

We want to show that this converges to zero as $t \to 0$. The first term vanishes since

$$
\left| \int (z_t - z)d(K_t - K) \right| \leq 2M \int |z_t - z|du = 2M\|z_t - z\|_B.
$$

Since $z$ is integrable, for every $\varepsilon > 0$ there exists a small number $\delta > 0$ such that

$$
\left( \int_0^\delta + \int_{1-\delta}^1 \right) |z|du + \int_\delta^{1-\delta} (|z| - (|z| \wedge \delta^{-1}))du \leq \varepsilon.
$$

90
This gives the inequality

\[
\left| \int zd(K_t - K) \right| \leq \left| \int_{\delta}^{1-\delta} (-\delta^{-1} \vee z \wedge \delta^{-1})d(K_t - K) \right| \\
+ \left| \int zd(K_t - K) - \int_{\delta}^{1-\delta} (-\delta^{-1} \vee z \wedge \delta^{-1})d(K_t - K) \right| \\
\leq \left| \int_{\delta}^{1-\delta} (-\delta^{-1} \vee z \wedge \delta^{-1})d(K_t - K) \right| + 2M\varepsilon.
\]

Let \( \tilde{z} := -\delta^{-1} \vee z \wedge \delta^{-1} \). Since \( \tilde{z} \) is ladcag on \([\delta, 1 - \delta]\), there exists a partition \( \delta = t_0 < t_1 < \cdots < t_m = 1 - \delta \) such that \( \tilde{z} \) varies less than \( \varepsilon \) on each interval \((t_{i-1}, t_i]\).

Let \( \bar{z} \) be the piecewise constant function that equals \( \tilde{z}(t_i) \) on each interval \((t_{i-1}, t_i]\).

Then

\[
\left| \int_{\delta}^{1-\delta} \tilde{z}d(K_t - K) \right| \leq 2M \sup_{u \in [\delta, 1-\delta]} |\tilde{z} - \bar{z}| + |\bar{z}(\delta)||K_t - K|(\{\delta\}) | \\
+ \sum_{i=1}^{m} |\bar{z}(t_i)||K_t - K|(t_{i-1}, t_i]).
\]

The first term is arbitrarily small by the choice of \( \varepsilon \), and the second and third terms are collectively bounded by \((2m + 1)\delta^{-1}\|K_t - K\|_\infty = (2m + 1)\delta^{-1}t\|\kappa_t\|_\infty\), which converges to 0 regardless of the choice of \( K \).

The proof for \( \tilde{\lambda} \) is basically the same as that for \( \lambda \).

Next, we give conditions under which the random measure \( K_n \) converges in \( L_Q \).

Not surprisingly, this convergence hinges on the tail behavior around 0 and 1. Roughly speaking, if \( Q \) has a \((2 + c)\)th moment, then weak convergence of \( \frac{X_\alpha}{u^r(1-u)^r} \) to \( \frac{X}{u^r(1-u)^r} \) in \( L_\infty \) for some \( r > \frac{1}{2+c} \) implies weak convergence of \( X_\alpha \) to \( X \) in \( L_Q \).

**Lemma 1.F.2.** Let \( Q : (0, 1) \to \mathbb{R} \) be a quantile function whose probability measure has a \((2 + c)\)th moment for some \( c > 0 \). If for a net of processes \( X_\alpha : \Omega_\alpha \to L_Q \) there exists \( r > \frac{1}{2+c} \) such that for every \( \eta > 0 \) there exists \( M \) satisfying

\[
\limsup_{\alpha} P^\alpha \left( \left\| \frac{X_\alpha}{u^r(1-u)^r} \right\|_\infty > M \right) < \eta,
\]

91
then there exists a semimetric $\rho$ on $(0, 1)$ such that $(0, 1)$ is totally bounded, $QX_\alpha$ is asymptotically uniformly $\rho$-equicontinuous in probability, and $X_\alpha$ is asymptotically $(\rho, Q)$-equiintegrable in probability.

**Proof.** Assume $r < 1$ first. Define $\rho$ by

$$\rho(s, t) := \int_{(s, t)} u^r(1 - u)^r dQ.$$

We show that $(0, 1)$ is totally bounded with respect to $\rho$. Observe that Lemma 1.C.1 and $r > \frac{1}{2+c}$ imply $u^r(1 - u)^r Q(u) \to 0$ as $u \to \{0, 1\}$. Therefore, integrating by parts,

$$\rho(0, 1) \leq \int_{(0, 1)} u^r \wedge (1 - u)^r dQ \leq |Q| \left(\frac{1}{2}\right) + \int_0^{\frac{1}{2}} u^{r-1} |Q| du + \int_{\frac{1}{2}}^1 (1 - u)^{r-1} |Q| du.$$

Since $Q \in L_{2+c}$ and $u^{r-1} \wedge (1 - u)^{r-1} \in L_q$ for every $q < 1/(1 - r)$, in particular for $q = (2 + c)/(1 + c)$, this integral is finite by Hölder’s inequality. This means that the diameter of $(0, 1)$ is finite, concluding that $(0, 1)$ is totally bounded.

Note that $|Q|$ is eventually smaller than $1/w^r(1 - u)^r$ near 0 and 1, so that for every $\eta$ there exists $M$ such that

$$\limsup_{\alpha} P^*\left(\|(|Q| \vee 1)X_\alpha\|_\infty > M\right) \leq \limsup_{\alpha} P^*\left(\left\|\frac{X_\alpha}{u^r(1 - u)^r}\right\|_\infty > M\right) < \eta.$$

This shows uniform equicontinuity.

Next, for every $0 < s \leq t < 1$,

$$\int_{(s, t)} |X_\alpha| dQ \leq \left(\sup_{u \in (0, 1)} \frac{|X_\alpha(u)|}{u^r(1 - u)^r}\right) \int_{(s, t)} u^r(1 - u)^r dQ \leq \left\|\frac{X_\alpha}{u^r(1 - u)^r}\right\|_\infty \rho(s, t).$$

Therefore,

$$P^*\left(\sup_{t \in (0, 1)} \int_{0 < \rho(s, t) < \delta} |X_\alpha| dQ(s) > \varepsilon\right) \leq P^*\left(\left\|\frac{X_\alpha}{u^r(1 - u)^r}\right\|_\infty > \frac{\varepsilon}{\delta}\right).$$

---

23 This semimetric is reminiscent of the condition for $\|\cdot /q\|_p$-metric convergence in Csörgő et al. (1993).
By assumption, this can be however small by the choice of \( \delta \). Conclude that \( X_\alpha \) is asymptotically \((\rho, Q)\)-equiintegrable in probability.

Finally, if \( r \geq 1 \), replace every \( r \) appeared in the proof by \( 1/2 \). Then the result follows since
\[
\| \frac{X_\alpha}{u(1-r)^r} \|_\infty \geq \| \frac{X_\alpha}{u^{1/2}(1-u)^{1/2}} \|_\infty.
\]

Next, we apply this lemma to show that most “well-behaved” sample selection measures satisfy the condition. Let \( X_1, \ldots, X_n \) be independent continuous random variables and \( X_{1,n}, \ldots, X_{m,n} \) be subset of \( X_1, \ldots, X_n \) that are selected by some (possibly random) criterion. Then, roughly speaking, if the empirical distribution of the selected sample \( X_{1,n}, \ldots, X_{m,n} \) converges in \( L_{\infty} \) to a smooth distribution, then the selection measure \( K_n \) defined in the text converges in \( \mathbb{L}_Q \).

**Proposition 1.F.3.** Let \( U_1, \ldots, U_n \) be independent uniformly distributed random variables on \((0, 1)\) and \( w_{1,n}, \ldots, w_{n,n} \) random variables bounded by some constant \( M \) whose distribution can depend on \( U_1, \ldots, U_n \) and \( n \). Define
\[
\mathbb{F}_n(u) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{U_i \leq u\}, \quad \mathbb{G}_n(u) := \frac{1}{n} \sum_{i=1}^{n} w_{i,n} \mathbb{1}\{U_i \leq u\}.
\]

Let \( I(u) := u \) and assume that \( K(u) := \lim_{n \to \infty} \mathbb{E}[\mathbb{G}_n(u)] \) exists and is Lipschitz and differentiable. If \( \sqrt{n}(\mathbb{G}_n - K) \) weakly converges jointly with \( \sqrt{n}(\mathbb{F}_n - I) \) in \( L_{\infty} \), then for the selection measure
\[
K_n(u) := \frac{1}{n} \sum_{i=1}^{n} w_{i,n} \mathbb{1}\{0 \vee (nu - n\mathbb{F}_n(U_i) + 1) \wedge 1\},
\]
we have \( \sqrt{n}(K_n - K) \) converge weakly in \( \mathbb{L}_Q \) for every quantile function \( Q \) whose distribution has a \((2 + c)\)th moment for some \( c > 0 \).

**Proof.** Assume without loss of generality that \( M = 1 \). Define \( U_{(0)} := 0 \). Let \( \tilde{\mathbb{F}}_n \) and
Lemma 1.F.4 (Inverse composition map). Let $\mathbb{L}_Q$ contain the identity map $I(u) := u$. Let $\mathcal{D}$ be the subset of $\mathbb{L}_Q \times \mathbb{L}_Q$ such that every $(A, B) \in \mathcal{D}$ satisfies $A(u_1) - A(u_2) \geq B(u_1) - B(u_2) \geq 0$ for every $u_1 \geq u_2$, the range of $A$ contains $(0, 1)$, and $B$ is differentiable and Lipschitz. Let $\mathbb{L}_{Q, UC}$ be the subset of $\mathbb{L}_Q$ of uniformly continuous functions. Then, the map $\chi : \mathcal{D} \to \mathbb{L}_Q$, $\chi(A, B) := B \circ A^{-1}$, is Hadamard differentiable at $(A, B) \in \mathcal{D}$ for $A = I$ tangentially to $\mathbb{L}_Q \times \mathbb{L}_{Q, UC}$. The derivative is
given by
\[ x'_{I,B}(a, b)(u) = b(u) + B'(u)a(u), \quad u \in (0, 1). \]

Proof. For \((A, B) \in \mathcal{D}\) and \(u_1 \geq u_2\), denote \(v_1 := A(u_1)\) and \(v_2 := A(u_2)\). By assumption we have \(v_1 - v_2 \geq B(A^{-1}(v_1)) - B(A^{-1}(v_2)) \geq 0\) for every \(v_1 \geq v_2\). Therefore, \(B \circ A^{-1}\) is monotone and bounded by the identity map up to a constant. This implies
\[ \int_{(0,1)} |\tilde{B} \circ A^{-1}||dQ| \leq \int_{(0,1)} |\tilde{I}||dQ| < \infty \]
and \(\|Q(B \circ A^{-1})\|_\infty < \infty\); it follows that \(B \circ A^{-1}\) is in \(\mathbb{L}_{Q,1}\).

Let \(a_t \to a\) and \(b_t \to b\) in \(\mathbb{L}_{\nu_2}\) and \((A_t, B_t) := (I + ta_t, B + tb_t) \in \mathcal{D}\). We want to show that
\[ \left\| \frac{B_t \circ A_t^{-1} - B \circ I^{-1}}{t} - b - B'a \right\|_{L_Q} \to 0 \quad \text{as} \quad t \to 0. \]

That \(\| \cdot \|_{Q,\infty} \to 0\) follows by applying Van der Vaart and Wellner (1996, Lemma 3.9.27) to \((A^{-1}, QB)\) as elements in \(L_\infty\). Thus, it remains to show that \(\| \cdot \|_Q \to 0\).

In the assumed inequality, substitute \((u_1, u_2)\) by \((u, A_t^{-1}(u))\) to find that
\[ |A_t(u) - u| \geq |B_t(A_t^{-1}(u)) - B_t(u)| \geq 0. \]

Therefore, the following inequality holds pointwise:
\[ |B_t \circ A_t^{-1} - B| \leq |B_t \circ A_t^{-1} - B_t| + |B_t - B| \leq |A_t - I| + |B_t - B| = |ta_t| + |tb_t|. \]

For \(\varepsilon > 0\), write \(\| \cdot \|_Q\) as
\[ \left( \int_{0}^{\varepsilon} + \int_{\varepsilon}^{1-\varepsilon} + \int_{1-\varepsilon}^{1} \right) \left| \frac{B_t \circ A_t^{-1} - B}{t} - b - B'a \right||dQ|. \]

For any fixed \(\varepsilon > 0\) the middle term vanishes as \(t \to 0\) since \(\| \cdot \|_{Q,\infty} \to 0\). It remains to show that the first term can be however small by the choice of \(\varepsilon\) since then by symmetry the third term must likewise be ignorable. Using the inequality obtained
above, write

\[
\int_0^\varepsilon B_t \circ A_t^{-1} - B - B' a \, dQ \leq \int_0^\varepsilon \left( |a_t| + |b_t| + |b| + |B'a| \right) |dQ|.
\]

Since \(|a_t - a|_Q \to 0\) and \(|b_t - b|_Q \to 0\), this integral should be arbitrarily small by the choice of \(\varepsilon\), as desired. ■

Now we are ready to give the main conclusion of this chapter.

**Proposition 1.F.5 (L-statistic).** Let \(m_1, m_2 : \mathbb{R} \to \mathbb{R}\) be continuously differentiable functions and \(F : \mathbb{R}^2 \to [0, 1]\) be a distribution function on (a rectangular of) \(\mathbb{R}^2\) with marginal distributions \((F_1, F_2)\) that have at most finitely many jumps and are otherwise continuously differentiable with strictly positive marginal densities \((f_1, f_2)\) such that \(m_1(X_1)\) and \(m_2(X_2)\), \((X_1, X_2) \sim F\), have \((2 + c)\)th moments for some \(c > 0\). Along with i.i.d. random variables \(X_{1,1}, \ldots, X_{n,1}\) and \(X_{1,2}, \ldots, X_{n,2}\), let \(w_{1,1,1}, \ldots, w_{n,1,1}\) and \(w_{1,1,2}, \ldots, w_{n,1,2}\) be random variables bounded by a constant \(M\) whose distributions can depend on \(n\) and all of \(X_{1,1}, \ldots, X_{n,1}\) and \(X_{1,2}, \ldots, X_{n,2}\) such that the empirical distributions of \(X_{i,1}, X_{i,2}, w_{i,1,1}, X_{i,1}\), and \(w_{i,1,2}, X_{i,2}\) converge uniformly jointly to continuously differentiable distribution functions. Then, the normalized \(L\)-statistics

\[
\sqrt{n} \left( \frac{\mathbb{E}_n[m_1(X_{i,1})w_{i,1,1}] - \mathbb{E}[m_1(X_{i,1})w_{i,1,1}]}{\mathbb{E}_n[m_2(X_{i,2})w_{i,2,2}] - \mathbb{E}[m_2(X_{i,2})w_{i,2,2}]} \right)
\]

\[
= \sqrt{n} \left( \frac{\int_0^1 m_1(Q_{n,1})dK_{n,1} - \int_0^1 m_1(Q_1)dK_1}{\int_0^1 m_2(Q_{n,2})dK_{n,2} - \int_0^1 m_2(Q_2)dK_2} \right)
\]

where

\[
K_j(u) := \lim_{n \to \infty} \mathbb{E}[w_{i,n,j} \mid F_j(X_{i,j}) \leq u],
\]

\[
K_{n,j}(u) := \frac{1}{n} \sum_{i=1}^n w_{i,n,j} \mathbb{1} \left\{ 0 \vee (nu - n\mathbb{E}_{n,j}(X_i) + 1) \right\} \wedge 1,
\]

96
converge weakly in $\mathbb{R}^2$ to a normal vector $(\xi_1, \xi_2)$ with mean zero and (co)variance

$$\text{Cov}(\xi_j, \xi_k) = \int_0^1 \int_0^1 \left( m_j'(Q_j(s))Q_j'(s)m_k'(Q_k(t))Q_k'(t) \times \right)$$

$$\left( [F_{jk}^Q(s, t) - st] + [K_{jk}(s, t)F_{jk}^Q(s, t) - stK_j(s)K_k(t)] \right)$$

$$- K_j(s)[F_{jk}^Q(s, t) - st] - K_k(t)[F_{jk}^Q(s, t) - st] \right) dsdt,$$

where

$$F_{jk}^Q(s, t) := \Pr(X_{i,j} \leq Q_j(s), X_{i,k} \leq Q_k(t)),$$

$$K_{jk}(s, t) := \lim_{n \to \infty} \mathbb{E}[w_{i,n,j}w_{i,n,k} | X_{i,j} \leq Q_j(s), X_{i,k} \leq Q_k(t)].$$

If $F$ has no jumps, this is equal to

$$\text{Cov}(\xi_j, \xi_k) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left( [1 - K_j^F(x) - K_k^F(y)][F_{jk}(x, y) - F_j(x)F_k(y)] \right.$$

$$+ [K_{jk}^F(x, y)F_{jk}(x, y) - K_j^F(x)K_k^F(y)F_j(x)F_k(y)] \left. \right) dm_j(x)dm_k(y),$$

where

$$F_{jk}(x, y) := \Pr(X_{i,j} \leq x, X_{i,k} \leq y),$$

$$K_{jk}^F(x, y) := \lim_{n \to \infty} \mathbb{E}[w_{i,n,j}w_{i,n,k} | X_{i,j} \leq x, X_{i,k} \leq y].$$

Given $m_j$ and $m_k$ known, this can be consistently estimated by its sample analogue

$$\hat{\text{Cov}}(\xi_j, \xi_k) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left( [1 - \mathbb{K}_{n,j}^F(x) - \mathbb{K}_{n,k}^F(y)][\mathbb{F}_{n,jk}(x, y) - \mathbb{F}_{n,j}(x)\mathbb{F}_{n,k}(y)] \right.$$

$$+ [\mathbb{K}_{n,jk}^F(x, y)\mathbb{F}_{n,jk}(x, y) - \mathbb{K}_{n,j}^F(x)\mathbb{K}_{n,k}^F(y)\mathbb{F}_{n,j}(x)\mathbb{F}_{n,k}(y)] \left. \right) dm_j(x)dm_k(y),$$

where $\mathbb{F}_{n,jk}(x, y) := \mathbb{E}_n[\mathbb{1}\{X_{i,j} \leq x, X_{i,k} \leq y\}]$ and $\mathbb{K}_{n,jk}^F(x, y) := \mathbb{E}_n[w_{i,n,j}w_{i,n,k} | X_{i,j} \leq x, X_{i,k} \leq y].$

The derivative formulas give us

\[
\text{Cov}(\xi_j, \xi_k) = \int_0^1 \int_0^1 \int_0^1 m_j'(Q_j(s))Q_j'(s)m_k'(Q_k(t))Q_k'(t)[F_{jk}(s, t) - st]dsdt
\]

\[
+ \int_0^1 \int_0^1 \int_0^1 m_j'(Q_j(s))Q_j'(s)m_k'(Q_k(t))Q_k'(t)[K_{jk}(s, t)F_{jk}(s, t) - stK_j(s)K_k(t)]dsdt
\]

\[
- \int_0^1 \int_0^1 \int_0^1 m_j'(Q_j(s))Q_j'(s)m_k'(Q_k(t))Q_k'(t)K_j(s)F_{jk}(s, t)dsdt
\]

\[
- \int_0^1 \int_0^1 \int_0^1 m_j'(Q_j(s))Q_j'(s)m_k'(Q_k(t))Q_k'(t)K_k(t)[F_{jk}(s, t) - st]dsdt,
\]

where \(K_{jk}(s, t) := \lim_{n \to \infty} \mathbb{E}[w_{i,n,j}w_{i,n,k} | X_{i,j} \leq Q_j(s), X_{i,k} \leq Q_k(t)]\) and \(F_{jk}(s, t) := \Pr(X_{i,j} \leq Q_j(s), X_{i,k} \leq Q_k(t))\). Apply the change of variables \(x = Q_j(s)\) and \(y = Q_k(t)\) to obtain the second formula. Consistency of the sample analogue estimator follows by uniform convergence of \(K_{n,j}^F\) and \(K_{n,k}^F\) and Addendum 1.H.1.  

1.G Validity of Nonparametric Bootstrap

In this section, we establish the validity of nonparametric bootstrap. Since we have proved the Hadamard differentiability of maps involved, it remains to show the weak convergence of bootstrap processes conditional on the original observations.

The bootstrap process is given by

\[
\hat{Z}_n(x) := \sqrt{n}(\hat{F}_n - F_n)(x) := \frac{1}{\sqrt{n}} \sum_{i=1}^n (M_{ni} - 1) \mathbb{1}\{X_i \leq x\}
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^n (M_{ni} - 1)(\mathbb{1}\{X_i \leq x\} - F(x))
\]

where \(M_{ni}\) is the number of times \(X_i\) is drawn in the bootstrap sample. We want to prove that this process converges weakly to the same limit as \(Z_n := \sqrt{n}(\hat{F}_n - F)\) conditional on the observations \(X_i\). The strategy is very similar to the one employed in Van der Vaart and Wellner (1996, Chapter 3.6) and goes as follows: since \(M_{ni}\) sums up to \(n\), it is slightly dependent to each other; we replace \(M_{ni}\) with independent Poisson random variables \(\xi_i\) by Poissonization, that is, we show equiva-
lence of weak convergence of the bootstrap process \( \hat{Z}_n \) and the multiplier process 
\[ Z'_n := n^{-1/2} \sum \xi_i (\mathbb{1}\{X_i \leq x\} - F) \] (Lemma 1.G.1; then, we prove unconditional convergence of \( Z'_n \) (so randomness comes from both \( X_i \) and \( \xi_i \)) by symmetrization, which replaces \( \xi_i \) with independent Rademacher random variables \( \varepsilon_i \) (Lemma 1.G.2); next, we show conditional convergence of \( Z'_n \) conditional on \( Z_n \) (so randomness comes only from \( \xi_i \)) by discretizing \( Z'_n \) (Lemma 1.G.3).

We observe that many results in Van der Vaart and Wellner (1996, Chapters 2.9, 3.6, and A.1) translate directly to our norm \( \lambda_{\mu} \); therefore, we will not reproduce the entire discourse but rather note when this is the case and prove results that require modification to accommodate our norm. Additionally, in order to show validity of bootstrap for \( L \)-statistics, we also need to show conditional weak convergence of the bootstrap sample selection process \( \sqrt{n}(\hat{K}_n - K_n) \). We restrict attention to the case of sample selection based on sample quantiles and show validity by writing the sample selection process as a function of the sum of i.i.d. random variables (Lemma 1.G.4).

We first prove the key lemma used in Poissonization, the counterpart of Van der Vaart and Wellner (1996, Lemma 3.6.16).

**Lemma 1.G.1.** For each \( n \), let \((W_{n1}, \ldots, W_{nn})\) be an exchangeable nonnegative random vector independent of \( X_1, X_2, \ldots \) such that \( \sum_{i=1}^{n} W_{ni} = 1 \) and \( \max_{1 \leq i \leq n} |W_{ni}| \) converges to zero in probability. Then, for every \( \varepsilon > 0 \), as \( n \to \infty \),

\[
\Pr_{W} \left( \left\| \sum_{i=1}^{n} W_{ni} (\mathbb{1}\{X_i \leq x\} - F(x)) \right\|_{\mu} > \varepsilon \right) \xrightarrow{\text{as}} 0.
\]

**Proof.** Assume without loss of generality that \( \mu \) is a positive measure, that is, \( m \) is increasing (we may do so since \( m \) is of locally bounded variation). Since Van der Vaart and Wellner (1996, Lemma 3.6.7) can be restated with our norm \( \| \cdot \|_{\lambda_{\mu}} \), the proof of this lemma is almost identical to that of Van der Vaart and Wellner (1996, Lemma 3.6.16) modulo the norm. Essentially, the only part that requires modification is the boundedness of \( n^{-1} \sum_{i=1}^{n} \| \mathbb{1}\{X_i \leq x\} - F(x) \|_{\mu} \). Note that

\[
|\mathbb{1}\{X_i \leq x\} - F(x)| \leq |\mathbb{1}\{X_i \leq x\} - \mathbb{1}\{0 \leq x\}| + |\tilde{F}(x)|.
\]
Therefore,
\[ \| \mathbb{1}\{X_i \leq x\} - F(x)\|_\mu \leq |m(X_i) - m(0)| + \|\tilde{F}\|_\mu. \]

Find that
\[ \frac{1}{n} \sum_{i=1}^{n} \| \mathbb{1}\{X_i \leq x\} - F(x)\|_\mu^r \leq \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - m(0)|^r + \|\tilde{F}\|_\mu^r, \]
which converges almost surely to \( \mathbb{E}[|m(X_i) - m(0)|^r] + \|\tilde{F}\|_\mu^r \), which is finite.

Given this lemma, we may infer by the same arguments as Van der Vaart and Wellner (1996, Theorem 3.6.1) that conditional weak convergence of the bootstrap process \( \hat{Z}_n \) follows from conditional weak convergence of the multiplier process \( Z'_n \).

Before moving on to the conditional convergence, however, we need to show the unconditional convergence of the multiplier process \( Z'_n \) in our norm. For a random variable \( \xi \), we use the notation
\[ \|\xi\|_{2,1} := \int_0^\infty \sqrt{\mathbb{P}(|\xi| > x)}dx. \]
That \( \|\xi\|_{2,1} < \infty \) means that \( \xi \) has slightly more than a variance. The following is a modification of Van der Vaart and Wellner (1996, Theorem 2.9.2).

**Lemma 1.G.2.** Let \( m : \mathbb{R} \rightarrow \mathbb{R} \) be a function of locally bounded variation and \( \mu \) the Lebesgue-Stieltjes measure associated with \( m \). Let \( \xi_1, \ldots, \xi_n \) be i.i.d. random variables with mean zero, variance 1, and \( \|\xi\|_{2,1} < \infty \), independent of \( X_1, \ldots, X_n \). For a probability distribution \( F \) on \( \mathbb{R} \) such that \( m(X) \) has a \((2 + c)\)th moment for \( X \sim F \) and some \( c > 0 \), the process \( Z'_n(x) := n^{-1/2} \sum_{i=1}^{n} \xi_i[\mathbb{1}\{X_i \leq x\} - F(x)] \) converges weakly to a tight limit process in \( \mathbb{L}_n \) if and only if \( Z_n := n^{-1/2} \sum_{i=1}^{n} [\mathbb{1}\{X_i \leq x\} - F(x)] \) does. In that case, they share the same limit processes.

**Proof.** By Proposition 1.D.1 and Van der Vaart and Wellner (1996, Theorem 2.9.2), marginal convergence and asymptotic equicontinuity of \( Z'_n \) are trivial. It remains to show the equivalence of asymptotic equiintegrability of \( Z'_n \) and \( Z_n \).

Note that the proofs of Van der Vaart and Wellner (1996, Lemmas 2.3.1, 2.3.6,
and 2.9.1 and Propositions A.1.4 and A.1.5) do not depend on the specificity of the norm \(\|\cdot\|_x\); therefore, they continue to hold with \(\|\cdot\|_{\mu}\). Given these, Van der Vaart and Wellner (1996, Lemma 2.3.11) also holds with \(\|\cdot\|_{\mu}\) (and \(\|\cdot\|_{\mu,\delta_n}\)). Finally, rewriting the proof of Van der Vaart and Wellner (1996, Theorem 2.9.2) in terms of \(\|\cdot\|_{\mu}\) yields the proof of this lemma.

Third, we show conditional convergence of the multiplier process \(Z'_n\) using the above result. Van der Vaart and Wellner (1996, Theorem 2.9.6).

**Lemma 1.G.3.** Let \(m : \mathbb{R} \to \mathbb{R}\) be a function of locally bounded variation and \(\mu\) the Lebesgue-Stieltjes measure associated with \(m\). Let \(\xi_1, \ldots, \xi_n\) be i.i.d. random variables with mean zero, variance 1, and \(\|\xi\|_{2,1} < \infty\), independent of \(X_1, \ldots, X_n\). For a probability distribution \(F\) on \(\mathbb{R}\) such that \(m(X)\) has a \((2 + c)\)th moment for \(X \sim F\) and some \(c > 0\), the process \(Z'_n(x) = n^{-1/2} \sum_{i=1}^n \xi_i \mathbb{1}\{X_i \leq x\} - F(x)\) satisfies

\[
\sup_{h \in \text{BL}_1(\mu)} \left| \mathbb{E} \xi h(Z'_n) - \mathbb{E} h(Z) \right| \longrightarrow 0
\]

in outer probability, and the sequence \(Z'_n\) is asymptotically measurable.

**Proof.** By Lemma 1.G.2, \(Z'_n\) is asymptotically measurable. Assume without loss of generality that \(m\) is continuous and strictly monotone (see footnote 22), and define a semimetric \(\rho\) on \(\mathbb{R}\) by

\[
\rho(s, t) = \left( \int_s^t \left( (m(x))^{1+c} \vee 1 \right) |\bar{F}(x)| d\mu(x) \right)^{1/2}.
\]

For \(\delta > 0\), \(t_1 < \cdots < t_p\) be such that \(\rho(-\infty, t_1) \leq \delta\), \(\rho(t_j, t_{j+1}) \leq \delta\), and \(\rho(t_p, \infty) \leq \delta\). Define \(Z_\delta\) by

\[
Z_\delta(x) := \begin{cases} 
0 & x < t_1 \text{ or } x \geq t_p, \\
Z(t_i) & t_i \leq x \leq t_{i+1}, i = 1, \ldots, p - 1.
\end{cases}
\]

Define \(Z'_{n,\delta}\) analogously. By the continuity and integrability of the limit process \(Z\),
we have $Z_\delta \to Z$ almost surely as $\delta \to 0$. Therefore,

$$\sup_{h \in \text{BL}_1(\mathbb{L}_\mu)} \left| \mathbb{E}h(Z_\delta) - \mathbb{E}h(Z) \right| \to 0 \quad \text{as} \quad \delta \to 0.$$ 

Second, by [Van der Vaart and Wellner (1996, Lemma 2.9.5)],

$$\sup_{h \in \text{BL}_1(\mathbb{L}_\mu)} \left| \mathbb{E}_\xi h(Z'_{n,\delta}) - \mathbb{E}h(Z_\delta) \right| \to 0 \quad \text{as} \quad n \to \infty$$

for almost every sequence $X_1, X_2, \ldots$ and fixed $\delta > 0$. Since $Z_\delta$ and $Z'_{n,\delta}$ take only on a finite number of values and their tail values are zero, one can replace the supremum over $\text{BL}_1(\mathbb{L}_\mu)$ with a supremum over $\text{BL}_1(\mathbb{R}^p)$. Observe that $\text{BL}_1(\mathbb{R}^p)$ is separable with respect to the topology of uniform convergence on compact sets; this supremum is effectively over a countable set, hence measurable. Third,

$$\sup_{h \in \text{BL}_1(\mathbb{L}_\mu)} \left| \mathbb{E}_\xi h(Z'_{n,\delta}) - \mathbb{E}_\xi h(Z'_n) \right| \leq \sup_{h \in \text{BL}_1(\mathbb{L}_\mu)} \mathbb{E}_\xi |h(Z'_{n,\delta}) - h(Z'_n)|$$

$$\leq \mathbb{E}_\xi \|Z'_{n,\delta} - Z'_n\|_{\mathbb{L}_\mu} \leq \mathbb{E}_\xi \|Z'_n\|_{\mathbb{L}_\mu,\delta}.$$ 

This implies that the outer expectation of the LHS is bounded above by $\mathbb{E}^*\|Z'_n\|_{\mathbb{L}_\mu,\delta}$, which vanishes as $n \to \infty$ by the modified [Van der Vaart and Wellner (1996, Lemma 2.9.1)] as in Lemma 1.G.2.

These results show that nonparametric bootstrap works for the empirical process $\sqrt{n}(F_n - F)$ and the empirical quantile process $\sqrt{n}(Q_n - Q)$. However, we also need to verify validity for the sample selection process $\sqrt{n}(K_n - K)$. The key in the proof is to represent $K_n$ in terms of Hadamard differentiable functions of “$F_n$” in Proposition 1.F.3.

**Lemma 1.G.4.** Let $U_1, \ldots, U_n$ be independent uniformly distributed random variables on $(0, 1)$ and $\xi_1, \ldots, \xi_n$ be i.i.d. random variables with mean zero, variance 1, and $\|\xi\|_{2,1} < \infty$, independent of $U_1, \ldots, U_n$. Define the bootstrapped empirical process of
\[ U \] by

\[ F'_n(u) := \frac{1}{n} \sum_{i=1}^{n} \xi_i \mathbb{1}\{U_i \leq u\}, \]

and let \( w'_{i,n} \) be the indicator of whether \( U_i \) is above the \( \alpha \)-quantile of the bootstrapped sample, that is, \( w'_{i,n} = \mathbb{1}\{U_i > F'_n^{-1}(\alpha)\} \). Define

\[ G'_n(u) := \frac{1}{n} \sum_{i=1}^{n} \xi_i w'_{i,n} \mathbb{1}\{U_i \leq u\}. \]

Then, for \( F(u) = 0 \vee u \wedge 1 \) and \( G(u) = 0 \vee (u - \alpha) \wedge (1 - \alpha) \),

\[
\sup_{h \in BL_1(L_{\nu_2})} \left| \mathbb{E}_x h(\sqrt{n}(F'_n - F)) - \mathbb{E} h(\sqrt{n}(F_n - F)) \right| \to 0,
\]

\[
\sup_{h \in BL_1(L_{\nu_2})} \left| \mathbb{E}_x h(\sqrt{n}(G'_n - G)) - \mathbb{E} h(\sqrt{n}(G_n - G)) \right| \to 0,
\]

in outer probability, and the sequences \( \sqrt{n}(F_n - F) \) and \( \sqrt{n}(G'_n - G) \) are asymptotically measurable.

**Proof.** Noting that \( G'_n(u) = 0 \vee [F'_n(u) - F'_n \circ F'_n^{-1}(\alpha)] \), weak convergence of \( \sqrt{n}(F'_n - F) \) and \( \sqrt{n}(G'_n - G) \) follows from Lemmas 1.F.4 and 1.G.3. \[\blacksquare\]

**Remark.** Note that Lemma 1.G.4 immediately implies the validity of bootstrap for any selection of samples based on a finite number of empirical quantiles.

Now we show the validity of nonparametric bootstrap when sample selection is based on empirical quantiles.

**Proposition 1.G.5 (Validity of nonparametric bootstrap).** In the assumptions stated in Proposition 1.F.5, assume further that \( w_{i,n,j} \) represents sample selection based on a fixed number of empirical quantiles. Then, the joint distribution of \((\hat{\beta}_1, \ldots, \hat{\beta}_d)\) can be consistently estimated by nonparametric bootstrap. The algorithm is as follows.

**Here,** \( X_i \) denotes a vector \((X_{i,1}, \ldots, X_{i,d})\).

**i.** Bootstrap \( n \) (or fewer) random observations from \( X_1, \ldots, X_n \) with replacement.

\[24\] The assumption on convergence must be extended (from bivariate) to joint over all processes involved.
ii. Compute the statistics \( (\hat{\beta}_1^*, \ldots, \hat{\beta}_d^*) \) for the bootstrapped sample.

iii. Repeat the above steps \( S \) times.

iv. Use the empirical distribution of \( (\hat{\beta}_1^*, \ldots, \hat{\beta}_d^*) \) as the approximation to the theoretical asymptotic distribution of \( (\hat{\beta}_1, \ldots, \hat{\beta}_d) \).


1.H Glivenko-Cantelli Type Results

We state some of the Glivenko-Cantelli type results for distribution functions and quantile functions when we have additional information about their expectations. The results are generally stronger than the classical Glivenko-Cantelli theorems. It is noteworthy that the Glivenko-Cantelli theorem of quantile functions holds for the \( L_p \) spaces (see Appendix [1.E] for discussion). Related results are found in Parzen (1980) and Csörgő and Horváth (1993).

The first addendum provides a stronger Glivenko-Cantelli result for the distribution function when the underlying distribution has a finite expectation when transformed by a function \( m \).

Addendum 1.H.1. Let \( m : \mathbb{R} \to \mathbb{R} \) be a function of locally bounded variation and \( \mu \) the associated Lebesgue-Stieltjes measure. For a probability measure \( F \) on \( \mathbb{R} \) such that \( m(X) \) has a finite expectation for \( X \sim F \), we have

\[
\left\| m(t) \bar{F}_n(t) - m(t)F(t) \right\|_{\infty} \xrightarrow{\text{as*}} 0, \quad \left\| \int_{[s,t]} |m|d\bar{F}_n - \int_{[s,t]} |m|dF \right\|_{\infty} \xrightarrow{\text{as*}} 0, \\
\left\| \int_{[s,t]} |\bar{F}_n|d\mu - \int_{[s,t]} |\bar{F}|d\mu \right\|_{\infty} \xrightarrow{\text{as}} 0, \quad \int_{\mathbb{R}} |\bar{F}_n - F||d\mu| \xrightarrow{\text{as}} 0,
\]

where the supremum is taken respectively over \( t \in \mathbb{R} \), \( (s,t) \in \mathbb{R}^2 \), and \( (s,t) \in \mathbb{R}^2 \).

Proof. Note first that since a function of bounded variation can be represented as a difference of two increasing functions, we may assume without loss of generality that
$m$ is increasing (and hence $\mu$ is a positive measure). Moreover, we may also assume $m(0) = 0$ for otherwise the residual terms $m(0)(F_n - F)$ of the first two quantities vanish by the classical Glivenko-Cantelli theorem.

In view of [Van der Vaart and Wellner (1996) Theorem 2.4.1], to prove the first two claims it suffices to show that the classes of functions,

$$\mathcal{F} = \left\{ f_t : \mathbb{R} \to \mathbb{R} : t \in \mathbb{R}, f_t(x) = m(t) \mathbbm{1}\{x \leq t\}\right\},$$

$$\mathcal{G} = \left\{ g_{s,t} : \mathbb{R} \to \mathbb{R} : s, t \in \mathbb{R}, g_{s,t}(x) = |m(x)| \mathbbm{1}\{s \leq x \leq t\}\right\},$$

have finite bracketing numbers with respect to $L_1(P)$, i.e., $N_\varepsilon(\mathcal{F}, L_1(F)) < \infty$ and $N_\varepsilon(\mathcal{G}, L_1(F)) < \infty$ for every $\varepsilon > 0$. For $\mathcal{F}$ take $-\infty = t_0 < t_1 < \cdots < t_m = \infty$ such that $|\int (f_{t_{i+1}} - f_{t_i}) dF| < \varepsilon$ for each $i$ and consider the brackets $\{f_t\}$

Such a partition is finite since $m(X)$ has a finite first moment and by Lemma 1.C.1. For $\mathcal{G}$ take $-\infty = t_0 < t_1 < \cdots < t_m = \infty$ such that $|\int_{(-\infty,t_{i+1}[} |m| dF - \int_{(-\infty,t_i[} |m| dF| < \varepsilon$ for each $i$, then consider the brackets $\{g_{s,t}\}$ for every pair $s, t \in \{t_0, \ldots, t_m\}$

Such a partition is finite by the assumption that $m(X)$ has a finite first moment.

To prove the third claim, observe that by integration by parts,

$$\int_{[s,t]} |m| dF_n = \int_{[s,t]} |m| d\bar{F}_n = \left[ |m| \bar{F}_n \right]_s^t + \int_{[s,t]} |\bar{F}_n| d\mu.$$  

Then the claim follows in observation of the following triangle inequality and the two claims proved so far:

$$\| \int_{[s,t]} |\bar{F}_n| d\mu - \int_{[s,t]} |\bar{F}| d\mu \|_{\infty} \leq 2 \| m(t) \bar{F}_n(t) - m(t) F(t) \|_{\infty}$$

$$+ \| \int_{[s,t]} |m| dF_n - \int_{[s,t]} |m| dF \|_{\infty}. $$

To prove the last claim, observe that Lemma 1.C.1 and the preceding claim imply

\[\text{25If } F\text{ has a probability mass at } t\text{, then for small } \varepsilon\text{ take, instead of } f_t, \tilde{f}_{t,c}(x) = m(t)\mathbbm{1}\{x \leq t\} + (1 - c)\mathbbm{1}\{x < t\}\text{ for appropriately chosen } c.\]

\[\text{26Again, if } F\text{ has a mass, similar adjustments are required.}\]

\[\text{27Measurability of the sup on the LHS follows by continuity of the Lebesgue integrals.}\]
that for $\varepsilon > 0$ there exists $M < \infty$ such that

$$
\left( \int_{(-\infty,-M]} + \int_{[M,\infty)} \right) |\tilde{F}_n| d\mu + \left( \int_{(-\infty,-M]} + \int_{[M,\infty)} \right) |F| d\mu < \varepsilon
$$

with probability tending to 1. By the triangle inequality,

$$
\int_{\mathbb{R}} |\tilde{F}_n - \tilde{F}| d\mu \leq \int_{(-M,M)} |\tilde{F}_n - \tilde{F}| d\mu + \varepsilon \leq \|\tilde{F}_n - F\|_{\infty} \mu((-M,M)) + \varepsilon.
$$

Then the assertion follows by the Glivenko-Cantelli theorem. 

We next provide the Glivenko-Cantelli results for the quantile functions. Two points are interesting. First, despite the fact that we were unable to prove Hadamard differentiability of $F \mapsto Q$ as a map from $L_p$ to $L_p$, the Glivenko-Cantelli result still holds for the $L_p$ norm. Second, the addendum gives the “sup norm” for quantile functions. Although the quantile function for an unbounded random variable is unbounded, we still have a reasonable pseudo-uniform convergence when $Q$ is continuous.

**Addendum 1.H.2.** Let $F$ be a probability distribution on $\mathbb{R}$ with a $p$th moment for $p > 0$ and $Q := F^{-1}$. Then

$$
\left( \int_0^1 |Q_n - Q|^p du \right)^{1/p} \xrightarrow{as} 0.
$$

Moreover, if and only if $Q$ is continuous, we have

$$
\|u^{1/p}(1-u)^{1/p}(Q_n - Q)\|_{\infty} \xrightarrow{as} 0,
$$

where the supremum is taken over $u \in (0,1)$.

**Proof.** By the strong law of large numbers on $Y = |X|^p$, $X \sim F$, we have

$$
\int_{-\infty}^{\infty} |x|^p d\tilde{F}_n - \int_{-\infty}^{\infty} |x|^p dF \xrightarrow{as} 0.
$$
Applying the change of variables,
\[
\int_0^1 |Q_n|^p du - \int_0^1 |Q|^p du \xrightarrow{\text{as}} 0.
\]

In view of this (and since \(Q\) is in \(L_p\) by Lemma 1.C.1), for \(\varepsilon > 0\) one can take \(\delta > 0\) such that
\[
\left(\int_0^{2\delta} + \int_{1-2\delta}^1\right) |Q_n|^p du < \varepsilon,
\]
\[
\left(\int_0^{2\delta} + \int_{1-2\delta}^1\right) |Q|^p du < \varepsilon,
\]
with probability tending to 1. Combination with the triangle inequality allows us to bound the target as
\[
\int_0^1 |Q_n - Q|^p du \leq \int_{2\delta}^{1-2\delta} |Q_n - Q|^p du + 2\varepsilon.
\]

Moreover, observe
\[
\int_{2\delta}^{1-2\delta} |Q_n - Q|^p du \leq \varepsilon^p + \int_{2\delta}^{1-2\delta} |Q_n - Q|^p \mathbb{1}\{|Q_n - Q| > \varepsilon\} du.
\]

We want to bound the last integrand by a multiple of \(|Q_n - Q|\) to eliminate the \(p\)th power, since then we may further apply Fubini’s theorem to eliminate the inverse. Toward this goal, we aim to use the following inequality: for \(p > 0\) and \(M \geq 0\),
\[
|x|^p \mathbb{1}\{|x| > \varepsilon\} \leq \left(\varepsilon^{p-1} \vee M^{p-1}\right) |x| \quad \text{for every} \quad |x| \leq M.
\]

But to do that, we need to find the bound \(M\) on \(|Q_n - Q|\) over \((2\delta, 1 - 2\delta)\) that does not depend on \(n\).

By the classical Glivenko-Cantelli theorem, we may take \(n\) large enough so that
\[
\|F_n - F\|_\infty < \delta \quad \text{with outer probability at least} \quad 1 - \varepsilon.
\]

This implies \(F - \delta \leq F_n \leq F + \delta\), and since \(F\) and \(F_n\) are nondecreasing,
\[
(F - \delta)^{-1} \leq F_n^{-1} = Q_n \leq (F + \delta)^{-1}.
\]
This yields the following bounds on $Q_n$ over the region of integration.

\[
Q_n(1 - 2\delta) \leq (F + \delta)^{-1}(1 - 2\delta) = Q(1 - \delta),
\]

\[
Q_n(2\delta) \geq (F - \delta)^{-1}(2\delta) = Q(\delta).
\]

Note that nondecreasingness of $F$ implies analogous inequalities for $Q$ itself, namely, $Q(1 - 2\delta) \leq Q(1 - \delta)$ and $Q(2\delta) \geq Q(\delta)$. Therefore, the difference $|Q_n - Q|$ is bounded by $M := Q(1 - \delta) - Q(\delta)$ over the region of integration.

Given this, we can successfully bound the last integral by

\[
\left(\varepsilon^{p-1} \vee M^{p-1}\right) \int_{2\delta}^{1-2\delta} |Q_n - Q| du.
\]

Finally, invoke Fubini’s theorem to find

\[
\int_{2\delta}^{1-2\delta} |Q_n - Q| du = \int_{2\delta}^{1-2\delta} \int_{Q \cap Q_n} dx du \leq \int_{Q(\delta)}^{Q(1-\delta)} |F_n - F| dx \leq M \|F_n - F\|_{\infty},
\]

which vanishes outer almost surely.

Next, we prove the second claim. Since $Q \in L_p$, $Q_n \in L_p$, and $Q_n$ converges to $Q$ in $L_p$, for every $\varepsilon > 0$ there exists $\delta$ such that

\[
\left(\int_0^{\delta} + \int_{1-\delta}^1\right) \left(|Q_n|^p + |Q|^p\right) du < \varepsilon
\]

with probability tending to 1. When this inequality holds, we argue that $|Q_n|^p$ and $|Q|^p$ never exceed the function $\varepsilon/[u(1-u)]$ on $u \in (0, \delta) \cup (1 - \delta, 1)$. Suppose otherwise (without loss of generality consider $Q$ only) and let $u$ be the point of exceedance. Since $Q$ is monotone, the area of $|Q|^p$ must contain either the left rectangle $(0, u) \times (0, \varepsilon/[u(1-u)])$ or the right rectangle $(u, 1) \times (0, \varepsilon/[u(1-u)])$. In either case, the integral of $|Q|^p$ over $(0, \delta) \cup (1 - \delta, 1)$ must exceed $\varepsilon$, since each rectangle has area bigger than $\varepsilon$. This is a contradiction to the assumption that the integral is less than $\varepsilon$.

This ensures that the difference $|Q_n - Q|^p$ is less than $2^p \varepsilon/[u(1-u)]$ on this region.
In other words, the supremum of $u(1-u)|Q_n - Q|^p$ over $(0, \delta) \cup (1-\delta, 1)$ is less than $2^p\varepsilon$. Meanwhile, if $Q$ is continuous, then $Q_n$ converges to $Q$ pointwise on any fixed closed interval of $(0, 1)$ since $Q_n$ and $Q$ are monotone. Therefore, the supremum of $u(1-u)|Q_n - Q|^p$ over $[\delta, 1-\delta]$ must vanish as $n \to \infty$. Combining the results, sufficiency follows.

On the other hand, suppose $Q$ is discontinuous at $u \in (0, 1)$. By construction $Q_n$ can only have discontinuity points on $\{1/n, \ldots, n/n\}$. But there exist infinitely many $n$ such that $u \notin \{1/n, \ldots, n/n\}$. Therefore, $|Q_n - Q|$ can infinitely often be as large as half the jump height at $u$. ■
Bibliography


Chapter 2

Theory of Weak Identification in Semiparametric Models

Weak identification is a widespread concern in empirical economics, yet theoretical work has proposed valid asymptotic approximations and robust inference procedures only on a case-by-case basis. This chapter provides a general formulation of weak identification in semiparametric models and a meaningful efficiency concept for weakly identified parameters. The general formulation reveals that weak identification occurs as a result of a parameter depending asymptotically on the model score; we call such a parameter weakly regular. This dependence is fundamentally nonlinear, hence the non-normality of asymptotic distributions observed in the literature. As a result, consistent or equivariant (e.g., pivotal) estimation when a parameter is weakly regular is shown to be impossible. We further show that behind every weakly regular parameter there exists an underlying parameter that is regular and controls the limit behavior of the weakly regular parameter. While this parameter is not unique, concepts of sufficiency and minimality help pin down the desirable underlying regular parameter, which can then be used to define an efficiency concept for weakly regular parameters. In particular, the presence of noise in an estimator of the minimal sufficient underlying parameter introduces noise in the corresponding estimator of the weakly regular parameter, whence we can construct an estimator more concentrated toward the same mean by local asymptotic Rao-Blackwellization. Thus, we call an estimator
of a weakly regular parameter *weakly efficient* if it attains an asymptotic distribution that does not admit such an improvement. We apply our results to heteroskedastic linear IV models and demonstrate that 2SLS, GMM, Fuller, and unbiased estimators can be improved (made less dispersed) under the availability of an efficient estimator for the reduced-form coefficients and other conditions.

### 2.1 Introduction

Weak identification arises in a wide range of empirical settings. A leading example is the linear instrumental variables (IV) model in which the instruments and endogenous regressors are barely correlated (Nelson and Startz, 1990; Bound et al., 1995). When this happens, even with a large sample, classical asymptotic theory is known to yield poor approximations to the behavior of familiar statistics (Staiger and Stock, 1997), causing problems in both estimation and inference. We encounter this problem in various other contexts: Stock and Wright (2000) analyze weak identification in generalized method of moments (GMM) models; Guggenberger and Smith (2005, 2008) and Otsu (2006) in generalized empirical likelihood (GEL) models; Andrews and Cheng (2012), Han and McCloskey (2017), and Cox (2017) in extremum estimation models; Iskrev (2008), Ruge-Murcia (2007), and Canova and Sala (2009) in dynamic stochastic general equilibrium (DSGE) models; Armstrong (2016) in differentiated products demand estimation models. Many estimators of weakly identified parameters exhibit inconsistency and bias, and, as a consequence, standard inference procedures such as *t*- and Wald tests may have substantially distorted sizes (Phillips, 1984, 1989; Dufour, 1997; Hirano and Porter, 2015 as well as aforementioned papers).

Following these practically challenging problems, a vast amount of theoretical work has been published.

The theoretical literature on weak identification is largely confined to specific estimation and inference procedures in specific models. Many papers consider consequences of particular asymptotic embeddings on particular statistics, find statistics that are well-behaved, and derive various types of robust statistical procedures in
many practically relevant models, especially in the linear IV model. In contrast, many fundamental questions—such as what is the common cause of known instances of weak identification, what is a general guideline to look for well-behaved statistics, and what is the semiparametric efficiency in the presence of weak identification—have been largely left unanswered. Such exploration is essential, however, in order not only to facilitate unified understanding of the phenomenon but to measure semiparametric performance of different procedures and develop general systematic construction methods for estimation and inference. This is more important than it has ever been, especially now that numerous robust and efficient inference procedures have been developed in many individual settings.

This chapter studies weak identification from the perspective of semiparametric theory. We explore how weak identification emerges and how it behaves in the classical framework of Bickel et al. (1993), Van der Vaart (1998 Chapter 25), and Kosorok (2008 Part III). We find that weak identification occurs as a result of the parameter’s asymptotic dependence on the score; we call such a parameter \textit{weakly regular}. This is in stark contrast to the classical \textit{regular} parameters, whose derivatives (local parameters) depend on the score, not the parameters themselves. As an immediate consequence of this observation, we derive—without reference to a specific estimation or inference procedure—that there exists neither a consistent estimator, a consistent test, nor an equivariant (hence pivotal) estimator when the parameter is weakly regular. The dependence on the score is homogeneous of degree zero and essentially nonlinear, and this nonlinearity is the root cause of many non-Gaussian nonpivotal asymptotic distributions witnessed throughout the literature (Staiger and Stock, 1997; Stock and Wright 2000; Guggenberger and Smith, 2005; Andrews and Cheng, 2012; Cox, 2017). To circumvent the problem of almost arbitrary nonlinearity, we seek ways to explore weak regularity from the standpoint of \textit{regular} parameters.

We show that behind every weakly regular parameter there exists an underlying parameter that is regular and controls the limit behavior of the weakly regular parameter. In other words, a weakly regular parameter can be represented as a nonlinear transformation of the local parameter of some regular parameter. Finding
such a parameter allows us to reformulate the model in such a way that it consists only of regular parameters and thus provides a tractable foundation on which to discuss estimation and inference easily. This consorts with the repeated observation in the literature that reduction to regular parameters (usually referred to as “reduced-form parameters”) can substantially simplify the problems (Staiger and Stock, 1997; Stock and Wright, 2000; Chernozhukov et al., 2009; Magnusson and Mavroeidis, 2010; Magnusson, 2010; Guerron-Quintana et al., 2013; Andrews and Mikusheva, 2016a,b; Andrews, 2016; Cox, 2017, among many others); we generalize this observation to arbitrary semiparametric models and show that there exists an underlying regular parameter for every weakly regular parameter. However, underlying regular parameters are not unique, and statistical analyses based on different underlying parameters may yield different performances. This gives rise to the need for criteria to choose which underlying parameter to use.

We consider desirable properties of underlying parameters from two perspectives. Intuitively, a good underlying regular parameter would exhaustively contain all information about the weakly regular parameter that can be inferred by the model, and it would contain no irrelevant information that may lead to noisy analyses; this intuition parallels that of efficient influence functions of classical nuisance parameter theory. In light of this, we define an underlying parameter to be sufficient if knowing the value of its local parameter reveals as much information as knowing the weakly regular parameter. The key is to understand that information about the weakly regular parameter comes from two sources: the value of the weakly regular parameter and the very fact that it is identified. A sufficient underlying parameter would contain both pieces of information. Next, we define an underlying parameter to be minimal if knowing its local parameter does not reveal more information than knowing the weakly regular parameter. If it does, its estimation would create additional noise in an effort to estimate its unnecessary “nuisance” component. In short, the best underlying regular parameter would be minimal and sufficient. We show existence of minimal sufficient underlying parameters, provide a way to assess their sufficiency and minimality in general setups, and present examples of minimal sufficient underlying
parameters.

With these concepts, we define a new notion of efficiency for estimating weakly regular parameters. Efficiency of estimation under weak identification has received little treatment in the literature. This is because non-Gaussianity and nonpivotality of the asymptotic distributions render the classical efficiency concepts, namely the convolution and minimax theorems, inapplicable. Our formulation, on the other hand, enables us to decompose estimation of weakly regular parameters into estimation of the minimal sufficient underlying regular parameters and their transformation. As the underlying regular parameters admit the classical convolution theorem, efficiency of their estimation can be discussed through the classical theory. Moreover, if the estimators of the underlying parameters contain unnecessary noise, then their transformations would also contain unnecessary noise. Such noise can then be eliminated by taking expectation with respect to it since the noise and the asymptotic distributions of efficient estimators are asymptotically independent. Conceptually, this corresponds to applying the Rao-Blackwell theorem to the local asymptotic representations of the estimators, exploiting the fact that the efficient asymptotic distributions of regular parameters are “sufficient” in the local expansion. The resulting conditional expectation estimators are, as a consequence, more concentrated toward the same means without affecting the size of the biases. We formalize this idea as a theorem and name it local asymptotic Rao-Blackwellization (LAR). If such improvement is impossible, we call the estimators weakly efficient. We put the qualifier “weakly” as weakly efficient estimators are not unique. We also discuss relationship between weak efficiency and classical efficiency.

Most estimators in the literature can be written as functions of estimators of some underlying regular parameters and thus fall within the class of estimators covered by our results. We apply our results to heteroskedastic linear IV models and present examples of weakly efficient estimators. Conventional estimators such as two-stage least squares (2SLS), GMM, and Fuller as well as the unbiased estimator of Andrews and Armstrong (2017) are shown to be inefficient in the presence of heteroskedasticity and, under the availability of an efficient estimator of the reduced-form coefficients,
admit transformations into weakly efficient estimators by LAR. We give simulation studies illustrating how much the weakly efficient estimators can outperform their original estimators.


The rest of the chapter is organized as follows. Section 2.2 provides examples of weak identification in economics. Section 2.3 overviews main results drawing examples from linear IV models. Section 2.4 defines weak identification in semiparametric models, proves impossibility results on consistent and equivariant estimation, and introduces the notion of underlying regular parameters. Section 2.5 elaborates on desired properties of underlying regular parameters, namely sufficiency and minimality, and characterizes minimal sufficient underlying regular parameters. Section 2.6 considers estimation of weakly regular parameters through underlying regular parameters.
and derives local asymptotic Rao-Blackwellization, whence an efficiency concept for estimators of weakly regular parameters is defined. Section 2.7 discusses application of LAR to heteroskedastic linear IV models and carries out simulation studies showing how 2SLS, GMM, Fuller, and unbiased estimators can be improved upon under some conditions. Section 2.8 concludes. The Appendices contain all proofs and discussion of miscellaneous topics.

2.2 Weak Identification in Economics

We encounter the problem of weak identification in many different places in economics, including both micro- and macroeconomic contexts. This section describes prominent examples of weak identification. Among them, the linear IV model (Example 2.1) is the most important one; it branches into all types of special cases that capture various aspects of weak identification.

Example 2.1 (Linear IV). Consider the IV regression model:

\[
\begin{align*}
    y_i &= x'_i \beta + \varepsilon_i, & \mathbb{E}[\varepsilon_i \mid z_i] &= 0, \\
    x'_i &= z'_i \pi + v'_i, & \mathbb{E}[v_i \mid z_i] &= 0,
\end{align*}
\]

where \( y_i \) and \( \varepsilon_i \) are scalars, \( x_i, \beta, \) and \( v_i \) are \( d \times 1 \) vectors, \( z_i \) is a \( k \times 1 \) vector, \( \pi \) is a \( k \times d \) full column rank matrix, and \( k \geq d \). The first equation is called the second-stage equation and the second the first-stage equation; they are collectively called the structural equations. On the other hand, the reduced-form equations are obtained by substituting the first-stage equation into the second-stage:

\[
\begin{align*}
    y_i &= z'_i \pi \beta + u_i, & \mathbb{E}[u_i \mid z_i] &= 0, \\
    x'_i &= z'_i \pi + v'_i, & \mathbb{E}[v_i \mid z_i] &= 0.
\end{align*}
\]

The first equation of the reduced-form equations may sometimes be referred to as the second-stage equation when there is no confusion. The model is said to be just-identified if \( k = d \) and overidentified if \( k > d \). We are interested in the parameter \( \beta \),
which is called the *structural parameter*.

The conditional moment restrictions $E[u_i \mid z_i] = 0$ and $E[v_i \mid z_i] = 0$ are the key identifying assumptions; they are sometimes replaced by the weaker versions of the unconditional moment restrictions $E[z_i u_i] = 0$ and $E[z_i v'_i] = 0$. The theory developed in this chapter subsumes both cases while the demonstration of our theory in later sections will be based on the conditional moment restrictions. The second moments of all variables are assumed to be finite. The model is called *homo*skedastic if $E[u_i^2 \mid z_i]$, $E[u_i v_i \mid z_i]$, and $E[v_i v'_i \mid z_i]$ do not depend on $z_i$; otherwise, it is *hetero*skedastic.$^1$

Weak identification of the structural parameter $\beta$ occurs when the correlation between the endogenous regressors $x_i$ and the instruments $z_i$ is weak, which is captured by conventional asymptotic embedding (Staiger and Stock, 1997):

$$\pi_n = O\left(\frac{1}{\sqrt{n}}\right).$$

Another possibility of weak identification is the case in which $\pi_n$ does not vanish but approaches a rank deficient matrix at root-$n$:

$$\pi_n = \pi_0 + O\left(\frac{1}{\sqrt{n}}\right),$$

where $\pi_0$ is nonzero but not of full column rank (Andrews and Guggenberger, 2017, call this *joint weak identification*). Through suitable reparametrization and normalization, this model reduces to one where $\pi_0$ is a diagonal matrix whose upper $\ell \times \ell$ submatrix equals identity for some $\ell < d$ and all other elements zero (Appendix 2.B).

In the main text, we further assume that $\pi_0$ is a zero matrix as in Staiger and Stock (1997). Let

$$\pi_n = \frac{\hat{\pi}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$$

for a $k \times d$ matrix $\hat{\pi}$ that is of full rank.$^2$ This embedding induces the following

---

$^1$With unconditional moment restrictions, homoskedasticity means the following weaker version: $E[u_i^2 z_i z_i']$, $E[u_i v_i \otimes z_i z_i']$, and $E[v_i v'_i \otimes z_i z_i']$ are proportional to $E[z_i z_i']$ (up to the Kronecker structure).

$^2$In fact, it is not even necessary that $\pi_n$ approach zero no faster than root-$n$, in which case it will induce asymptotic partial identification; see Appendix 2.C for further discussion.
asymptotics to the second-stage coefficients,
\[\pi_n \beta_n = \frac{\hat{\pi} \beta}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)\]
for a \(d \times 1\) vector \(\beta\). When the model is weakly identified, the 2SLS estimator \(\hat{\beta}\) is no longer consistent for \(\beta\). Staiger and Stock (1997) show that the 2SLS converges in distribution to a Cauchy-like distribution that depends on the local parameter. It is important to understand that \(\hat{\beta}\) itself is \(O_P(1)\), not the inflated version \(\sqrt{n}(\hat{\beta} - \beta)\).

**Example 2.2** (Nonlinear GMM). Many structural models in economics identify parameters of interest in the form of a nonlinear moment equation. In particular, the parameter of interest \(\beta \in \mathbb{R}^d\) is identified as a unique solution to
\[E[Z_i h(X_i; \beta)] = m(\beta) = 0\]
for some vectors of random variables \(X_i\) and \(Z_i\) and a known function \(h\) indexed by \(\beta\). Weak identification of \(\beta\) happens when the moment function \(m\) converges uniformly to a function that has multiple zeros at rate \(\sqrt{n}\). Under weak identification, the nonlinear GMM estimator \(\hat{\beta}\) is not consistent and converges in distribution to a nonstandard nonpivotal distribution (Stock and Wright, 2000).

**Example 2.3** (Extremum estimation). Another popular specification that arises from structural models is the extremum estimation model. The parameter of interest \(\beta \in \mathbb{R}^d\) is identified as the minimizer of an unknown function of which we have an observable random estimator; the minimizer of the random function yields the estimator of \(\beta\). This model is closely related to Example 2.2 and some models (including linear IV models) admit both representations as GMM models and extremum estimation models. Weak identification of \(\beta\) occurs when the objective function flattens out (partially or fully) as the sample size tends to infinity. Andrews and Cheng (2012) consider cases when one of the parameters parameterizes the identification strength of other parameters; Han and McCloskey (2017) offer a way to reparameterize some extremum estimation models into the framework of Andrews and Cheng (2012) when
the source of identification failure is known; and Cox (2017) considers models that are doubly parameterized by structural and reduced-form parameters where the reduced-form parameters are always identified.

Example 2.4 (Differentiated product demand estimation). Endogeneity resulting from the simultaneous determination of prices and quantities poses a problem in demand estimation in industrial organization. In a situation in which one observes characteristics of many markets (but not individual purchasing behaviors), this endogeneity is often solved by using characteristics of other products as instruments for the endogenous prices (Berry et al., 1995). To invoke asymptotic approximation, the limit is often considered in the number of products tending to infinity, the so-called "large market asymptotics."

Armstrong (2016) shows, however, that the strength of these instruments is sensitive to how many products there are; in particular, when the number of products diverges along with the number of markets, the demand parameters may exhibit behaviors of strong identification, weak identification, or identification failure, depending on the relative rate of growth of the numbers of products and markets. A distinct feature of this example is that weak or non-identification asymptotics arises as a consequence of an equilibrium outcome, rather than as a purely statistical consideration of approximation.

Example 2.5 (Limited information macroeconomic models). Modern DSGE models contain three key equations: a Phillips curve, an Euler equation, and a monetary policy rule. When we estimate parameters of a DSGE model, we often take one of the equations individually and estimate via GMM with a particular choice of instruments. As this makes use of only a part of the full structure of the DSGE model, it is called the limited information approach and allows one to conduct estimation with a minimal set of assumptions about the nature of macroeconomic dynamics.

Increasing attention is given to the fact that these instruments have the tendency to be weak. To name a few examples, Dufour et al. (2006) and Nason and Smith (2008) analyze the case of potential weak identification in the context of New Keynesian

Weak identification is thus widely observed in economics. There are also other instances of “weak identification” that actually entail partial identification in the limit. Our main focus in this chapter is on “pure” weak identification; see Appendix 2.C for more discussions on this phenomenon.

**Example 2.6 (Partially weakly identified models).** It has been observed that weak identification and partial identification in linear IV models arise from similar sources (Poskitt and Skeels, 2013). In the context of weak instruments, if the first-stage coefficients approach degeneracy too quickly, then even if the structural parameter $\beta$ is identified at each $n$, it suffers partial identification in the corresponding local expansion (Appendix 2.C). In extremum estimation models, Cox (2017) uses a higher-order Taylor expansion to characterize the asymptotic behavior of the extremum estimator under the asymptotic embedding he calls “super-weak sequences of parameters” which cause similar phenomena. Another example of partial weak identification appears in DSGE models.

DSGE models can be estimated in their entirety by maximum likelihood (or the method of simulated moments). While this relies on a strong assumption that the full model is correctly specified, it is considered more efficient than the limited information approach if it is correct, hence called the *full information approach*. Not surprisingly, if the parameters are close to the point of identification failure, this full model is, again, subject to weak identification. Iskrev (2008) studies identification conditions in this context. Ruge-Murcia (2007) and Canova and Sala (2009) give a summary and comparison of weak identification in limited and full information approaches. Andrews and Mikusheva (2014, 2015) consider the maximum likelihood estimation and inference of DSGE models. Andrews and Mikusheva (2016a) consider minimum distance inference in general models containing DSGE models; they note in their supplementary material that a simple DSGE model approaches a limit at which only four out of six parameters are identified.
In Appendix 2.C we show that their DSGE model accompanies partial identification in the asymptotic embedding of weak identification they consider.

2.3 Overview of Main Results using Linear IV Models

This section illustrates our main results in relation to linear IV models (Example 2.1). We focus on intuitive verbal explanation and leave most of the technical details to later sections.

Let $\mathcal{P}$ be a set of distributions on $(x, y, z)$ such that they have second moments and there exists $(\pi, \beta) \in \mathbb{R}^{k \times d} \times \mathbb{R}^d$ such that $E[y - x' \beta | z] = 0$ and $E[x' - z' \pi | z] = 0$. This $\mathcal{P}$ contains all probability distributions that we deem possible, and hence is called the model. A key idea of semiparametric theory is to regard a semiparametric model as a collection of parametric models $\{Q_t\}, Q_t \in \mathcal{P}$, parameterized by the “sample size” $t = 1/\sqrt{n}$ (Section 2.4). Each parametric model, called a parametric submodel or a path, is assumed to be well-behaved in the sense that it admits a score and is locally asymptotically normal at a fixed point $P \in \mathcal{P}$, around which we approximate our unknown population.

In linear IV models, a parametric submodel corresponds to a particular asymptotic embedding, such as $\pi_n = \pi_0 + \hat{\pi}/\sqrt{n}$; the semiparametric nature of this approach comes from the fact that other (infinite-dimensional) components of the model, such as distributions of errors, are left unspecified. Thus, in a semiparametric model, we have a set $\hat{\mathcal{P}}_P$ of scores induced by each of the submodels; this is set called the tangent space.

Note that the moment conditions in the linear IV model can be written as $E[y - z' \pi \beta | z] = 0$ and $E[x' - z' \pi | z] = 0$. Thus, if $\pi$ that satisfies the moment conditions is degenerate, then there is more than one $\beta$ that satisfies the moment conditions; in other words, the structural parameter $\beta$ is not identified at such $P \in \mathcal{P}$. Let $\mathcal{P}_\beta \subset \mathcal{P}$ be the subset of distributions on $(x, y, z)$ for which $\beta$ is identified, viz., $\pi$ is
nondegenerate.

Weak identification of $\beta$ is known to arise when the population is close to a point of identification failure. Correspondingly, the key to modeling weak identification in semiparametric models is to consider parametric submodels $\{Q_t\}, \ Q_t \in \mathcal{P}_\beta$, that asymptote to some point of identification failure $\ P \notin \mathcal{P}_\beta$. As in classical theory, this limiting distribution $P$ is considered not known to us but held fixed when developing the local theory around it. Since parametric submodels cannot take values outside of $\mathcal{P}_\beta$ (although they can asymptote to a point outside of $\mathcal{P}_\beta$), the tangent set we consider becomes a subset $\mathcal{P}_{P,\beta}$ of $\mathcal{P}_P$, which can be shown to be a (usually nonconvex) cone.

Our first observation is that, along such a parametric submodel $Q_t$, the parameter $\beta(Q_t)$ asymptotes to a function of the score $\beta_P(g), \ g \in \mathcal{P}_{P,\beta}$, that is continuous and homogeneous of degree zero (Section 2.4.1); we call such a parameter weakly regular. This puts a severe restriction on what can be done in regard to estimation of a weakly regular parameter; many, though not all, of impossibility results known in linear IV models are shown to generalize to arbitrary semiparametric models. In Theorem 2.2, we prove that no consistent estimator of $\beta$ exists, so that any reasonable estimator $\hat{\beta}$ will remain random in the limit as the sample size tends to infinity. Also, there is no equivariant estimator, meaning that the asymptotic distribution of $\hat{\beta} - \beta$ necessarily changes as we consider different parametric submodels. As we often base our “robust” inference procedures on quantities whose asymptotic distributions do not depend on particular asymptotic embeddings, this result implies that we may not base our inference—even confidence intervals—on virtually any estimator of a weakly regular parameter. This roughly explains why most of the known “robust” inference procedures in linear IV models rely on estimators of the reduced-form coefficients, not of the structural parameter itself. This impossibility also yields an unfortunate implication that the classical efficiency concepts, namely the convolution and minimax theorems, are inapplicable, at least in their direct forms, to estimators of weakly regular parameters.

Given the impossibility results, we seek ways to formalize the notion of “reduced-
form coefficients” in general semiparametric models (Section 2.4.3). In linear IV models, it is known that the reduced-form coefficients \((\pi, \pi\beta)\) are always strongly identified regardless of the identification of \(\beta\) and that they completely characterize \(\beta\) by their “ratio” \(\beta = (\pi' \pi)^{-1} \pi' (\pi\beta)\); as such, they serve a key role in the derivation of asymptotic distributions of various estimators of \(\beta\) as well as in most “robust” inference procedures. Thus, if we could find such well-behaved statistics in other models, it would become a powerful toolkit to the development of valid statistical procedures in those models. We first formalize the notion of reduced-form coefficients in general semiparametric models by the name underlying regular parameters, and then show that for any weakly regular parameter there exists an underlying regular parameter. The first property of the reduced-form coefficients—that they are strongly identified—can be captured in semiparametric models as being regular at the limiting distribution \(P\). A regular parameter is differentiable (in the sense defined below) and its derivative is called the local parameter. In linear IV models under the Staiger-Stock asymptotics \(\pi_n = \hat{\pi}/\sqrt{n}\), \((\hat{\pi}, \hat{\pi}\beta)\) is the local parameter of \((\pi, \pi\beta)\). The second important property of the reduced-form coefficients is that it fully characterizes the weakly regular parameter; this is equivalent to requiring that there exists a map from the local parameter of the underlying regular parameter to the weakly regular parameter. Continuing on the linear IV model, this map is given by \((\hat{\pi}, \hat{\pi}\beta) \mapsto (\hat{\pi}' \hat{\pi})^{-1} \hat{\pi}' (\hat{\pi}\beta) = \beta\). If a linear IV model is overidentified, throwing away extra instruments does not impair identification of \(\beta\). Thus, there are many combinations of reduced-form coefficients that “fully characterize” \(\beta\). In this sense, underlying regular parameters are not unique.

However, it is not unreasonable to expect that the “best” underlying regular parameter would be the one that contains all available reduced-form coefficients (hence all instruments). But then can we clarify in what sense it is good? If so, when and how can we identify the “best” underlying regular parameter in a general semiparametric model? Using information theory of scores, we first identify two desirable properties of underlying regular parameters, namely sufficiency and minimality (Section 2.5); then, we show that a minimal sufficient underlying regular parameter (the “best”
one) exists for every weakly regular parameter. In particular, sufficiency means that the underlying parameter contains all available information about $\beta$; in the context of linear IV models, the set of all reduced-form coefficients is shown to constitute a sufficient underlying parameter. Sufficiency is desirable in that rewriting an estimation or inference problem on a weakly regular parameter to that on a sufficient underlying parameter would not miss any important information about $\beta$. Minimality means that the underlying parameter contains no irrelevant information about $\beta$; in linear IV models, any parameter other than the reduced-form coefficients is shown to be irrelevant to the structural coefficient $\beta$, hence should be excluded from any minimal underlying parameter. Minimality is desirable in that estimation or inference on a minimal underlying parameter will not involve unnecessary estimation of (a part of) a nuisance parameter. Thus, the set of all and only reduced-form coefficients gives the minimal sufficient underlying parameter for the structural coefficient $\beta$.

Next, we consider estimation of a weakly regular parameter as a problem of estimation of a minimal sufficient underlying parameter and its transformation (Section 2.6). In linear IV models, we thus focus on estimators of $\beta$ that can be interpreted asymptotically as functions of regular estimators of reduced-form coefficients $(\pi, \pi\beta)$; we call such estimators regular. Regularity of an estimator $\hat{\beta}$ allows us to think about efficiency of $\hat{\beta}$ from the perspective of efficiency of corresponding $(\hat{\pi}, \hat{\pi}\beta)$. In particular, if there is noise in estimators of the underlying parameter, their transformations should also contain noise. Considering that an efficient estimator of a regular parameter is “asymptotically sufficient,” we may invoke the Rao-Blackwell theorem to their asymptotic representation in order to improve efficiency of estimators of a weakly regular parameter. Technically, the “improvement” is in terms of any continuous convex loss, and the improved estimator is more concentrated toward the same mean as the original estimator. Thence, we define an estimator that does not admit such improvement as weakly efficient.

Many popular estimators in linear IV models, including 2SLS, two-step GMM, Fuller (Fuller, 1977), and unbiased (Andrews and Armstrong, 2017) estimators, can be represented asymptotically as functions of OLS estimators of the reduced-form
coefficients; thus, they are all regular. For example, consider 2SLS

\[ \tilde{\beta}_{2SLS} = (\tilde{\pi}'(Z'Z)\tilde{\pi})^{-1}\tilde{\pi}'(Z'Z)\tilde{\pi}\beta, \]

where \( \tilde{\pi} \) and \( \tilde{\pi}\beta \) are OLS estimators of the reduced-form coefficients. If reduced-form errors exhibit heteroskedasticity, however, OLS is not efficient in its own right. In particular, if an efficient estimator \( (\hat{\pi}, \hat{\pi}\beta) \), say feasible GLS, is available, then the OLS estimator is asymptotically equal to feasible GLS plus independent noise, that is, \( \sqrt{n}(\hat{\pi}, \hat{\pi}\beta) = \sqrt{n}(\tilde{\pi}, \tilde{\pi}\beta) + (U_\pi, U_{\pi\beta}) + o_P(1) \). Then, we may improve 2SLS by taking conditional expectation with respect to the noise:

\[ \mathbb{E}[(\sqrt{n}\hat{\pi} + U_\pi)'(Z'Z)\sqrt{n}\hat{\pi} + U_\pi)^{-1}(\sqrt{n}\hat{\pi} + U_\pi)'(Z'Z)(\sqrt{n}\hat{\pi}\beta + U_{\pi\beta}) \mid \hat{\pi}, \hat{\pi}\beta]. \]

Since this estimator does not admit further improvement along this line, it is by itself weakly efficient. Figure 2-1a shows an example of asymptotic distribution of a 2SLS estimator under heteroskedasticity; Figure 2-1b shows its improvement under availability of feasible GLS. Note that the asymptotic distribution of either estimator is not consistently estimable, implying that one cannot construct a valid confidence
interval from either distribution. Therefore, the fact that the improvement has “less chance” of actually hitting the true value does not seem to pose a problem, since in any event we need to use some other alternative procedure to construct a valid confidence interval. Section 2.7 also gives improvements of two-step GMM, Fuller, and unbiased estimators.

2.4 Weak Identification in Semiparametric Models

In this section, we define the weakly regular parameter that forms the basis of our theory. As a minor caveat, whenever we talk about measurability, it is understood in the sense that is appropriate in the context (for example, g below needs to be “P-measurable”).

Suppose we observe i.i.d. random variables $X_1, \ldots, X_n$ from the sample space $(\mathcal{X}, \mathcal{A})$. The set of possible distributions of each $X$ is denoted by $\mathcal{P}$ and is called the model. To obtain fruitful asymptotics around a distribution $P \in \mathcal{P}$, we consider a path of distributions $Q_t \in \mathcal{P}$ indexed by a real number $t \in (0, 1]$ that is differentiable in quadratic mean (DQM) at $P$, that is, there exists a measurable function $g : \mathcal{X} \to \mathbb{R}$ such that\(^4\)

$$\int_{\mathcal{X}} \left[ \frac{dQ_t^{1/2}}{t} - dP^{1/2} - \frac{1}{2} g dP^{1/2} \right]^2 \to 0 \quad \text{as} \quad t \to 0.$$ 

This convergence is denoted by $Q_t \to^{\text{DQM}} P$, and we call $g$ the (model) score induced by the path $\{Q_t\}^{\mathcal{P}}$. The idea behind asymptotic approximation theory is that the path of “alternatives” $\{Q_t\}$ that approaches $P$ at the same rate as the path of “samples” $\{\hat{P}_n\}$ is not deterministically discriminable in the limit and hence yields an approximation that reflects finite sample uncertainty; therefore, in many examples, it is suggestive to understand $t = 1/\sqrt{n}$ and in a minor abuse of notation to denote $Q_{1/\sqrt{n}}$ by $Q_n$.

\(^3\)A path is also called a (parametric) submodel.

\(^4\)The integral is understood with respect to some $\sigma$-finite measure dominating $P$ and $Q_t$, and $dP$ and $dQ_t$ are the Radon-Nikodym derivatives of $P$ and $Q_t$ with respect to it.

\(^5\)Throughout the chapter, dependence of $g$ on $\{Q_t\}$ will be implied by the context.
We often do not consider every possible path in $\mathcal{P}$; let $\mathcal{P}_P$ denote the set of paths we consider that tend to $P$ in DQM. Since there is little chance of misunderstanding, we hereafter denote $\{Q_t\}$ simply by $Q_t$, for example, $Q_t \in \mathcal{P}_P$; therefore, $Q_t$ can refer to the entire path $\{Q_t\}$ or an element $Q_t$ of the path for a specific $t$, depending on the context. The set $\hat{\mathcal{P}}_P$ of scores $g$ induced by the paths in $\mathcal{P}_P$ is called the tangent set at $P$. It is clear from the definition of scores that $\hat{\mathcal{P}}_P$ is a subset of $L_2(P)$. Depending on the structure of $\mathcal{P}_P$, the tangent set might be a linear space, a cone, or just a set without much structure; we assume that we can always augment $\mathcal{P}_P$ linearly so that the induced tangent set will be linear. For this reason, we call the tangent set the tangent space. The tangent space can be considered the local approximation of the model by a linear vector space. Finally, a parameter $\psi : \mathcal{P} \to \mathbb{D}$ is defined as a map from the model $\mathcal{P}$ to a Banach space $\mathbb{D}$.

If the parameter $\psi : \mathcal{P} \to \mathbb{D}$ is differentiable in a suitable sense, by the chain rule we may approximate the change in the parameter along any path by a linear map from the tangent space $\hat{\mathcal{P}}_P$ to the parameter space $\mathbb{D}$. Any infinitesimal perturbation of distribution $P$ then leads to a linear perturbation of parameter $\psi$. Such a parameter is known to behave well and is said to be regular. This case is well studied in the literature (Groeneboom and Wellner, 1992; Bickel et al., 1993; Van der Vaart 1988, 1998; Kosorok 2008), and we will make good use of it in the study of weak identification. The appropriate notion of differentiability is given as follows.

**Definition** (Regular parameter). A parameter $\psi : \mathcal{P} \to \mathbb{D}$ is regular (or differentiable) at $P$ relative to $\mathcal{P}_P$ if there exists a continuous linear map $\hat{\psi}_P : \hat{\mathcal{P}}_P \to \mathbb{D}$ such that

$$\frac{\psi(Q_t) - \psi(P)}{t} \to \hat{\psi}_P g \quad \text{for every} \quad Q_t \in \mathcal{P}_P.$$ 

The derivative map $\hat{\psi}_P$ is called the local parameter of $\psi$. The adjoint map $\hat{\psi}_P^* : \mathbb{D}^* \to \hat{\mathcal{P}}_P$ is called the efficient influence map of $\psi$, where $\mathbb{D}^*$ is the dual space of $\mathbb{D}$ and $\hat{\mathcal{P}}_P$.

---

7 In this sense, $\mathcal{P}$ is the set of equivalence classes of scores, to be precise.
8 A subset $X$ of a linear space is called a cone if $x \in X$ implies $ax \in X$ for every $a > 0$.
9 If $\mathcal{P}_P$ is the set of all possible paths in $\mathcal{P}$, then regularity of $\psi$ is equivalent to Hadamard differentiability of $\psi$.
the completion of $\hat{P}_P$.\footnote{The function $\tilde{\psi}_P : \mathcal{X} \rightarrow \mathbb{D}$ such that $\tilde{\psi}_P^* \delta^* = \delta^* \tilde{\psi}_P$ for every $\delta^* \in \mathbb{D}^*$ is called the \textit{efficient influence function} of $\psi$ \cite{Bickel1993} Section 5.2. The qualifier \textit{efficient} is justified in the context of the convolution theorem as remarked in Section 2.6 \cite{Kosorok2008} Section 18.1) also gives alternative definitions (interpretations) of efficient influence functions in the context of functional parameters.}

Remark. In the classical context, the tangent set “represents” the set of paths \cite[Section 18.1]{Kosorok2008}, so regularity (differentiability) is often defined “relative to the tangent set” \cite[Chapter 25]{Van der Vaart1998}; \cite[Section 18.1]{Kosorok2008}. In the context of weak identification, however, the corresponding tangent set does not represent the set of paths (see the next section); therefore, we keep the original wording “relative to the set of paths” from \cite{Van der Vaart1991b}. The word “regular” is taken from \cite{Van der VaartWellner1996} Chapter 3.11.

2.4.1 Weakly Regular Parameters

Now we define a weakly identified parameter. When we talk about weak identification, often have we in mind a situation in which the sequence of distributions converges to a point of identification failure. In this respect, the weakly identified parameter may only be defined on a subset $\mathcal{P}_\beta$ of $\mathcal{P}$ where the difference $\mathcal{P} \setminus \mathcal{P}_\beta$ represents all points of identification failure. Accordingly, the path cannot fall outside of the submodel $\mathcal{P}_\beta$, and the tangent set must be restricted in order for the weakly identified parameter to be well defined. Computing the corresponding tangent subset requires care, however, since too rapid an approach to the point of identification failure must be avoided.

Somewhat counterintuitively, the tangent set pertinent to $\mathcal{P}_\beta$ cannot be defined as the set of all scores induced by the paths taking values in $\mathcal{P}_\beta$; it is the set of all scores that are \textit{not} induced by the paths \textit{not} taking values in $\mathcal{P}_\beta$.

Definition (Pertinent tangent cone). The tangent set $\hat{\mathcal{P}}_{P,\beta} \subset \hat{\mathcal{P}}_P$ \textit{pertinent to} the submodel $\mathcal{P}_\beta$ at $P \in \mathcal{P}$, possibly $P \in \mathcal{P} \setminus \mathcal{P}_\beta$, is the set of scores $g \in \hat{\mathcal{P}}_P$ such that there does not exist a path in $\mathcal{P}_P$ that takes values in $\mathcal{P} \setminus \mathcal{P}_\beta$ and induces $g$. Define $\mathcal{P}_{P,\beta}$ to be the set of paths in $\mathcal{P}_P$ that take values in $\mathcal{P}_\beta$ and induce scores in $\hat{\mathcal{P}}_{P,\beta}$.

The following example illustrates why we need this circuitous definition.
Example 2.1 (Linear IV, continuing from p. 125). Consider a simple parametric linear IV model with \( d = k = 1 \) and \((u, v) \sim N(0, I_2)\). So \( \mathcal{P} \) is the set of distributions of \((x, y, z)\) such that \((y - z\pi\beta, x - z\pi) \sim N(0, I_2)\) for some \( \pi, \beta \in \mathbb{R} \) and \( \mathcal{P}_\beta \) is the subset of \( \mathcal{P} \) such that \( \pi \neq 0 \). Consider the asymptotic embedding \( \pi_n = \hat{\pi}/n \) instead of \( \pi_n = \hat{\pi}/\sqrt{n} \). The path of this embedding is

\[
dQ_n = \frac{1}{2\pi} \exp\left( -\frac{(y - z\hat{\pi}\beta/n)^2 + (x - z\hat{\pi}/n)^2}{2} \right),
\]

which converges in DQM to the point of identification failure

\[
dP = \frac{1}{2\pi} \exp\left( -\frac{y^2 + x^2}{2} \right)
\]

with the score

\[
\sqrt{n} \frac{dQ_n - dP}{dP} \rightarrow 0.
\]

Although \( Q_n \) takes values only on \( \mathcal{P}_\beta \), its score can also be induced by the path \( \tilde{Q}_n \equiv P \), which should be excluded. Therefore, \( \hat{\mathcal{P}}_{P,\beta} \) cannot be taken as the set of all scores induced by paths in \( \mathcal{P}_\beta \), but as the set of scores not induced by paths not in \( \mathcal{P}_\beta \).

From the observation that \( P \) is not in \( \mathcal{P}_\beta \), we see that \( \hat{\mathcal{P}}_{P,\beta} \) is only a cone.

Lemma 2.1. \( \hat{\mathcal{P}}_{P,\beta} \) and \( \hat{\mathcal{P}}_P \setminus \hat{\mathcal{P}}_{P,\beta} \) are cones.

Remark. In classical asymptotic theory, the limit distribution \( P \) is often regarded as the “null hypothesis” and the path \( Q_t \) as a drifting sequence of “alternatives.” When it comes to weak identification, in contrast, both the null and alternatives reside as paths in \( \mathcal{P}_{P,\beta} \); \( P \) is merely a point of reference of identification failure.

If the set of paths \( \mathcal{P}_P \) is much richer than \( \mathcal{P}_{P,\beta} \) in a way that \( \text{Span} \hat{\mathcal{P}}_{P,\beta} \) is a strict subset of \( \hat{\mathcal{P}}_P \), then there exists a superfluously rich side of the model on which \( \beta \) is not even defined. Since it is meaningless to consider such parts of the model when one’s focus is on the parameter \( \beta \), we assume innocuously that \( \text{Span} \hat{\mathcal{P}}_{P,\beta} = \hat{\mathcal{P}}_P \).\textsuperscript{11}

\textsuperscript{11}Later on we define the underlying regular parameter on the whole of \( \mathcal{P} \), so it is actually harmful to require that that parameter be regular on the unconsidered realm of the model.
Now we define the weakly identified parameter under the name *weakly regular parameter*. We henceforth shun the use of the qualifier “weakly identified” since weak identification in the literature may not always exclude cases of in fact *no* identification (e.g., Moreira [2009], Andrews and Cheng [2013, 2014], Han and McCloskey [2017]). In this chapter, we assume that weakly regular parameters are identified at every fixed $n$ in that there exists a unique value of the parameter for any given distribution $Q_n$ belonging to $\mathcal{P}_\beta$. Moreover, we assume that the parameters remain identified in the limit in the sense that there exists a unique value of the parameter for each score $g$ in $\hat{\mathcal{P}}_{P,\beta}$. See Example 2.6 and Appendix 2.C for cases that entail partial identification in the limit. Let $\mathbb{B}$ be another Banach space on which a weakly regular parameter will be defined.

**Definition** (Zero-degree homogeneity). Let $\mathcal{X}$ be a linear space and $\mathcal{Y}$ a set. A map $f : A \to \mathcal{Y}$ defined on a cone $A$ of $\mathcal{X}$ is said to be *homogeneous of degree zero* if

$$f(a) = f(ba) \quad \text{for every} \quad b > 0, \quad a \in A.$$

**Definition** (Weakly regular parameter). A parameter $\beta : \mathcal{P}_\beta \to \mathbb{B}$ is *weakly regular* at $P \in \mathcal{P}$, possibly $P \in \mathcal{P} \setminus \mathcal{P}_\beta$, relative to $\mathcal{P}_{P,\beta}$ if there exists a continuous map $\beta_P : \hat{\mathcal{P}}_{P,\beta} \to \mathbb{B}$ that is homogeneous of degree zero such that

$$\beta(Q_t) \longrightarrow \beta_P(g) \quad \text{for every} \quad Q_t \in \mathcal{P}_{P,\beta}.$$

The idea behind the definition is that a weakly identified parameter is necessarily accompanied by the fundamental uncertainty of the model summarized by the score. This captures our observation that the estimator of a weakly identified parameter does not converge to the true value but retains some randomness in the limit. In this sense, asymptotics of weakly regular parameters is “global” in nature, and nonlinearity often observed in the literature (e.g., Cox [2017]) arises compellingly from nonlinearity of the map $\beta_P$, resulting in a nonstandard limit distribution. The definition clarifies the

---

12 Not to be confused with *weak regularity* of an estimator defined in [Van der Vaart (1988, Section 2.2)](https://doi.org/10.1017/CBO9780511802257) or [Bickel et al. (1993, Definition 5.2.6)](https://doi.org/10.1007/978-1-4612-4380-6).
extent of such randomness and nonlinearity that are inherent to the model (and hence may not be compensated for by the choice of estimation methods); see Theorem 2.2 for the consequential, fundamental impossibility results.

Remark. Since we regard the reference point of identification failure $P$ and the sets of paths $\mathcal{P}_P$ and $\mathcal{P}_{P,\beta}$ as fixed, we often mention (weak) regularity of a parameter without referring to the point of evaluation and the relative collection of paths.

Remark. Being a continuous map, a regular parameter is trivially weakly regular; that is, if $\psi: \mathcal{P} \to \mathbb{D}$ is regular, then $\psi(Q_t) \to \psi_P(g)$ where $\psi_P(g) \equiv \psi(P)$. Also, if $\beta$ is a nontrivial weakly regular parameter, i.e., $\beta_P$ is nonconstant, then $\beta_P$ cannot be linear since a linear function that is homogeneous of degree zero must be identically zero.

Remark. A weakly regular parameter is not only nondifferentiable at $P$; it is even discontinuous at $P$ (unless $\beta_P$ is trivially constant). This is in contrast to the literature on continuous but not (fully) differentiable parameters [Hirano and Porter 2012, Fang 2015, 2016; Fang and Santos 2015; Hong and Li 2017]. This discontinuity plays a key role in one of the impossibility results in Theorem 2.2; we exploit the “continuity” of asymptotic distributions implied by Le Cam’s third lemma.

Remark. Homogeneity of $\beta_P$ is a natural consequence of dependence on $g$. Since $\beta(Q_{kt})$ for fixed $k > 0$ converges to the same limit as $\beta(Q_t)$, we have $\beta_P(kg) = \beta_P(g)$.

Continuity of $\beta_P$ is required only on its domain $\hat{\mathcal{P}}_{P,\beta}$; it is not possible to extend $\beta_P$ continuously to the whole of $\hat{\mathcal{P}}_P$ unless $\beta_P$ is trivially constant.

Now we show how linear IV models fall into our framework. We construct paths that encapsulate the asymptotic embeddings discussed in Section 2.2 and show that the parameters can be written as continuous and homogeneous functions of the model scores.

Example 2.1 (Linear IV, continuing from p. 138). Let $\mathcal{P}_{uvw}$ be the set of probability distributions $P_{uvw}$ on $(u, v', z)$ with second moments such that $\mathbb{E}[u \mid z] = 0$, $\mathbb{E}[v \mid z] =$

\footnote{Fang and Santos (2015) observe a related fact that a directional derivative must be homogeneous of degree one.}
0, $\mathbb{E}[zz']$ is invertible, and $dP_{uvz}$ differentiable almost everywhere in $(u, v')$. The model $\mathcal{P}$ is the set of probability distributions $P$ on the observable elements $(x, y, z)$ such that

$$dP(x, y, z) = dP_{uvz}(y - z'\pi\beta, x' - z'\pi, z)$$

for some $P_{uvz} \in \mathcal{P}_{uvz}$, $\pi \in \mathbb{R}^{k \times d}$, $\beta \in \mathbb{R}^{d}$.

The distribution of $z$ is characterized by $P_{uvz}$; that of $y$ by $P_{uvz}$ and $\pi\beta$; that of $x$ by $P_{uvz}$ and $\pi$; thus, we have “parameterized” the semiparametric model $\mathcal{P}$ by three parameters $P_{uvz}$, $\pi$, and $\beta$. The submodel $\mathcal{P}_\beta$ is the subset of $\mathcal{P}$ of all distributions with $\det(\pi\pi') \neq 0$. For simplicity, assume $\pi(P) = 0$ at $P \in \mathcal{P} \setminus \mathcal{P}_\beta$. We consider a path $Q_t$ toward $P$ such that $[\pi(Q_t) - 0]/t$ converges to some element $\dot{\pi}$ in $\mathbb{R}^{k \times d}$. If $\det(\pi'\pi) = 0$, then there exists a path taking values in $\mathcal{P} \setminus \mathcal{P}_\beta$ that yields the same limit of $\pi$; this means that for every $Q_t \in \mathcal{P}_{P,\beta}$ we have $\det(\pi'\pi) \neq 0$. Such a path can be represented as

$$dQ_t(x, y, z) = dQ_{t,uvz}(y - z'(t\dot{\pi}_t\beta_t), x' - z'(t\dot{\pi}_t), z)$$

for some path $Q_{t,uvz}$ in $\mathcal{P}_{uvz}$, and $\dot{\pi}_t \rightarrow \dot{\pi}$ and $\beta_t \rightarrow \beta$. Being a probability distribution by itself, $Q_{t,uvz}$ has its own “model score” $g_{uvz}$. To see what it is like, note that the only essential restriction of $Q_{t,uvz}$ is $\int zudQ_{t,uvz} = 0$ and $\int zvdQ_{t,uvz} = 0$. Therefore,

$$0 = \frac{1}{t} \left( \int zudQ_{t,uvz} - \int zudP_{uvz} \right) \rightarrow \int zug_{uvz}dP_{uvz} = \mathbb{E}_P[zug_{uvz}].$$

Similarly, one sees that $\mathbb{E}_P[z'g_{uvz}] = 0$. Therefore, the set of scores of the parameter $P_{uvz}$ consists of all appropriate scores that satisfy these two restrictions.\footnote{Differentiability is not necessary as long as each one-dimensional parametric submodel is differentiable in quadratic mean (see, e.g., Pollard [1997] Van der Vaart [1998], Section 7.2). Here we assume this for illustration of derivation of scores. See also Van der Vaart (1988, Section 1.2 and Appendix A.2).}

Using this,
the model score for the path of interest $Q_t$ can be calculated as

$$
\frac{dQ_t - dP}{tdP} = \frac{dQ_{t,uvz}(y - z'(t\tilde{\tau}_t\beta_t), x' - z'(t\tilde{\tau}_t))}{tdP} - dP_{uz}(y - z'(t\tilde{\tau}_t\beta_t), x' - z'(t\tilde{\tau}_t)) \\
+ \frac{dP_{zu}(y - z'(t\tilde{\tau}_t\beta_t), x' - z'(t\tilde{\tau}_t))}{tdP} - dP_{uz}(y, x) \\
\rightarrow g = g_{uvz} - z'\tilde{\tau}_t \frac{dP_{uvz,u}}{dP} - z'\tilde{\tau}_t \frac{dP_{uvz,v}}{dP},
$$

where $P_{uvz,u}$ and $P_{uvz,v}$ represent the partial derivatives of $P_{uvz}$ with respect to $u$ and $v$. Observe that by integration by parts

$$
\mathbb{E}_P[zug] = -\int zuz'\tilde{\tau}_t dP_{uvz,u} - \int zuz'\tilde{\tau}_t dP_{uvz,v} = \int zz'\tilde{\tau}_t dP = \mathbb{E}_P[zz']\tilde{\tau}_t.
$$

Similarly, $\mathbb{E}_P[zv'g] = \mathbb{E}_P[zz']\tilde{\tau}_t$. Therefore, $\beta_t$ converges to

$$
\beta = (\mathbb{E}_P[zz']^{-1}\mathbb{E}_P[zv'g])^{-\gamma}(\mathbb{E}_P[zz']^{-1}\mathbb{E}_P[zug]) =: \beta_P(g),
$$

where $A^\rightarrow$ denotes the left inverse of $A$, that is, $A^\rightarrow := (A'A)^{-1}A'$. This map is continuous on $\hat{\mathcal{P}}_{P,\beta}$ and homogeneous of degree zero but nonlinear.

Thus, we have shown that $\beta$ defined on $\mathcal{P}_\beta$ converges to a continuous and homogeneous function $\beta_P$ of a score along every path in $\mathcal{P}_{P,\beta}$, meaning that $\beta$ is weakly regular at $P$ relative to $\mathcal{P}_{P,\beta}$. □

**Example 2.7** (Testing local hypotheses). For a regular parameter $\psi : \mathcal{P} \rightarrow \mathbb{D}$, consider the hypothesis

$$
H_0 : \psi(P) \in \mathbb{D}_0 \quad \text{vs} \quad H_1 : \psi(P) \in \mathbb{D}_1,
$$

where $\mathbb{D}_0 \cap \mathbb{D}_1 = \emptyset$ and $\mathbb{D}_0 \cup \mathbb{D}_1 \subset \mathbb{D}$. This induces a local testing problem at a boundary $P$ of the following form

$$
H_0 : \dot{\psi}Pg \in \mathbb{D}_{P,0} \quad \text{vs} \quad H_1 : \dot{\psi}Pg \in \mathbb{D}_{P,1}
$$

with $\mathbb{D}_{P,0} \cap \mathbb{D}_{P,1} = \emptyset$ and $\mathbb{D}_{P,0} \cup \mathbb{D}_{P,1} \subset \mathbb{D}$. These testing problems can be represented
by a weakly regular parameter $\beta : \mathcal{P}_\beta \rightarrow [0, 1]$ such that

$$
\beta(P) := \begin{cases}
1 & \psi(P) \in \mathbb{D}_0, \\
0 & \psi(P) \in \mathbb{D}_1,
\end{cases}
$$

where $\mathcal{P}_\beta := \psi^{-1}(\mathbb{D}_0 \cup \mathbb{D}_1)$, and its corresponding limit

$$
\beta_P(g) := \begin{cases}
1 & \dot{\psi}_P g \in \mathbb{D}_{P,0}, \\
0 & \dot{\psi}_P g \in \mathbb{D}_{P,1}.
\end{cases}
$$

If $\mathbb{D}_{P,0}$ and $\mathbb{D}_{P,1}$ are cones and their boundary is excluded, $\beta$ can be considered weakly regular.

\[\square\]

2.4.2 Fundamental Impossibility

The utility of our theoretical formalism can be readily harvested in the following theorem. It gives a formal proof to the conventional wisdom that a “weakly identified” parameter cannot be estimated consistently or pivotally—but not as a characteristic of a specific estimation method—as a direct consequence of the characteristic of the model (see, inter alia, Phillips 1984, 1989; Staiger and Stock 1997; Stock and Wright 2000; Guggenberger and Smith 2005; Andrews and Cheng 2012; Cox 2017). This result can also be viewed as a generalized proof of nonexistence of a consistent test conjectured by Hahn et al. (2011). Distinct but related are the impossibility results by Dufour (1997) and Hirano and Porter (2015); their setup is a generalization of the weak linear IV structure whereas our setup is a generalization of the weak identification phenomena. Indeed, Dufour (1997) shows nonexistence of bounded confidence sets (which is “stronger” than nonexistence of consistent estimators) while there exist weakly regular parameters that admit bounded confidence sets (Example 2.7);

\[\text{\footnotesize \textsuperscript{16}}\text{Consistent estimation may be possible in linear IV models if the number of weak instruments tends to infinity and some other conditions are met (Chao and Swanson 2005; Newey and Windmeijer 2009). In this case, the structural parameter is not weakly regular.}\]

\[\text{\footnotesize \textsuperscript{17}}\text{Their setup can be translated into ours by taking $\mathcal{B}$ to be the product space for two estimators compared in the Hausman test, observing that a regular parameter is trivially weakly regular.}\]
Hirano and Porter (2015) show the impossibility of unbiased estimation while there exist weakly regular parameters that admit unbiased estimation (Andrews and Armstrong, 2017).

Theorem 2.2 (Impossibility of consistent and equivariant estimation). There is no consistent sequence of estimators of a nontrivial weakly regular parameter; there is no consistent sequence of nontrivial tests of a nontrivial weakly regular parameter; there is no equivariant-in-law sequence of estimators of a nontrivial weakly regular parameter with a separable limit law.

Remark. The assumption of separability of the limit law is without “great loss of generality;” we treat it as general impossibility of equivariant estimation in the main text. See discussions of Van der Vaart and Wellner (1996, Theorem 1.3.10).

Impossibility of equivariant estimation implies that the asymptotic distribution of any estimator of a weakly regular parameter, when centered at the true value, is nonpivotal and not consistently estimable. However, it does not preclude the possibility that there exist test statistics whose distributions are pivotal or consistently estimable (Kleibergen, 2002, 2005). In fact, almost any reasonable inference procedure would be based on statistics whose asymptotic distributions are known or at least estimable; hence, the problem of estimation and the problem of inference bear quite distinct aspects when it comes to weakly regular parameters. This partly explains the specialty of current literature on inference problems pertaining to weak identification.

2.4.3 Underlying Regular Parameters

The idea on analyzing the weak regularity of a parameter is that in many cases there exists another parameter that is regular and whose local parameter controls the limit behavior of the weakly regular parameter. In the literature, such a parameter is...
known as the “reduced-form parameter” and is considerably utilized in various robust inference procedures under weak identification (inter alia, [Magnusson and Mavroeidis 2010, Mavroeidis 2010, Guerron-Quintana et al. 2013, Andrews and Mikusheva 2016a, Armstrong 2016, Cox 2017]). Then, the weakly regular parameter acts by itself as (a transformation of) the local parameter of some “underlying” regular parameter; in other words, it is sufficient to know the value of (the local parameter of) the underlying regular parameter in order to infer the value of the weakly regular parameter in the local expansion around the point of identification failure. We now formalize this idea, starting with the following definition.

**Definition (Underlying regular parameter).** Let $\beta : \mathcal{P}_\beta \rightarrow \mathbb{B}$ be weakly regular at $P \in \mathcal{P}$ relative to $\mathcal{P}_{P,\beta}$. The parameter $\psi : \mathcal{P} \rightarrow \mathbb{D}$ is an underlying (regular) parameter for $\beta$ at $P$ relative to $\mathcal{P}_P$ if it is regular at $P$ relative to $\mathcal{P}_P$ and there exists a continuous map $\beta_{P,\psi} : \mathbb{D}_\beta \rightarrow \mathbb{B}$ that is homogeneous of degree zero such that

$$
\beta(Q_t) \rightarrow \beta_{P,\psi}(\dot{\psi}_P g) \quad \text{for every} \quad Q_t \in \mathcal{P}_{P,\beta},
$$

where $\mathbb{D}_\beta$ is the subset of $\mathbb{D}$ on which the local parameter of $\psi$ takes values, that is, $\{\delta \in \mathbb{D} : \delta = \dot{\psi}_P g \text{ for some } g \in \dot{\mathcal{P}}_{P,\beta}\}$.

**Remark.** There exists a map $\beta_\psi : \mathbb{D} \rightarrow \mathbb{B}$ such that $\beta(Q_t) = \beta_\psi(\psi(Q_t)) + o(1)$ and thus $\beta_\psi$ admits approximation at $\psi(P)$ by a homogeneous function $\beta_\psi$; set, e.g., $\beta_\psi(\delta) = \beta_{P,\psi}(\delta - \psi(P))$. In many applications, moreover, there exists an exact direct representation $\beta(Q_t) = \beta_\psi(\psi(Q_t))$ with some function $\beta_\psi : \mathbb{D} \rightarrow \mathbb{B}$ that is “locally homogeneous” at $\psi(P)$. For instance, we show below in Example 2.1 that the structural parameter $\beta$ in linear IV models has a direct representation by the underlying regular parameter taken to be the reduced-form coefficients.

**Remark.** An underlying parameter is regular and hence susceptible to various estimation and inference techniques developed in statistics ([Bickel et al., 1993, Van der Vaart 1998, Kosorok 2008]).

\[^{19}\text{On the other hand, the weakly regular parameter is often referred to as the “structural parameter.”}\]
Remark. In the context of extremum estimation, Cox (2017) defines “reduced-form parameters” as functions of “structural parameters.” We take the opposite route: a weakly regular parameter approaches a function of (the local parameter of) an underlying regular parameter.

This definition requires that knowing the local parameter of the underlying regular parameter is enough to recover the value of the weakly regular parameter; the reduction of information from knowing \( g \) to knowing \( \hat{g}_P g \) does not impair the ability to discern \( \beta \) in the limit. With this definition, several questions arise: Does an underlying parameter always exist? How do we find an underlying regular parameter? How can we check whether a particular parameter is an underlying regular parameter? Which underlying regular parameter is better than another? We answer the first two questions in the remainder of this section and the rest in the next section.

The first question turns out to be straightforward. If one takes the root likelihood ratio \( Q \mapsto dQ^{1/2}/dP^{1/2} \) to be a parameter, one can trivially claim that there always exists an underlying regular parameter for any weakly regular parameter. However, whether there exists an underlying regular parameter that admits root-\( n \) consistent estimation is a different matter. For this, we need to search for a good underlying parameter in each model separately.

Lemma 2.3 (Existence of underlying regular parameter). Let \( \beta: \mathcal{P}_\beta \to \mathbb{B} \) be weakly regular. Then, there exist a Banach space \( \mathbb{D} \) and an underlying regular parameter \( \psi: \mathcal{P} \to \mathbb{D} \) for \( \beta \).

Now we look into underlying regular parameters in examples. We see that the natural parameters that appear in each example above constitute underlying regular parameters; however, the linear IV case contains other interesting and equally natural underlying parameters that deserve attention.

Example 2.1 (Linear IV, continuing from p. 140). Define \( \psi := (\psi_1, \text{vec}(\psi_2)) := (\pi \beta, \text{vec}(\pi)) \) to be the \((k + kd) \times 1\) parameter. This is the so-called “reduced-form coefficients” in linear IV models. Many papers on weak instruments start from considering reduced-form coefficients and their estimators. Let us verify that \( \psi \) is indeed an
underlying regular parameter for $\beta$. Recall from Example 2.1 in the previous section that $\hat{\pi}_\beta = \mathbb{E}_P[z']^{-1}\mathbb{E}_P[zu g]$ and $\bar{\pi} = \mathbb{E}_P[z]^{-1}\mathbb{E}_P[z'g]$, that is, the local parameter of $(\pi_\beta, \text{vec}(\pi))$ is a continuous linear functional of the score; therefore, $\psi$ is regular with $\dot{\psi}_P g = (\dot{\psi}_1, \text{vec}(\dot{\psi}_2)) = (\hat{\pi}_\beta, \bar{\pi})$. The submodel $\mathcal{P}_\beta$ contains all distributions in $\mathcal{P}$ that satisfy $\det(\dot{\psi}_2^\top \dot{\psi}_2) \neq 0$ and $(I_k - \dot{\psi}_2 \dot{\psi}_2^\top)\dot{\psi}_1 = 0$. Since

$$\beta_P(g) = \dot{\psi}_2^\top \dot{\psi}_1,$$

$\psi$ is an underlying regular parameter for $\beta$ with $\beta_{P,\psi}(\dot{\psi}_1, \dot{\psi}_2) = \dot{\psi}_2^\top \dot{\psi}_1$ defined on $D_\beta = \{(\dot{\psi}_1, \text{vec}(\dot{\psi}_2)) \in \mathbb{R}^k \times \mathbb{R}^{k \times d} : \det(\dot{\psi}_2^\top \dot{\psi}_2) \neq 0, (I_k - \dot{\psi}_2 \dot{\psi}_2^\top)^{-1} \dot{\psi}_2^\top = 0\}$. In fact, this underlying parameter admits the direct representation $\beta(Q_t) = \psi_2(Q_t)^\top \psi_1(Q_t)$.

There are other formulations of an underlying regular parameter. Let $\pi_d$ be the first $d \times d$ submatrix of the $k \times d$ matrix $\pi$. Then $\psi(d) := (\pi_d \beta, \text{vec}(\pi_d))$ is also an underlying regular parameter for $\beta$ since

$$\beta_P(g) = \dot{\psi}_{(d),2}^\top \dot{\psi}_{(d),1}$$

with an analogous definition of $\dot{\psi}_{(d)}$. Or, the $d \times d$ matrix can in fact be any (non-degenerate) combination of coefficients on $k$ instruments, as long as one can recover the value of $\beta$. This is to say that in overidentified linear IV models $(k > d)$, there are many natural choices of underlying regular parameters. □

### 2.5 Minimal Sufficient Underlying Regular Parameters

There are many possible ways to parameterize an underlying regular parameter. This section addresses the remaining questions: (i) What makes an underlying parameter a good one? (ii) Is there a best choice? (iii) How can we find an underlying regular parameter? Question (i) leads to the concept of informational sufficiency of the underlying parameter. We say that an underlying parameter is sufficient if it contains
all the information captured by the weakly regular parameter: its identification and value. Question (ii) leads to the concept of informational minimality of the underlying parameter. We say that an underlying parameter is \textit{minimal} if it does not contain information that is irrelevant to the weakly regular parameter; in other words, a minimal underlying parameter does not contain a “nuisance parameter.” Putting these together, once we find an underlying parameter that is sufficient and minimal, we can “forget” about the weakly regular parameter and concentrate on the model that consists only of regular parameters. Question (iii) depends on applications.

Although the proofs in Lemma \ref{lemma2.3} and Theorem \ref{theorem2.2} contain explicit construction, it is not a reasonable method to identify the natural parameterization of an underlying parameter. Therefore, this question will be addressed later by examples.

In order to answer these questions, we develop the theory of information for the parameters in the model. In doing so, we utilize the insight from classical semiparametric theory: there are many paths that yield the same parameter value and some are less informative than others; while we generally want the score to be large—so the information is large—we also want to allow as many paths as possible in order to accommodate the desired level of semiparametric generality, thereby running the risk of having scores that are small. It is essential, therefore, to identify the maximum amount of information that is contained in \textit{any} path that is allowed within the generality of the model. In this spirit, we establish a way to discover the minimal tangent subspace that is “spanned” by the parameter of interest.

### 2.5.1 Nuisance Tangent Spaces

By definition, a weakly regular parameter satisfies $\beta(Q_t) \rightarrow \beta_P(g)$ for every path $Q_t$. To illuminate the idea in the coming definition, let us assume for the sake of argument that $\beta_P : \hat{P}_{\beta,P} \rightarrow \mathbb{B}$ is “invertible” in the sense that the equation $\beta_P(g) = b$ can be written as $g = g_\eta + g_\beta(b)$, as is often the case in many models. The first term $g_\eta$ does not affect the value of $\beta_P$, so it is nuisance; the second term is the important component of the score that contains information about $\beta_P$. Thus, the tangent space consists of two subspaces: one spanned by $\{g_\eta\}$ and the other by $\{g_\beta(b) : b \in \mathbb{B}\}$. 

148
In the nuisance parameter literature, the “efficient tangent space” that represents the maximal amount of relevant information contained in the model is derived through the projection of \( \{g_\beta(b)\} \) onto the orthocomplement of \( \{g_\eta\}\).\(^{20}\) Then, we naturally expect that the efficient tangent space for a good underlying regular parameter should contain and only contain the efficient tangent space for \( \beta \), representing the right amount of information in the local expansion.

What makes the analysis nonstandard is the involvement of the nonlinear map \( \beta_P \) in the local expansion around \( P \). In the classical semiparametric theory, the score and the local parameters are related linearly to each other, thereby leading to a very nice use of the theory of linear operators \([Bickel et al., 1993]\). The following definition extends the key notions from this literature to a nonlinear map defined on a cone of a linear space in order to restore the similar structure in our local expansion.

**Definition.** Let \( \mathcal{X} \) be a linear space and \( \mathcal{Y} \) a set. For a map \( f : A \rightarrow \mathcal{Y} \) defined on a cone \( A \) in \( \mathcal{X} \), define the *range* \( R \) and *kernel* \( N \) by

\[
R(f) := \{y \in \mathcal{Y} : y = f(x) \text{ for some } x \in A\},
\]

\[
N(f) := \{\tilde{x} \in \mathcal{X} : x + \tilde{x} \in A \text{ and } f(x + \tilde{x}) = f(x) \text{ for every } x \in A\}.
\]

**Remark.** If \( f \) is linear and \( A = \mathcal{X} \), \( N(f) \) reduces to the standard definition of a kernel for linear maps.

Now we are ready to define the smallest tangent set for \( \beta \), using the notion of the kernel instead of “invertibility.” The key is that any score that does not affect identification or the value of \( \beta \) only contains information about the path that is irrelevant to \( \beta \); such a score can be deemed nuisance. Since the tangent space \( \hat{P}_P \) is linear, we separate the space into the space spanned by nuisance scores and the residual space. That residual space will, by construction, only contain scores that are relevant to either identification or value of \( \beta \). This is the minimal tangent set, which we will show shortly is a cone.

\(^{20}\)For more intuition on this point, see \cite{Van der Vaart (1998) Section 25.4}.
Definition (Nuisance tangent space). Let $\beta : \mathcal{P}_\beta \to \mathbb{B}$ be weakly regular. Call the kernel $N(\beta_P) \subset \hat{\mathcal{P}}_P$ of $\beta_P$ the nuisance tangent space. Denote by $\Pi_{-\beta}$ and $\Pi_\beta$ the projection operators onto $N(\beta_P)$ and $N(\beta_P)^\perp$ in $L_2(P)$.

The definition of $N(\beta_P)$ tells that $\tilde{g} \in N(\beta_P)$ means $g + \tilde{g} \in \dot{\mathcal{P}}_{P,\beta}$ and $\beta_P(g + \tilde{g}) = \beta_P(g)$ for every $g \in \hat{\mathcal{P}}_{P,\beta}$; the first condition is the preservation of identification and the second the preservation of the value of $\beta$. The flip side of this is that if $\tilde{g} \notin \dot{\mathcal{P}}_{P,\beta}$, then there exists $g \in \hat{\mathcal{P}}_{P,\beta}$ such that either $g + \tilde{g} \notin \dot{\mathcal{P}}_{P,\beta}$ or $\beta_P(g + \tilde{g}) \neq \beta_P(g)$ is true. Therefore, such $\tilde{g}$ can be considered to hold information of either identification or distinction of $\beta$.

The following lemma verifies that the minimal tangent cone is indeed a cone and the nuisance tangent space is a linear space.

Lemma 2.1. The following hold.

i. $N(\beta_P)$ is a linear space.

ii. If $P \in \mathcal{P} \setminus \mathcal{P}_\beta$, then $N(\beta_P) \subset \dot{\mathcal{P}}_P \setminus \dot{\mathcal{P}}_{P,\beta}$.

iii. If $P \in \mathcal{P} \setminus \mathcal{P}_\beta$, then $g \in \dot{\mathcal{P}}_{P,\beta}$ implies $\Pi_\beta g \neq 0$.

Now we clarify these concepts in linear IV models.

Example 2.1 (Linear IV, continuing from p. 146). Recall the score formula:

$$g = g_{uvz} - z'\hat{\pi}'\beta \frac{dP_{uvz,u}}{dP} - z'\hat{\pi}' \frac{dP_{uvz,v}}{dP},$$

where $\mathbb{E}_P[zug_{uvz}] = 0$ and $\mathbb{E}_P[zv'g_{uvz}] = 0$. This implies that the space spanned by $g \in \dot{\mathcal{P}}_P$ such that $\mathbb{E}_P[zug] = 0$ and $\mathbb{E}_P[zv'g] = 0$ must be contained in $N(\beta_P)$. On the other hand, any other score will change the value of either $\hat{\pi}$ or $\beta$. Suppose that for $g_1 \in \hat{\mathcal{P}}_{P,\beta}$, adding the score $\tilde{g} \in \hat{\mathcal{P}}_P$ would change the value of $\hat{\pi}$ but not $\beta$. So one may write

$$g_1 = g_{uvz} - z'\hat{\pi}'\beta \frac{dP_{uvz,u}}{dP} - z'\hat{\pi}' \frac{dP_{uvz,v}}{dP}, \quad \tilde{g} = \tilde{g}_{uvz} - z'\hat{\pi}'\beta \frac{dP_{uvz,u}}{dP} - z'\hat{\pi}' \frac{dP_{uvz,v}}{dP}.$$
Then take $g_2$ to be such that

$$g_2 = -z'(2\beta) \frac{dP_{uvz,a}}{dP} - z' \pi \frac{dP_{uvz,v}}{dP}$$

so that $\beta_P(g_2) = 2\beta$. Then adding $\tilde{g}$ to $g_2$ will change the value of $\beta$ since $(\tilde{\pi} + \pi)^- (2\pi + \tilde{\pi}) \gamma \neq \beta$; it can even be that $(2\pi + \tilde{\pi}) \gamma$ falls outside of the column space of $\pi + \tilde{\pi}$. Therefore, such $\tilde{g}$ cannot be in $N(\beta_P)$. Thus we see that $N(\beta_P)$ equals the set of scores $g$ such that $E_P[zug] = 0$ and $E_P[zv'g] = 0$, or, the set of scores induced by $\mathcal{P}_{uvz}$.

**Example 2.8 (Regular parameter).** A regular parameter $\psi : \mathcal{P} \to \mathbb{D}$ can be considered trivially weakly regular. Since $\psi$ is defined on the whole of $\mathcal{P}$, $\mathcal{P}_\beta = \mathcal{P}$ and $\mathcal{P}_{P,\beta} = \mathcal{P}_P$. Since $\psi$ is differentiable at $P$, it is continuous at $P$, that is, $\psi(Q_t) \to \psi(P)$ for every $Q_t \in \mathcal{P}_{P,\beta}$. Therefore, the limit of $\psi(Q_t)$ can be trivially written as a constant function $\psi_P : \mathcal{P}_P \to \mathbb{D}$ such that $\psi_P(g) \equiv \psi(P)$ for every $g \in \mathcal{P}_P$. Any change in the score cannot affect the value of $\psi_P$, so $N(\psi_P) = \mathcal{P}_P$ and $\mathcal{P}_{P,\psi} = \{0\}$.

### 2.5.2 Sufficiency and Minimality of Underlying Regular Parameters

The underlying regular parameters are characterized by the span of their “scores,” or equivalently, of their efficient influence maps. The first property we want in the underlying regular parameter is that it contain all relevant information about $\beta$ that is available in the model. Here, “information” is captured by the ability to discern distinct scores in the limit.

**Definition** (Sufficiency of underlying regular parameter). Let $\beta : \mathcal{P}_\beta \to \mathbb{D}$ be weakly regular. An underlying regular parameter $\psi : \mathcal{P} \to \mathbb{D}$ for $\beta$ is **sufficient** if

$$N(\psi_P) \subset N(\beta_P),$$

---

21 See Van der Vaart (1991b) and Bickel et al. (1993) for equivalence.
or equivalently, $N(\beta_P)^\perp \subset R(\dot{\psi}_P)$\footnote{Use the property of an adjoint operator: $N(\dot{\psi}_P)^\perp = R(\dot{\psi}_P)$ (Kosorok 2008, Equation 17.3).}

The efficient influence map $\dot{\psi}_P^*$ of an underlying parameter $\psi$ summarizes the set of scores that the local parameter of $\psi$ can distinguish. Therefore, if $\psi$ is sufficient, then knowing the local parameter of $\psi$ gives a sufficient amount of information that a score would contain about the identification or distinction of $\beta$. The equivalent formulation says that the score that $\psi$ cannot distinguish is never used in identification or valuation of $\beta$.

The following example shows that an underlying regular parameter need not be sufficient.

**Example 2.1** (Insufficiency in linear IV, continuing from p. 150). Let $d = 1$ and $k > 1$. Consider the underlying regular parameter

$$\psi(Q) = (\pi_1 \beta, \pi_1)$$

that induces $\dot{\psi}_P g = (\dot{\pi}_1 \beta, \dot{\pi}_1)$. This parameter only uses the first instrument and abandons information from all other instruments available in the model. Therefore, $N(\dot{\psi}_P)$ contains elements $g$ that change the value of $\dot{\pi}_2$. However, changing the value of $\dot{\pi}_2$ without adjusting for the values of $\dot{\pi}_1 \beta$ and $\dot{\pi}_1$ will make $\beta$ undefined and push the score outside of $\hat{P}_{P, \beta}$, so we have $g \notin N(\beta_P)$. Hence, $\psi$ is not sufficient. \hfill $\square$

Not surprisingly, a sufficient underlying regular parameter contains information of all instruments.

**Example 2.1** (Sufficiency in linear IV, continuing from p. 152). The underlying regular parameter

$$\psi(Q) = (\pi \beta, \text{vec}(\pi))$$

is sufficient. To see this, note that $\dot{\psi}_P g = (\dot{\pi} \beta, \text{vec}(\dot{\pi}))$ and take $g_q \in N(\psi_P)$. The values of $\dot{\pi} \beta$ and $\dot{\pi}$ do not change by adding $g_q$ to the score. Recalling the score formula

$$g = g_{uvw} - z' \dot{\pi} \beta \frac{dP_{uvw,u}}{dP} - z' \dot{\pi} \frac{dP_{uvw,v}}{dP},$$

we see that

$$N(\dot{\psi}_P)^\perp = R(\dot{\psi}_P)$$

(Kosorok 2008, Equation 17.3).
one sees that $g + g_\eta \in \hat{\mathcal{P}}_{P,\beta}$ whenever $g \in \hat{\mathcal{P}}_{P,\beta}$ and $\beta_P(g + g_\eta) = \beta_P(g)$, that is, $g_\eta \in N(\beta_P)$. Therefore, $\psi$ is sufficient.

The next property we want in an underlying regular parameter is that it has only relevant information for the weakly regular parameter. Otherwise, the underlying parameter contains some information of a “nuisance parameter” and estimating it may capture unwanted noise that is irrelevant to estimation of the weakly regular parameter.

**Definition** (Minimality of underlying regular parameter). Let $\beta : \mathcal{P}_\beta \to \mathbb{B}$ be weakly regular. An underlying regular parameter $\psi : \mathcal{P} \to \mathbb{D}$ for $\beta$ is minimal if

$$N(\beta_P) \subset N(\hat{\psi}_P),$$

or equivalently, $R(\hat{\psi}_P) \subset N(\beta_P)^\perp$.

Minimality of $\psi$ requires the opposite inclusion to sufficiency between $N(\beta_P)$ and $N(\hat{\psi}_P)$. This is to say that the score irrelevant to identification or distinction of $\beta$ is also irrelevant to distinction of the local parameter of $\psi$. Equivalently, the range of the efficient influence map of $\psi$ does not contain a score that is unrelated to $\beta$. In this sense, a minimal underlying parameter is “free” of potential nuisance parameters.

**Example 2.1** (Minimality in linear IV, continuing from p.152). As seen in Example 2.1 in the previous subsection, $N(\beta_P)$ is the set of scores $g_{uwz}$ induced by $\mathcal{P}_{uwz}$. Again, recalling the score formula, one sees that adding such scores does not change the values of $\hat{\pi}\beta$ and $\hat{\pi}$, which implies $N(\beta_P) \subset N(\hat{\psi}_P)$ for both choices of the underlying regular parameter we discussed: $(\pi\beta, \text{vec}(\pi))$ and $(\pi_1\beta, \pi_1)$.

Note that $\psi = (\pi\beta, \text{vec}(\pi))$ is still minimal even in the homoskedastic model. Homoskedasticity helps simply actual efficient estimation, but it does not help much in simplifying the semiparametric structure itself.

**Remark.** Minimal sufficiency in our definition is of a parameter, while minimal sufficiency in the context of sufficient statistics is of a statistic.
The following theorem ensures that an underlying regular parameter that is minimal and sufficient exists.

**Theorem 2.2** (Existence of minimal sufficient underlying regular parameter). For a weakly regular parameter $\beta : \mathcal{P}_\beta \to \mathbb{B}$, there exists a minimal sufficient underlying regular parameter.

Minimal sufficiency *per se* is not strong enough to pin down the underlying parameter uniquely. However, underlying parameters that are both minimal and sufficient are almost equivalent in terms of the information they contain. The next theorem characterizes minimal sufficient underlying parameters and establishes this “equivalence.”

**Theorem 2.3** (Characterization of minimal sufficient underlying regular parameter). Let $\beta : \mathcal{P}_\beta \to \mathbb{B}$ be weakly regular and $\psi : \mathcal{P} \to \mathbb{D}$ a sufficient underlying regular parameter for $\beta$. Then $\psi$ is minimal if and only if for any sufficient underlying regular parameter $\phi : \mathcal{P} \to \mathbb{E}$ for $\beta$ on a Banach space $\mathbb{E}$ there exists a linear map $\tau : \mathbb{E} \to \mathbb{D}$ such that

$$\tau(\dot{\phi}_P g) = \dot{\psi}_P g \quad \text{for every} \quad g \in \dot{\mathcal{P}}_P.$$ 

*Remark.* Theorem 2.3 can be understood as “almost uniqueness” of observational information regarding minimal sufficient underlying regular parameters. If the linear map $\tau$ between two minimal sufficient underlying regular parameters is bicontinuous, efficiency in one parameterization implies efficiency in the other (see, e.g., [Van der Vaart](1991a) [1998] Section 25.7).

Although the proof of Theorem 2.2 contains the explicit construction of the minimal sufficient underlying parameter, it is not the most natural way to find one in virtually any context. However, reduced-form parameters are often minimal sufficient underlying parameters.

**Example 2.1** (Linear IV, continuing from p. 153). As seen earlier, $\psi = (\pi \beta, \text{vec}(\pi))$ is a natural choice of an underlying regular parameter that is minimal and sufficient.
Given the minimal sufficient underlying regular parameter, the problem of estimation or inference on a weakly regular parameter can be translated into a problem of estimation or inference on the local parameter of the minimal sufficient underlying parameter. Being a local parameter of a regular parameter, it will plausibly provide workable grounds for the intended statistical analysis.

One caveat applies. Unlike regular parameters in the classical theory, we “know” that the local parameter of the underlying regular parameter $\dot{\psi}_{P\theta}$ must lie in the strict subset $D_\beta$ of $D$. This constraint, next to the nonlinearity of $\beta_{P,\psi}$, may stand as a second source of complication in estimation and inference.

### 2.6 Weak Efficiency for Weakly Regular Parameters

This section describes efficiency of estimators of a weakly regular parameter. The difficulty in formulating a reasonable goodness criterion for weakly regular parameters lies in the fact that their asymptotic distribution is nonstandard. Both the convolution and minimax theorems in the classical theory require symmetry of the asymptotic distribution while many natural estimators of a weakly regular parameter do not lead to a symmetric distribution. In the context of our formulation, the limit of a weakly regular parameter is a nonlinear transformation of the local parameter of an underlying regular parameter. As the estimator of a local parameter often leads to a Gaussian random variable, one can anticipate that the asymptotic distribution of the estimators of a weakly regular parameter would lead to a similar nonlinear transformation of a Gaussian distribution. Moreover, a nonlinear transformation of a Gaussian distribution can, in general, be anything. Our idea of defining efficiency for weakly regular parameters lies in observing that the goodness criterion of the convolution theorem—whether the limit distribution involves an irrelevant noise—should carry over even with a nonlinear transformation. In particular, in light of

---

23 Technically, the minimax theorem can sometimes be applied to nonsymmetric distributions (Fang, 2016).
Jensen’s inequality, the involvement of an irrelevant noise should increase any convex loss.

Assuming that $\beta_{P,\psi}$ is smooth enough, the problem of accurately estimating $\beta$ translates into a problem of accurately estimating $\psi$. If the natural estimator of $\psi$ takes values so that the estimated local parameter falls into $\mathbb{D}_\beta$ (the range of $\dot{\psi}_P$ on the pertinent tangent cone $\mathcal{P}_{P,\beta}$) with probability one, then there is nothing that needs to be done (and indeed this is the case in many applications). If not, we need to accommodate the constraint $\dot{\psi}_P g \in \mathbb{D}_\beta$ in either of the following ways.

1. Estimate $\psi$ with the constraint, so we have $\dot{\psi}_P g \in \mathbb{D}_\beta$ almost surely.

2. Estimate $\psi$ without the constraint; then deal with values outside of $\mathbb{D}_\beta$.

Noting that the first approach can be considered to be estimating $\psi$ without the constraint and then reconstructing another estimator that satisfies the constraint, the two approaches are essentially equivalent. In this section, we employ the latter interpretation that estimators of $\beta$ are constructed from estimators of $\psi$ without accounting for the constraint.

In order to make use of the convolution theorem, we need to restrict our attention to the estimators of $\beta$ that are transformations of some regular estimators of $\psi$. Throughout this section, we assume that $\psi$ can be estimated regularly and efficiently at root-$n$.\footnote{Not all regular parameters admit root-$n$ consistent estimation, especially infinite-dimensional ones (Wellner et al., 2006; Giné and Nickl, 2015). We make this assumption here since the focus of this chapter is to demonstrate the power of reduction to regular parameters in the context of weak identification and many interesting examples in economics indeed admit root-$n$ consistent estimation.}

First, recall the definition of a regular estimator for a regular parameter.

**Definition** (Regular estimator for regular parameter). A sequence of estimators $\hat{\psi}_n$ for a regular parameter $\psi : \mathcal{P} \to \mathbb{D}$ is called regular at $P \in \mathcal{P}$ relative to $\mathcal{P}_P$ if there exists a tight Borel random element $L$ in $\mathbb{D}$ such that

$$\sqrt{n}(\hat{\psi}_n - \psi(Q_n)) \overset{Q_n}{\rightarrow} L \quad \text{for every} \quad Q_n \in \mathcal{P}_P.$$

This sequence is called (semiparametric) efficient at $P$ relative to $\mathcal{P}_P$ if it attains
the distributional lower bound (denote it by $L_\psi$) of the convolution theorem (Van der Vaart 1991a, Theorem 2.1; Kosorok 2008, Theorem 18.3).

**Remark.** The convolution theorem states that $L = L_\psi + L_\eta$ where $L_\psi$ and $L_\eta$ are independent tight Borel random elements in $D$ such that $\Pr(L_\psi \in R(\dot{\psi}_P)) = 1$ and

$$\delta^* L_\psi \sim N\left(0, \|\dot{\psi}_P^*\delta^*\|_{L_2(P)}^2\right) \quad \text{for every} \quad \delta^* \in \mathbb{D}^*.$$  

This is to say, the asymptotic distribution of any regular estimator of a regular parameter is the sum of a Gaussian variable with covariance being the “$L_2$ norm” of the efficient influence map and an independent noise. It is efficient when no noise is involved, that is, when $L_\eta \equiv 0$.

**Remark.** If one centers the estimator at $\psi(P)$, the convergence becomes $\sqrt{n}(\hat{\psi}_n - \psi(P)) \sim Q_n \dot{\psi}_P g + L$.

Now we define the class of estimators we consider. We focus on estimators of a weakly regular parameter that can be represented as transformations of estimators of a minimal underlying regular parameter. Many estimators in the literature fall within this class. Asymptotic randomness of such estimators must come from estimators of the local parameter of the underlying parameter $\sqrt{n}(\hat{\psi}_n - \psi(P))$ and possibly some irrelevant noise. Toward this end, we define functions that admit approximation by a root-$n$ normalization around the point of identification failure.

**Definition** (Local continuous approximability). Let $D_n$ be indexed subsets of $D$. A sequence of maps $T_n : D_n \times [0, 1] \to \mathbb{B}$ is locally continuously root-$n$ approximable at $\delta \in D$ tangentially to $D_\infty \subset D$ if there exists a measurable map $T_\delta : D_\infty \times [0, 1] \to \mathbb{B}$ such that for every $u \in [0, 1]$, $\delta + \delta_n / \sqrt{n} \in D_n$, $\delta_\infty \in D_\infty$, if a subsequence $\delta_n' \to \delta_\infty$, then

$$T_n \left(\delta + \frac{\delta_n'}{\sqrt{n}}, u\right) \to T_\delta(\delta_\infty, u) \quad \text{as} \quad n \to \infty.$$  

\footnote{If $X$ is a Gaussian random variable in a Euclidean space $\mathbb{R}^k$, then for every vector $a \in \mathbb{R}^k$ the inner product $a'X$ is Gaussian in $\mathbb{R}$. Analogously, if a random variable $X$ in a Banach space $D$ is Gaussian, then for every $\delta^*$ in the dual space $D^*$, $\delta^* X$ is Gaussian in $\mathbb{R}$.}
Denote this by $T_n \to^\delta T_\delta$ and call $T_\delta$ the approximating function of $T_n$ at $\delta$. We use these definitions even when $T_n$ and $T_\delta$ do not depend on the second argument since they can be considered trivially dependent on (i.e., constant with respect to) $u \in [0,1]$.

Remark. By construction $T_\delta$ is continuous on $D_\infty$.

Example 2.9. The sequence of functions $f_n : \mathbb{R} \times (\mathbb{R} \setminus \{b\}) \to \mathbb{R}$ such that $f_n(x, y) = (x - a)/(y - b)$ is locally continuously root-$n$ approximable at $(a, b)$ tangentially to $\mathbb{R} \times (\mathbb{R} \setminus \{0\})$ with an approximating function $f_{(a,b)}(h_x, h_y) = h_x/h_y$ if $h_y \neq 0$. Note that $f$ is not continuous at $(a, b)$. □

Now the class of estimators we consider is one that asymptotically admits approximation by a locally continuously approximable function.

Definition (Regular estimator for weakly regular parameter). A sequence of estimators $\hat{\beta}_n$ for a weakly regular parameter $\beta : \mathcal{P}_\beta \to \mathbb{B}$ is called regular at $P \in \mathcal{P}$ relative to $\mathcal{P}_{P,\beta}$ if there exist a minimal underlying regular parameter $\psi : \mathcal{P} \to \mathbb{D}$ for $\beta$, a regular sequence of estimators $\hat{\psi}_n$ of $\psi$ with $\sqrt{n}(\hat{\psi}_n - \psi(Q_n)) \overset{Q_n}{\sim} L$, a sequence of nonrandom maps $T_n : \mathbb{D} \times [0,1] \to \mathbb{B}$ that is locally continuously root-$n$ approximable at $\psi(P)$ tangentially to the range of $\dot{\psi}_P g + L$ for $g \in \dot{\mathcal{P}}_{P,\beta}$, and an independent noise $U \sim U[0,1]$ such that $T_{\psi(P)}(L, U)$ is Borel measurable and

$$\hat{\beta}_n = T_n(\hat{\psi}_n, U) + o_P(1) \quad \text{under every} \quad Q_n \in \mathcal{P}_{P,\beta}.$$

Remark. We may without loss of generality take $\psi$ as sufficient; for otherwise one can augment $\psi$ and have $T_n$ ignore the augmented part.

The asymptotic distribution of a regular estimator is a transformation of the asymptotic distribution of the underlying regular parameter.

Proposition 2.1. Let $\hat{\beta}_n = T_n(\hat{\psi}_n, U) + o_P(1)$ be a regular sequence of estimators for a weakly regular parameter $\beta : \mathcal{P}_\beta \to \mathbb{B}$ and $\sqrt{n}(\hat{\psi}_n - \psi(Q_n)) \overset{Q_n}{\sim} L$. Then,

$$\hat{\beta}_n \overset{Q_n}{\to} T_{\psi(P)}(\dot{\psi}_P g + L, U).$$
Remark. More generally, a regular estimator $\hat{\beta}_n$ can be viewed as a possibly random transformation $\hat{T}_n(\hat{\psi}_n, U)$ of $\hat{\psi}_n$ and $U$, and randomness of $\hat{T}_n$ vanishes in an appropriate sense. In the linear IV example below, we show directly that there exists a nonrandom transformation $T_n$ such that the difference $\hat{T}_n(\hat{\psi}_n, U) - T_n(\hat{\psi}_n, U)$ is $o_P(1)$.

To aid in other cases, we give a version of the continuous mapping theorem extended to random mappings in Addendum 2.1 in Appendix 2.A.

Example 2.1 (Linear IV, continuing from p. 154). Popular estimators of the linear IV model are regular. Consider the 2SLS. Observe that the reduced-form coefficients $(\pi \beta, \pi)$ are regular and the 2SLS can be written as a function of their estimators $\tilde{\pi}_n = (Z'Z)^{-1}Z'X$ and $\tilde{\pi}_n \beta_n = (Z'Z)^{-1}Z'Y$:

$$\tilde{\beta}_{2SLS} = (\tilde{\pi}_n'Z'Z\tilde{\pi}_n)^{-1}\tilde{\pi}_n'(Z'Z)\tilde{\pi}_n \beta_n = (\tilde{\pi}_n'\mathbb{E}[zz']\tilde{\pi}_n)^{-1}\tilde{\pi}_n'\mathbb{E}[zz']\tilde{\pi}_n \beta_n + o_P(1).$$

The residual is $o_P(1)$ since $(Z'Z)/n$ converges to $\mathbb{E}[zz']$ in probability under every path. This shows regularity of the 2SLS with the homogeneous approximating function $T : \mathbb{R}^k \times \mathbb{R}^{k \times d} \to \mathbb{R}^d$

$$T(\pi \beta, \pi) = (\pi'\mathbb{E}[zz']\pi)^{-1}\pi'\mathbb{E}[zz']\pi \beta.$$

Recall that a subset of the instruments is not sufficient (Example 2.1). However, the 2SLS estimator that uses only a part of the instruments is also regular. To see this, let the subscript $(d)$ denote the selection of $d$ instruments. The partial 2SLS estimator can be seen as a function of an estimator for the entire reduced-form coefficients as

$$\tilde{\beta}_{2SLS,(d)} = T_n(\tilde{\pi}_n, \tilde{\pi}_n \beta_n, U) = (\tilde{\pi}_n'(d)\mathbb{E}[z(d)z'(d)]\tilde{\pi}_n(d))^{-1}\tilde{\pi}_n'(d)\mathbb{E}[z(d)z'(d)]\tilde{\pi}_n \beta_n + o_P(1),$$

where $\tilde{\pi}_n(d)$ and $\tilde{\pi}_n \beta_n(d)$ are an estimator of the entire reduced-form coefficients using only the $(d)$ subset of instruments, and the remaining parts of $\tilde{\pi}$ and $\tilde{\pi} \beta$ can be anything (as long as $(\tilde{\pi} \beta, \tilde{\pi})$ is regular).

Similarly, GMM can be shown to be regular. Denote by $W$ the weighting matrix.
The GMM estimator $\tilde{\beta}_{GMM}$ with weighting $W$ solves
\[
\min_b \left[ \frac{Z'(Y - Xb)}{n} \right]' W \left[ \frac{Z'(Y - Xb)}{n} \right].
\]
Write the objective function as
\[
\frac{1}{n} \sqrt{n}(\tilde{\pi}_n b - \tilde{\pi}_n b)' W \frac{Z'Z}{n} \sqrt{n}(\tilde{\pi}_n b - \tilde{\pi}_n b).
\]
The oracle weighting matrix for the efficient GMM is
\[
W = \mathbb{E}[(y - x'\beta)^2 z z']^{-1},
\]
while it is estimated in case of the feasible GMM. In particular, the two-step GMM estimates $W$ by plugging in the 2SLS estimator for $\beta$ and taking its sample counterpart, i.e., $\tilde{W}_{2SGMM} = \mathbb{E}_n[(y - x'\tilde{\beta}_{2SLS})^2 z z']^{-1}$. The expectation involved in $\tilde{W}$ (other than the 2SLS estimator) can be consistently estimated. Moreover, minimization is invariant to scaling, so if one sees minimization as a function of the 2SLS estimators, it is homogeneous of degree zero (and continuous). This shows that the two-step GMM is regular.

The Fuller estimator proposed by Fuller (1977) is regular while we suspect that the heteroskedasticity-robust Fuller (HFUL) estimator proposed by Hausman et al. (2012) is not. Let $P := Z(Z'Z)^{-1}Z'$. For a constant $C$, let $\tilde{P}_{\text{Fuller}} := P + (C/n)(I - P)$. The Fuller estimator is given by
\[
\tilde{\beta}_{\text{Fuller}} = (X'\tilde{P}_{\text{Fuller}} X)^{-1}(X'\tilde{P}_{\text{Fuller}} Y)
\]
\[
= \left( \frac{C}{n} X'X + \left[ 1 - \frac{C}{n} \right] \tilde{\pi}_n (Z'Z) \tilde{\pi}_n \right)^{-1} \left( \frac{C}{n} X'y + \left[ 1 - \frac{C}{n} \right] \tilde{\pi}_n (Z'Z) \tilde{\beta} \right)
\]
\[
= \left( C \mathbb{E}[xx'] + \sqrt{n} \tilde{\pi}_n \mathbb{E}[zz'] \sqrt{n} \tilde{\pi}_n \right)^{-1} \left( C \mathbb{E}[xy] + \sqrt{n} \tilde{\pi}_n \mathbb{E}[zz'] \sqrt{n} \tilde{\beta} \right) + o_P(1).
\]
Thus, under weak identification, the Fuller estimator can be thought of as a “weighted combination” of OLS ($C = \infty$) and 2SLS ($C = 0$). On the other hand, due to its
jackknife form, the HFUL estimator requires calculation of the off-diagonal matrix of \( P \) (Hausman et al., 2012). While this is their source of robustness to heteroskedasticity (under different asymptotics), this makes it challenging, possibly infeasible, to represent HFUL only as a function of the OLS estimator.

The unbiased estimator by Andrews and Armstrong (2017) is also regular. For simplicity, let \( d = 1 \) and \( k = 1 \) and assume that \( \pi > 0 \) and \( \hat{\pi}_n \) and \( \hat{\pi}_\beta_n \) are asymptotically uncorrelated. Then the unbiased estimator of \( \beta \) is

\[
\hat{\beta}_{\text{unbiased}} = \frac{\sqrt{n} \hat{\pi}_n \beta_n}{\hat{\sigma}_{\pi,\beta}} \frac{1 - \Phi(\sqrt{n} \hat{\pi}_n / \hat{\sigma}_{\pi,\beta})}{\sigma_{\pi,\beta} \phi(\sqrt{n} \hat{\pi}_n / \hat{\sigma}_{\pi,\beta})} = \frac{\sqrt{n} \hat{\pi}_n \beta_n}{\hat{\sigma}_{\pi,\beta}} \frac{1 - \Phi(\sqrt{n} \hat{\pi}_n / \sigma_{\pi,\beta})}{\sigma_{\pi,\beta} \phi(\sqrt{n} \hat{\pi}_n / \sigma_{\pi,\beta})} + o_P(1),
\]

which is regular with

\[
T_n(\hat{\pi}_n, \hat{\pi}_\beta_n, U) = \frac{\sqrt{n} \hat{\pi}_n \beta_n}{\sigma_{\pi,\beta}} \frac{1 - \Phi(\sqrt{n} \hat{\pi}_n / \sigma_{\pi,\beta})}{\sigma_{\pi,\beta} \phi(\sqrt{n} \hat{\pi}_n / \sigma_{\pi,\beta})}, \quad T_{(0,0)}(\hat{\pi}, \hat{\pi}_\beta, U) = \frac{\hat{\pi} \beta}{\sigma_{\pi,\beta}} \frac{1 - \Phi(\hat{\pi} / \sigma_{\pi})}{\sigma_{\pi,\beta} \phi(\hat{\pi} / \sigma_{\pi})}.
\]

\[\square\]

2.6.1 Local Asymptotic Rao-Blackwellization

The following theorem states that for any regular estimator of a weakly regular parameter, there exists another regular estimator that is weakly better than the original one in terms of convex loss. A strict improvement is always possible unless our estimator is already a nonrandom transformation of an efficient estimator of the underlying parameter. In other words, whenever an estimator contains some “noise” irrelevant to the efficient estimation of \( \psi \), then one can always construct another estimator that shares the same expectation as the original estimator and is more concentrated around its expectation. This implies that one should use an efficient estimator for the underlying parameter whenever possible. We demonstrate the power of this improvement in Section 2.7.

**Theorem 2.2** (Local asymptotic Rao-Blackwellization). Let \( \beta : \mathcal{P}_\beta \to \mathbb{B} \) be weakly regular and \( \psi : \mathcal{P} \to \mathbb{D} \) a minimal underlying regular parameter for \( \beta \). Let \( \hat{\psi}_n \) be a regular sequence of estimators of \( \psi \) and \( \hat{\beta}_n = T_n(\hat{\psi}_n, U) + o_P(1) \) be a regular
sequence of estimators of $\beta$ with noise $U \sim U[0,1]$. Suppose that an efficient regular sequence of estimators $\hat{\psi}_n$ of $\psi$ exists and $\hat{T}_n(\delta) := \mathbb{E}[T_n(\delta + L_n/\sqrt{n}, U)]$ exists as a Bochner integral. Then $\hat{T}_n(\hat{\psi}_n)$ is a better regular estimator than $\hat{\beta}_n$ in the sense that for every convex continuous loss function $\ell : \mathbb{B} \to \mathbb{R}$ such that $\ell(\hat{\beta}_n - \beta(Q_n))$ and $\ell(\hat{T}_n(\hat{\psi}_n) - \beta(Q_n))$ are asymptotically equiintegrable under $Q_n \in \mathcal{P}_{\mathbb{P}, \beta}$:

$$\liminf_{n \to \infty} \mathbb{E}_{Q_n,*}[\ell(\hat{\beta}_n - \beta(Q_n))] - \mathbb{E}_{Q_n,*}^*[\ell(\hat{T}_n(\hat{\psi}_n) - \beta(Q_n))] \geq 0.$$ 

**Remark.** Theorem 2.2 is a kind of admissibility requirement for a convex loss. Unlike popular discussion of inadmissibility, however, it restricts attention to the class of regular estimators but provides an improvement method (Rao-Blackwellization) to achieve “admissibility.” If $\mathbb{B} = \mathbb{R}$, $\hat{T}_n$ first-order stochastically dominates $T_n$.

**Remark.** Efficiency is usually justified in the context of subconvex loss functions (Kosorok, 2008, Theorem 18.4; Van der Vaart and Wellner, 1996, Theorem 3.11.5). Theorem 2.2 is in the same spirit but restricts us to convex functions. This difference comes from the fact that our best asymptotic distribution is a nonlinear transformation of Gaussian; there is no symmetry of the distribution we can exploit to accommodate subconvexity.

**Example 2.1** (Linear IV, continuing from p.159). Suppose that the reduced-form errors are heteroskedastic, and the feasible GLS estimator $(\hat{\pi}_n, \hat{n}_n)$ is available. As seen in Example 2.1 in the previous section, all of the 2SLS, GMM, Fuller, and unbiased estimators are functions of the OLS estimator of the reduced-form coefficients. Then, Theorem 2.2 suggests in such cases that the use of the 2SLS estimator is asymptotically suboptimal in terms of the concentration of asymptotic distributions measured by convex loss functions. In the case of 2SLS, one can construct a better

---

26 See Bharucha-Reid (1972) for a discussion of Bochner integrals.

27 $X_n$ is asymptotically equiintegrable if $\lim_{M \to \infty} \limsup_{n \to \infty} \mathbb{E}^*[|X_n||\{\{|X_n| > M\}|] = 0$. This condition is referred to in this way in Van der Vaart and Wellner (1996, p. 421).

28 Technically, there is no implication between convexity and subconvexity of a function. In our context, subconvexity can be thought of as roughly weaker.
estimator $\hat{T}_n(\pi\beta, \pi)$ by

$$
\hat{T}_n(\pi\beta, \pi) := \mathbb{E}\left[\left(\left[\pi + \frac{U\pi}{\sqrt{n}}\right] \mathbb{E}[zz'] \left[\pi + \frac{U\pi}{\sqrt{n}}\right]\right)^{-1}\left(\left[\pi + \frac{U\pi}{\sqrt{n}}\right] \mathbb{E}[zz'] \left[\pi\beta + \frac{U\pi\beta}{\sqrt{n}}\right]\right)\right],
$$

where

$$
\left(\sqrt{n}(\pi\beta_n - \pi\beta_n), \sqrt{n}(\pi_n - \pi_n)\right) \rightsquigarrow \left(U\pi, U\pi\right).
$$

In Section 2.7, we compare the performance of 2SLS and its improvement. Interestingly, even with the oracle weighting matrix, GMM contains noise that can be removed if an efficient estimator of the reduced-form coefficients is available.

Note that the limited information maximum likelihood (LIML) estimator is known to have no moment (Chao et al., 2012), being outside the direct scope of Theorem 2.2 but it can be said to be regular by definition. LIML estimates $\hat{W}(b)$ assuming homoskedasticity, that is, $\hat{W}_{\text{LIML}}(b) = n(Z'Z)^{-1}/\hat{\sigma}^2(b)$ where $\hat{\sigma}^2(b) = \mathbb{E}_n[(y - x'b)^2]$ (Andrews, 2017). Since the second and cross moments of $y$ and $x$ can be consistently estimated, LIML is asymptotically only a function of the OLS estimators of the reduced-form coefficients. Similarly, although the continuously updating GMM is suspected to have no moment (Guggenberger, 2005), it is regular as it uses $\hat{W}_{\text{CUGMM}}(b) = \mathbb{E}_n[(y - x'b)^2zz']^{-1}$, which, again, admits consistent estimation. □

### 2.6.2 Weakly Efficient Estimators of Weakly Regular Parameters

Backed up by this result, we define an efficiency concept for estimating a weakly regular parameter. The idea is that when an estimator of a weakly regular parameter does not admit an improvement by Theorem 2.2 we want to call such an estimator “efficient.” The condition under which an estimator does not allow improvement is that it is already a nonrandom transformation of an efficient estimator of the minimal sufficient underlying regular parameter.

**Definition** (Weak efficiency for weakly regular parameter). A regular sequence of
estimators \( \hat{\beta}_n \) for a weakly regular parameter \( \beta \) is *weakly (semiparametric) efficient* at \( P \in \mathcal{P} \) relative to \( \mathcal{P}_{P,\beta} \) if there exist a minimal sufficient underlying regular parameter \( \psi : \mathcal{P} \to \mathbb{D} \), its efficient sequence of estimators \( \hat{\psi}_n \), and a sequence of nonrandom measurable maps \( T_n : \mathbb{D} \to \mathbb{B} \) that is locally continuously root-\( n \) approximable at \( \psi(P) \) tangentially to the range of \( \dot{\psi} g + L_\psi \) such that

\[
\hat{\beta}_n = T_n(\hat{\psi}_n) + o_P(1) \quad \text{under every} \quad Q_n \in \mathcal{P}_{P,\beta}.
\]

Weak efficiency is related to the classical concept of efficiency. Consider two regular parameters, \( \psi_1 \) and \( \psi_2 \), related to each other through a Hadamard differentiable function \( f \) such that \( \psi_2 = f(\psi_1) \). [Van der Vaart (1998, Theorem 25.47)] shows that efficiency of \( \hat{\psi}_1 \) for \( \psi_1 \) implies efficiency of \( f(\hat{\psi}_1) \) for \( \psi_2 \). Recall from a remark in Section 2.4.3 that a weakly regular parameter \( \beta \) is often related to an underlying regular parameter \( \psi \) as \( \beta = \beta_\psi(\psi) \). We define that \( \hat{\beta} \) is *weakly efficient* if it is a nonrandom transformation of an efficient \( \hat{\psi} \). The reason we do not restrict attention to the transformation by \( \beta_\psi \) is due to the impossibility of consistent and nonpivotal estimation (Theorem 2.2). In principle, there is no necessity to treat uncertainty contained in \( \hat{\psi} \) the same way we treat the true value \( \psi \) for the sake of yielding a “good” estimator, if the relationship is nonlinear. Thus, weakly efficient estimators are not unique; this looseness is the grounds on which we call it *weak*.

A straightforward but substantially helpful byproduct of our definition is that, if at some \( Q \in \mathcal{P}_\beta \) the transformation locally reduces to \( \beta_\psi \) and \( \beta_\psi \) is Hadamard differentiable, then a weakly efficient estimator constructed for \( P \in \mathcal{P} \setminus \mathcal{P}_\beta \) becomes *efficient* in the classical sense under the “strong identification asymptotics” at \( Q \in \mathcal{P}_\beta \), provided that \( \hat{\psi}_n \) remains efficient in both asymptotics. This is a direct consequence of the delta method ([Van der Vaart, 1991a, 1998, Section 25.7]). Therefore, weak efficiency can also be regarded as a generalization of efficiency to discontinuous but locally continuously approximable transformations of efficient estimators.

*Remark.* Technically speaking, our definition of weak efficiency does not require the existence of moments. If an estimator does not even have a first moment, then we
cannot apply Theorem 2.2 since any nontrivial convex function should grow at least linearly, but we still consider it weakly efficient if it is a nonrandom transformation of an efficient estimator.

**Example 2.1** (Linear IV, continuing from p. 162). As explained earlier, the 2SLS, GMM, Fuller, and unbiased estimators are inefficient in the presence of heteroskedasticity. If an efficient estimator of the reduced-form coefficients is available, then the Rao-Blackwellized 2SLS, GMM, Fuller, and unbiased estimators conditional on this efficient estimator are weakly efficient. See Section 2.7 for the comparison of the these estimators in simulation. □

**Remark.** A shortcoming of our efficiency concept comes from the fact that the constraint $\dot{\psi}_{Pp} \in \mathcal{D}_\beta$ is not explicitly accounted for. The classical convolution and minimax theorems require that the tangent set be a convex cone. If it is just a cone, the convolution theorem can only conclude the lower bound of the variance (Van der Vaart 1998, Theorem 25.20), which no longer allows invocation of Jensen’s inequality. In our setup, the tangent set $\dot{P}_{P,\beta}$ is a cone but not necessarily convex (Lemma 2.1).

### 2.7 Simulation of Weak Efficiency in Linear IV Models

In order to assess how much bite weak efficiency has, we conduct simulation studies of a linear IV model with heteroskedasticity. In Example 2.1, we consider discrete instruments that collectively take up only on four distinct values; since $z_i$ has a finite support, we can estimate the heteroskedastic covariance matrix without imposing any parametric assumption. This enables us to compute the feasible GLS estimator of the reduced-form coefficients without further restricting the model and use it to construct the Rao-Blackwellized improvements as suggested by Theorem 2.2. We carry out simulation of four estimators: the 2SLS, GMM, Fuller, and unbiased estimators.

In particular, we take $n = 10,000$, $d = 1$, $k = 2$, and $z_i = (z_{i1}, z_{i2})$ where $z_{ij}$ are independent Bernoulli dummy variables. We set $\beta = 0$ and $\pi = (1, 1)/\sqrt{n}$. For
each of the 4 possible values of \( z_i \), the covariance matrix of \((u_i, v_i')\) is randomly drawn and fixed at the beginning of the simulation. We call this the *model A* and iterate it for 5,000 times. We compare three estimators with their Rao-Blackwellization: the 2SLS, two-step GMM, and Fuller estimators. We also show the HFUL estimator for comparison, while we note that we conjecture the HFUL estimator to be non-regular.

Since unbiased estimators in overidentified models take complicated forms, we use just-identified models to compare the performance of the unbiased and Rao-Blackwellized unbiased estimators. We take \( n = 10,000 \) and \( d = k = 1 \) and let \( z_i \) distribute as a uniform discrete random variable over \( \{0, 1, 2\} \); then we compare the unbiased and Fuller estimators and their Rao-Blackwellization, assuming that the sign of the first-stage coefficient is known [Andrews and Armstrong, 2017]. We also show 2SLS for comparison, but note that 2SLS does not have the first moment in just-identified cases. Since this model is just-identified, GMM is numerically identical to 2SLS. In this model, we take \( \beta = 1 \) and \( \pi = 1/\sqrt{n} \). Similarly as before, the covariance matrix of \((u_i, v_i)\) is randomly determined for each of the 3 possible values of \( z_i \) and fixed afterwards. We call this the *model B* and iterate it for 5,000 times for two different specifications of heteroskedasticity.

To compute the Rao-Blackwellized versions of the four estimators, we derive the feasible GLS estimator of the reduced-form coefficients \((\pi \beta, \pi)\). The classical GLS is considered within a single-equation framework; to accommodate both the first-stage and second-stage equations, we map our reduced-form equations into one single-equation regression by

\[
\begin{bmatrix}
Y \\
\text{vec}(X)
\end{bmatrix} =
\begin{bmatrix}
\pi \beta \\
\pi
\end{bmatrix} +
\begin{bmatrix}
0 \\
\Pi_d \otimes Z
\end{bmatrix} \pi +
\begin{bmatrix}
u \\
\text{vec}(v)
\end{bmatrix} =
\begin{bmatrix}
Z \\
\Pi_d \otimes Z
\end{bmatrix} \begin{bmatrix} \pi \beta \\
\pi
\end{bmatrix} +
\begin{bmatrix} u \\
\text{vec}(v)
\end{bmatrix}.
\]

Consequently, the conditional covariance matrix \( \Omega \) of the error terms has some nonzero off-diagonal elements. We estimate it using the OLS coefficients, and thereby compute the feasible GLS estimator for \((\pi \beta, \pi)\). Note that variances of the OLS and GLS
estimators are given by

\[ \text{Var}(\hat{\psi}_{\text{OLS}} \mid Z) = \tilde{Z}' \tilde{Z} \left( \sum_{i=1}^{n} e_{i}^{2} \tilde{z}_{i} \tilde{z}'_{i} \right)^{-1}, \quad \text{Var}(\hat{\psi}_{\text{GLS}} \mid Z) = (\tilde{Z}' \Omega^{-1} \tilde{Z})^{-1}. \]

Since the GLS estimator is efficient, by orthogonality we have

\[ \text{Var}(\hat{\psi}_{\text{OLS}} - \hat{\psi}_{\text{GLS}} \mid Z) = \text{Var}(\hat{\psi}_{\text{OLS}} \mid Z) - \text{Var}(\hat{\psi}_{\text{GLS}} \mid Z). \]

This can be estimated by its sample counterpart. In the event that this estimate is not positive semidefinite, we use the nearest positive semidefinite matrix implied by Higham (1988). With this, we compute the conditional expectations of the 2SLS and GMM estimators conditional on the GLS estimator using a Monte Carlo method with 100,000 draws from

\[ \left( \begin{array}{c} U_{\pi \beta} \\ U_{\pi} \end{array} \right) \sim N \left( \begin{bmatrix} \pi_{FGLS,n} \\ \pi_{FGLS,n} \end{bmatrix}, \text{Var}(\hat{\psi}_{\text{OLS}} - \hat{\psi}_{\text{GLS}} \mid Z) \right). \]

Given this, the Rao-Blac kwellized 2SLS estimator of \( \beta \) is given by

\[ \mathbb{E}_{U}[(\hat{\pi}_{FGLS,n} + U_{\pi})'(Z'Z)(\hat{\pi}_{FGLS,n} + U_{\pi})^{-1}(\hat{\pi}_{FGLS,n} + U_{\pi})'(Z'Z)(\hat{\pi}_{FGLS,n} + U_{\pi})]; \]

where \( \mathbb{E}_{U} \) denotes expectation with respect to \( (U_{\pi \beta}, U_{\pi}) \). The Rao-Blac kwellized two-step GMM estimator of \( \beta \) is given by

\[ \mathbb{E}_{U}[(\hat{\pi}_{FGLS,n} + U_{\pi})'(Z'Z)\hat{W}(U_{\pi \beta}, U_{\pi})(Z'Z)(\hat{\pi}_{FGLS,n} + U_{\pi})^{-1}[(\hat{\pi}_{FGLS,n} + U_{\pi})'(Z'Z)\hat{W}(U_{\pi \beta}, U_{\pi})(Z'Z)(\hat{\pi}_{FGLS,n} + U_{\pi})]]; \]

where

\[ \hat{W}(U_{\pi \beta}, U_{\pi}) = \mathbb{E}_{n}[(y - x'(\hat{\pi}_{FGLS,n} + U_{\pi})^{-1}(\hat{\pi}_{FGLS,n} + U_{\pi})^2 z z')]. \]

\footnote{Note that \( U_{\pi \beta} \) and \( U_{\pi} \) here are already denormalized by \( \sqrt{n} \).}
The Rao-Blackwellized Fuller estimator of $\beta$ is given by

$$
\mathbb{E}_U \left[ \left( \frac{C}{n} X'X + \left[ 1 - \frac{C}{n} \right] (\hat{\pi}_{FGLS,n} + U_\pi)'(Z'Z)(\hat{\pi}_{FGLS,n} + U_\pi) \right)^{-1} \left( \frac{C}{n} X'y + \left[ 1 - \frac{C}{n} \right] (\hat{\pi}_{FGLS,n} + U_\pi)'(Z'Z)(\hat{\pi}_{FGLS,n} + U_\pi) \right) \right].
$$

The Rao-Blackwellized unbiased estimator of $\beta$ does not require Monte Carlo computation of expectations; Andrews and Armstrong (2017) show uniqueness of the unbiased estimator; the Rao-Blackwellized version of the unbiased estimator constructed with the OLS estimator equals the unbiased estimator constructed with the GLS estimator

$$
\sqrt{n} \hat{\beta}_{FGLS,n} / \hat{\sigma}_{\pi,\beta,FGLS,n} \Phi(\sqrt{n} \hat{\beta}_{FGLS,n} / \hat{\sigma}_{\pi,\beta,FGLS,n}).
$$

Figure 2-1a is the histogram of 5,000 iterations of the 2SLS and Rao-Blackwellized 2SLS estimators in model A. The dotted line in the middle indicates the true value, i.e., $\beta = 0$. It shows that the distribution of the Rao-Blackwellized 2SLS estimator is more concentrated than the 2SLS estimator, illustrating the power of Rao-Blackwellization in the limit. Note that since Rao-Blackwellization does not affect its mean, both estimators have the same bias. Figure 2-1b is the histogram of the GMM and Rao-Blackwellized GMM estimators for the same run as Figure 2-1a. Their distributions are very close to their 2SLS counterparts. Figure 2-1c is the histogram of the Fuller, Rao-Blackwellized Fuller, and HFUL estimators for the same run. Due to the shrinkage property, the Rao-Blackwellized Fuller estimator is quite concentrated.

To connect to the claim of Theorem 2.2, consider two loss functions $\ell : \mathbb{R} \to \mathbb{R}$ such that $\ell(x) = x^2$ and $\ell(x) = |x|$. That is, the expected losses are measured by the mean squared error (MSE) and mean absolute error (MAE) of the estimator. The MSE of 2SLS in this simulation is 0.22 while that of Rao-Blackwellized 2SLS is 0.03; the “MSE” of GMM is 0.23 while that of Rao-Blackwellized GMM is 0.02; the MSE of Fuller is 0.07 while that of Rao-Blackwellized Fuller is 0.01 and that of HFUL is 0.22.

$^{30}$Although our simulation suggests that two-step GMM has as many moments as 2SLS, it is not known whether it is the case. Therefore, we place double quotes around $MSE$ and $MAE$ of GMM estimators.
Figure 2-1: Distributions of the 2SLS, GMM, and Fuller estimators, and their Rao-Blackwellization under heteroskedasticity (model A). Simulated with normal errors, 10,000 observations, and 5,000 iterations. Clusters at the boundaries indicate how many observations fall outside of the range.

The MAE of 2SLS is 0.36 while that of Rao-Blackwellized 2SLS is 0.16; the “MAE” of GMM is 0.36 while that of Rao-Blackwellized GMM is 0.13; the MAEs of Fuller, Rao-Blackwellized Fuller, and HFUL are 0.21, 0.07, and 0.42. Theorem 2.2 guarantees that the MSE of the Rao-Blackwellized version never exceeds that of the original one, at least asymptotically. It is, therefore, preferable to use a weakly efficient estimator if one is available.

The second simulation is the comparison of the unbiased estimator and its Rao-Blackwellization together with Fuller. For this purpose we use a just-identified model,
Figure 2-2: Distributions of the unbiased, Rao-Blackwellized unbiased, 2SLS, Fuller, Rao-Blackwellized Fuller, and HFUL estimators under heteroskedasticity (model B). Simulated with normal errors, 10,000 observations, and 5,000 iterations. Clusters at the boundaries indicate how many observations fall outside of the range.

since the closed-form expression is available in \cite{Andrews2017}.

\footnote{Unbiased estimation itself is possible in overidentified models \cite{Andrews2017}.}

We also present the 2SLS estimator for the sake of comparison, although 2SLS is not subject to Rao-Blackwellization due to the lack of its first moment.\footnote{Technically, 2SLS may have moments conditional on GLS, in which case Rao-Blackwellization makes some sense. Numerical Rao-Blackwellization indicates it can affect numerical losses either upward or downward.} Note that the GMM estimator coincides with the 2SLS estimator since it is just-identified. Figures 2-2a and 2-2b are the histograms of 5,000 iterations of the unbiased and Rao-Blackwellized unbiased, and 2SLS estimators for one case of heteroskedasticity.
Table 2.1: Mean absolute errors (MAEs) and mean squared errors (MSEs).

<table>
<thead>
<tr>
<th></th>
<th>Model A (1)</th>
<th>Model A (2)</th>
<th>Model A (3)</th>
<th>Model B (2)</th>
<th>Model B (3)</th>
<th>Model B (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MSE</td>
<td>MAE</td>
<td>MSE</td>
<td>MAE</td>
<td>MSE</td>
</tr>
<tr>
<td>2SLS</td>
<td>0.42</td>
<td>0.36</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
</tr>
<tr>
<td>Rao-Blackwellized 2SLS</td>
<td>0.19</td>
<td>0.04</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
</tr>
<tr>
<td>GMM</td>
<td>“0.46”</td>
<td>“0.61”</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
</tr>
<tr>
<td>Rao-Blackwellized GMM</td>
<td>“0.17”</td>
<td>“0.03”</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
<td>∞</td>
</tr>
<tr>
<td>Unbiased</td>
<td>—</td>
<td>∞</td>
<td>1.15</td>
<td>∞</td>
<td>5.28</td>
<td>∞</td>
</tr>
<tr>
<td>Rao-Blackwellized unbiased</td>
<td>—</td>
<td>∞</td>
<td>1.04</td>
<td>∞</td>
<td>1.59</td>
<td>∞</td>
</tr>
<tr>
<td>Fuller</td>
<td>0.21</td>
<td>0.07</td>
<td>1.357</td>
<td>1.94</td>
<td>1.31</td>
<td>2.47</td>
</tr>
<tr>
<td>Rao-Blackwellized Fuller</td>
<td>0.07</td>
<td>0.01</td>
<td>1.356</td>
<td>1.88</td>
<td>1.06</td>
<td>1.27</td>
</tr>
<tr>
<td>HFUL</td>
<td>0.42</td>
<td>0.22</td>
<td>1.62</td>
<td>2.85</td>
<td>1.41</td>
<td>2.64</td>
</tr>
</tbody>
</table>

Number of observations: 10,000 10,000 10,000 10,000 10,000 10,000
Number of iterations: 5,000 5,000 5,000 5,000 5,000 5,000

* (1) Randomly generated heteroskedasticity for Figures 2-1a to 2-1c; (2) for Figures 2-2a and 2-2b; (3) for Figures 2-2c and 2-2d. Quotes indicate that their finiteness is not known.

Blackwellized unbiased estimators, and of the Fuller and Rao-Blackwellized Fuller estimators in model B with one type of heteroskedasticity which shows a slight improvement by LAR, and Figures 2-2c and 2-2d are the histograms with another type of heteroskedasticity. Figures 2-2c and 2-2d show that some heterogeneity can admit a substantial improvement to the Rao-Blackwellized unbiased estimator. Also, the figures show that performances of unbiased and Fuller estimators (in terms of MAE) depend on the form of heteroskedasticity. Note that Fuller estimator is biased under our weak identification asymptotics. The dotted lines at the center represent the true value, i.e., \( \beta = 1 \). With this specification \( (d = k = 1) \), the unbiased estimator does not have a second moment (Andrews and Armstrong 2017), so we use only the MAE as the measure of dispersion. The 2SLS estimators do not even have first moments, but we show their sample MAEs in the figure, just for the sake of comparison. Although in Figure 2-2a, the distribution of the Rao-Blackwellized version does not necessarily look more concentrated than the original estimator, the MAE of the Rao-Blackwellized version is 1.0 while that of the original one is 1.1. Table 2.1 gives losses of other estimators; it confirms Theorem 2.2 by showing reduction in convex losses from Rao-Blackwellization.

171
Note that the conditional moment assumption $E[u_i \mid z_i] = 0$ and $E[v_i \mid z_i] = 0$ plays a crucial role in this exercise. GLS is more efficient than OLS because this assumption holds; if we only have the unconditional moment restriction $E[u_i z'_i] = 0$ and $E[v_i z'_i] = 0$, then the GLS estimator is not consistent to what this condition defines. Another important assumption is the availability of the efficient estimator of the reduced-form coefficients. A notable example in which the form of heteroskedasticity is known \emph{a priori} is when $y_i$ is binary and one has a conditional moment restriction $E[y_i \mid x_i] = f(x_i)$; the form of heteroskedasticity is uniquely determined by $f$ as $E[(y_i - f(x_i))^2 \mid x_i] = f(x_i) - f(x_i)^2$ using $y_i^2 = y_i$. In case $f$ can be estimated, for example for being linear, one may use feasible GLS with no additional loss of generality. In other linear models without the knowledge of the form of heteroskedasticity, feasible GLS with a nonparametric estimator of the conditional covariance structure is available under various assumptions \cite{Carroll1982, Robinson1987, Newey1994}. See also \cite{Romano2017} for recent reinvestigation of the use of GLS in practice.

\section{Conclusion}

This chapter studies weak identification in semiparametric models and investigates its consequences and implications to efficient estimation. First, we show that weak identification can be captured by the notion of \emph{weak regularity} with which the parameter value itself depends on the score asymptotically. This dependence is homogeneous of degree zero and nonlinear, and we show, as a direct consequence, the impossibility of consistent estimation and inference and equivariant estimation. Then, we show that for each weakly regular parameter there exists an underlying parameter that is regular and fully characterizes the weakly regular parameter locally. As underlying regular parameters are not unique, we stipulate two desirable properties of an underlying parameter, namely \emph{sufficiency} and \emph{minimality}. The minimal sufficient underlying regular parameter contains the necessary and sufficient amount of information of the weakly regular parameter in the tangent space.
Regarding the estimation problem of a weakly regular parameter as the estimation problem of the minimal sufficient underlying parameter plus its transformation, we argue that the “efficiency” of the final estimator of a weakly regular parameter can be cast in terms of the involvement of noise in the estimator of the underlying parameter. When such noise exists, we can construct its improvement by taking a conditional expectation of the estimator, shown as local asymptotic Rao-Blackwellization. Intuitively, this exploits the property that an efficient estimator of a regular parameter is “asymptotically sufficient” and applies the Rao-Blackwell theorem to the asymptotic representations of the local expansion.\footnote{Cattaneo et al. (2012) also exploits “asymptotic sufficiency” of efficient estimators in semiparametric models. See also Le Cam and Yang (2000) and Van der Vaart (2002) for related discussion of “asymptotic sufficiency” in parametric models.} We then carry out simulation studies of a linear IV model and demonstrate that the 2SLS, GMM, and Fuller estimators as well as the unbiased estimator of Andrews and Armstrong (2017) can be made more concentrated given the availability of a (feasible) GLS estimator of the reduced-form coefficients.
Appendices

2.A Proofs

Proof of Lemma 2.1. Since $\dot{\mathcal{P}}_P$ is assumed to be linear, if $g \in \dot{\mathcal{P}}_P$ then $ag \in \dot{\mathcal{P}}_P$ for every $a \in \mathbb{R}$. If $g$ is induced by a path $t \mapsto Q_t$ and $a > 0$, then $ag$ can be induced by the path $t \mapsto Q_{at}$, which is the same path up to a scaled index. Therefore, if $Q_t \in \mathcal{P}_P \setminus \mathcal{P}_{P,\beta}$ then $Q_{at} \in \mathcal{P}_P \setminus \mathcal{P}_{P,\beta}$, implying that if $g \in \dot{\mathcal{P}}_P \setminus \dot{\mathcal{P}}_{P,\beta}$ then $ag \in \dot{\mathcal{P}}_P \setminus \dot{\mathcal{P}}_{P,\beta}$. Being defined as a difference between a linear space and a cone, $\dot{\mathcal{P}}_{P,\beta}$ is a cone. ■

Proof of Theorem 2.2. Let $\beta : \mathcal{P}_\beta \to \mathbb{B}$ be weakly regular and $\beta_P$ nonconstant.

The first assertion. Suppose that $\hat{\beta}_n : \mathcal{X}^n \to \mathbb{B}$ is a consistent sequence of estimators, or even weaker, that there exist two paths $Q_{n1}, Q_{n2} \in \mathcal{P}_{P,\beta}$ inducing $g_1, g_2 \in \dot{\mathcal{P}}_{P,\beta}$ such that $\beta_P(g_1) \neq \beta_P(g_2)$ and

$$\hat{\beta}_n \xrightarrow{Q_{nj}^*} \beta_P(g_j) \quad \text{under each} \quad Q_{nj} \in \{Q_{n1}, Q_{n2}\}.$$ 

Define $2\varepsilon := \|\beta_P(g_1) - \beta_P(g_2)\|_\mathbb{B}$. Denote by $Q_{nj}^n$ the product measure of $Q_{nj}$ on the product sample space $\mathcal{X}^n$. By the portmanteau theorem (Van der Vaart and Wellner 1996 Theorem 1.3.4) and the assumption of convergence in outer probability,

$$\limsup_{n \to \infty} Q_{n1}^n(\|\hat{\beta}_n - \beta_P(g_1)\|_\mathbb{B} \geq \varepsilon) \leq 0.$$
while
\[ \liminf_{n \to \infty} Q_{n2}^n(\|\hat{\beta}_n - \beta_P(g_1)\|_B^* \geq \varepsilon) \geq \liminf_{n \to \infty} Q_{n2}^n(\|\hat{\beta}_n - \beta_P(g_1)\|_{B,*} > \varepsilon) \geq 1. \]

Therefore, $Q_{n2}^n$ is not contiguous to $Q_{n1}^n$. Being paths, however, $Q_{n2}^n$ must be contiguous to $P^n$ and $P^n$ to $Q_{n1}^n$ (Van der Vaart and Wellner 1996, Lemma 3.10.11 and Theorem 3.10.9), hence a contradiction.

The second assertion. Let $H_0 : \beta \in \mathbb{B}_0$ and $H_1 : \beta \in \mathbb{B}_1$ be the null and alternative hypotheses such that $\mathbb{B}_0$ and $\mathbb{B}_1$ are nonempty. Suppose that $\phi_n : \mathcal{X}^n \to [0, 1]$ is a consistent sequence of tests of $H_0$ of level $\alpha < 1$ so that there exist two paths $Q_{n0}, Q_{n1} \in \mathcal{P}_{P,\beta}$ with $\beta_P(g_0) \in \mathbb{B}_0$ and $\beta_P(g_1) \in \mathbb{B}_1$ such that $\phi_n \to Q_{n0*} \alpha$ and $\phi_n \to Q_{n1*} 1$. Then by the same reasoning a contradiction follows.

The third assertion. Let $\hat{\beta}_n$ be an equivariant-in-law sequence of estimators of $\beta$ with a separable limit law, that is, there exists a fixed separable Borel probability measure $L$ on $\mathbb{B}$ such that

\[ \hat{\beta}_n - \beta(Q_n) \overset{Q_n}{\to} L \quad \text{for every} \quad Q_n \in \mathcal{P}_{P,\beta}. \]

Pick $g_1, g_2 \in \hat{P}_{P,\beta}$ such that $\beta_P(g_1) \neq \beta_P(g_2)$. Denote $\beta_1 := \beta_P(g_1)$ and $\beta_2 := \beta_P(g_2)$. Since $\hat{P}_{P,\beta}$ is a cone (Lemma 2.1), $ag_1$ and $ag_2$ are also in $\hat{P}_{P,\beta}$ for every $a > 0$ and by homogeneity we have $\beta_P(ag_1) = \beta_1$. For each positive integer $k$, take $Q_{nk1}, Q_{nk2} \in \mathcal{P}_{P,\beta}$ to be paths that induce scores $g_1/k$ and $g_2/k$. Let $d_{Q_n}$ denote the metric that metrizes weak topology on $\mathbb{B}$ under $Q_n$ toward separable limits (Van der Vaart and Wellner 1996, p. 73). For each $k$, let $n_k$ be such that for every $n \geq n_k$,

\[ \int_{\mathcal{X}} \left[ \frac{dQ_{nk1}^{1/2} - dP^{1/2}}{1/\sqrt{n}} - \frac{g_1}{2k} dP^{1/2} \right]^2 \vee \int_{\mathcal{X}} \left[ \frac{dQ_{nk2}^{1/2} - dP^{1/2}}{1/\sqrt{n}} - \frac{g_2}{2k} dP^{1/2} \right]^2 < \frac{1}{k}, \]

\[ d_{Q_{nk1}}(\hat{\beta}_n - \beta(Q_{nk1}), L) \vee d_{Q_{nk2}}(\hat{\beta}_n - \beta(Q_{nk2}), L) < \frac{1}{k}. \]

Then one can take $n'_k$ so that $n'_k \geq n_k$ and $n'_{k+1} > n'_k$ for every $k$. Construct two paths $Q'_{n1}$ and $Q'_{n2}$ by $Q'_{nj} = Q_{nk_nj}$ where $k_n$ satisfies $n'_{k_n} \leq n < n'_{k_n+1}$. Then $Q'_{nj} \to^{DQM} P$.
with scores equal to zero and $\hat{\beta}_n - \beta(Q'_{nj})$ converges weakly to $L$ under $Q'_{nj}$. Now we want to show that $dQ'_{n2}^m / dQ'_{n1}$ converges to 1 and invoke Le Cam’s third lemma. For this, we adopt the same proof strategy as Van der Vaart (1998, Theorem 7.2). Observe that

$$\mathbb{E}_{Q'_{n1}} \left[ n \left( 1 - \frac{dQ'_{n2}^{1/2}}{dQ'_{n1}^{1/2}} \right)^2 \right] \leq \int_X \left[ \frac{dQ'_{n1}}{dQ'_{n1}^{1/2}} - \frac{dQ'_{n2}}{dQ'_{n2}^{1/2}} \right]^2 \to 0.$$  

By Taylor’s theorem, $\log x^2 = -2(1-x) - (1-x)^2 + (1-x)^2 R(1-x)$ for some function $R : \mathbb{R} \to \mathbb{R}$ such that $R(1-x) \to 0$ as $x \to 1$. Then,

$$\log \frac{dQ'_{n2}}{dQ'_{n1}}(X_1, \ldots, X_n) = \log \left( \frac{dQ'_{n2}}{dQ'_{n1}}(X_1) \cdots \frac{dQ'_{n2}}{dQ'_{n1}}(X_n) \right) = \sum_{i=1}^n \log \frac{dQ'_{n2}}{dQ'_{n1}} = -2 \sum_{i=1}^n W_{ni} - \sum_{i=1}^n W^2_{ni} + \sum_{i=1}^n W^2_{ni} R(W_{ni}),$$

where $W_{ni} := 1 - dQ'_{n2}^{1/2} / dQ'_{n1}^{1/2}$. We argue that all three terms converge to zero in probability. Under $Q'_{n1},$

$$\left| \mathbb{E} \sum_{i=1}^n W_{ni} \right| = n \left| 1 - \int \frac{dQ'_{n2}^{1/2}}{dQ'_{n1}^{1/2}} dQ'_{n1} \right| \leq \frac{1}{2} \int \left[ \frac{dQ'_{n1}}{dQ'_{n1}^{1/2}} - \frac{dQ'_{n2}}{dQ'_{n2}^{1/2}} \right]^2 \to 0,$$

$$\text{Var} \left( \sum_{i=1}^n W_{ni} \right) \leq \mathbb{E}[nW^2_{ni}] = \mathbb{E} \left[ n \left( 1 - \frac{dQ'_{n2}^{1/2}}{dQ'_{n1}^{1/2}} \right)^2 \right] \to 0.$$

These results imply that the expectation and variance of $\sum W_{ni}$ converge to zero; hence it converges to zero in probability. The second result implies that $nW^2_{ni}$ converges to zero in mean; by the law of large numbers $\sum W^2_{ni}$ converges to zero in probability. By Markov’s inequality,

$$\Pr \left( \max_{1 \leq i \leq n} |W_{ni}| > \varepsilon \right) \leq n \Pr(|W_{ni}| > \varepsilon) \leq n \Pr(nW^2_{ni} > n\varepsilon^2) \leq \frac{\mathbb{E}[nW^2_{ni}]}{\varepsilon^2} \to 0$$

for every $\varepsilon > 0$. Thus, $\max_{1 \leq i \leq n} |W_{ni}|$ converges to zero in probability, meaning that $\max_{1 \leq i \leq n} |R(W_{ni})|$ converges to zero in probability as well. Therefore, the third term $\sum W^2_{ni} R(W_{ni})$ converges to zero in probability. We conclude that $dQ'_{n2}^m / dQ'_{n1}^m.$
converges to 1 in probability under $Q_{n1}'$. Since $L$ is separable, by Slutsky’s lemma (Van der Vaart and Wellner 1996, Example 1.4.7),

$$
\left( \hat{\beta}_n, \frac{dQ_{n2}'}{dQ_{n1}'} \right) \xrightarrow{Q_{n1}'} (\beta_1 + L, 1).
$$

By Le Cam’s third lemma (Van der Vaart and Wellner 1996, Theorem 3.10.7), we have

$$(\beta_2 + L)(B) = \mathbb{E} \left\{ \beta_1 + L \in B \right\} 1 = (\beta_1 + L)(B)$$

for every Borel $B \subset \mathbb{B}$, which contradicts $\beta_1 \neq \beta_2$. ■

Proof of Lemma 2.3. Denote by $\mathbb{D}$ the Banach space $L_2(P)$ of square integrable functions on $\mathcal{X}$ and define $\psi : \mathcal{P} \to \mathbb{D}$ by $\psi(Q) = dQ^{1/2}/dP^{1/2}$. Note that $\psi$ is regular with the derivative $\dot{\psi}_P : \dot{\mathcal{P}}_P \to \mathbb{D},$

$$
\dot{\psi}_P g = g,
$$

which is apparently linear and continuous. Thus, we have $\beta_{P,\psi} = \beta_P$. ■

Proof of Lemma 2.1. (i) Trivially, $0 \in N(\beta_P)$. The definition of the kernel implies that if $\tilde{g} \in N(\beta_P)$, then $-\tilde{g} \in N(\beta_P)$. Take $\tilde{g} \in N(\beta_P)$ and $a > 0$. Since $\mathcal{P}_{P,\beta}$ is a cone (Lemma 2.1) and $\beta_P$ is homogeneous of degree zero, $\beta_P(g) = \beta_P(g/a) = \beta_P(g/a + \tilde{g}) = \beta_P(g + a \tilde{g})$ for every $g \in \mathcal{P}_{P,\beta}$. This means that $a \tilde{g} \in N(\beta_P)$. Therefore, $N(\beta_P)$ is linear.

(ii) If $P \in \mathcal{P} \setminus \mathcal{P}_\beta$, then $0 \notin \mathcal{P}_{P,\beta}$. Since $g \in N(\beta_P) \cap \mathcal{P}_{P,\beta}$ implies $\beta_P(g) = \beta_P(g - g) = \beta_P(0)$, $N(\beta_P) \cap \mathcal{P}_{P,\beta}$ must be empty.

(iii) If $\Pi_\beta g = 0$ then $g \in N(\beta_P)$, which implies $g \notin \mathcal{P}_{P,\beta}$ given that $P \in \mathcal{P} \setminus \mathcal{P}_\beta$ by (ii). ■

Proof of Theorem 2.2. Let $\mathbb{D} = L_2(P)$ and define $\psi : \mathcal{P} \to \mathbb{D}$ by $\psi(Q) = 2\Pi_\beta dQ^{1/2}/dP^{1/2}$. Then $\psi$ is regular with the derivative $\dot{\psi}_P : \dot{\mathcal{P}}_P \to \mathbb{D},$

$$
\dot{\psi}_P g = \Pi_\beta g.
$$
Note that $\beta_P(g) = \beta_P(\Pi_\beta g)$. This implies that $\psi$ is an underlying regular parameter for $\beta$ and that $N(\dot{\psi}_P) = N(\beta_P)$, which implies minimal sufficiency of $\psi$. □

**Proof of Theorem 2.3** Sufficiency. Assume that for any sufficient underlying regular parameter $\phi : \mathcal{P} \to \mathbb{E}$ for $\beta$ there exists a map $\tau : \mathbb{E} \to \mathbb{D}$ such that $\tau(\dot{\phi}_P g) = \dot{\psi}_P g$ for every $g \in \dot{\mathcal{P}}_{P,\beta}$. This means that $N(\dot{\phi}_P) \subset N(\dot{\psi}_P)$. Take $\phi$ to be minimal; then $N(\beta_P) = N(\dot{\phi}_P) \subset N(\dot{\psi}_P)$. On the other hand, since $\dot{\psi}$ is assumed to be a sufficient underlying parameter, we have $N(\beta_P) \supset N(\dot{\psi}_P)$.

**Necessity.** Assume that $\dot{\psi} : \mathcal{P} \to \mathbb{D}$ is a minimal sufficient underlying regular parameter for $\beta$. Take $\phi : \mathcal{P} \to \mathbb{E}$ to be another sufficient underlying regular parameter for $\beta$. Then $\beta_{P,\psi}(\dot{\psi}_P g) = \beta_{P,\phi}(\dot{\phi}_P g)$ for every $g \in \dot{\mathcal{P}}_{P,\beta}$ and $N(\dot{\psi}_P) = N(\beta_P) \supset N(\dot{\phi}_P)$. The first property implies $\dot{\psi}_P g \in \beta_{P,\psi}^{-1}(\dot{\phi}_P g)$ for every $g \in \dot{\mathcal{P}}_{P,\beta}$. The second property implies that if $\dot{\phi}_P g_1 = \dot{\phi}_P g_2$ then $\dot{\psi}_P g_1 = \dot{\psi}_P g_2$. Conclude that there exists a linear map $\tau : \mathbb{E}_0 \to \mathbb{D}$ such that $\dot{\psi}_P g = \tau(\dot{\phi}_P g)$ for $g \in \dot{\mathcal{P}}_0$ where $\mathbb{E}_0 := \dot{\phi}_P(\dot{\mathcal{P}}_{P,\beta})$. One can extend $\tau$ on the whole of $\mathbb{E}$ by letting $\tau(e) := \tau(\Pi_{\mathbb{E}_0} e)$. □

**Proof of Proposition 2.1.** Define $T_{\delta,n} : \mathbb{D} \times [0,1] \to \mathbb{B}$ by

$$T_{\delta,n}(\delta, u) := T_n\left(\delta + \frac{\delta}{\sqrt{n}} u\right).$$

Then, $\hat{\beta}_n = T_{\psi(P),n}(\sqrt{n}(\hat{\psi}_n - \psi(P)), U) + o_P(1)$. By the extended continuous mapping theorem [Van der Vaart and Wellner 1996] Theorem 1.11.1 and Problem 1.11.1), the claim follows. □

**Proof of Theorem 2.2** Observe that $\hat{T}_n$ is locally continuously root-\(n\) approximable at $\psi(P)$ tangentially to the range of $\dot{\psi}_P g + L_\psi$ with the approximating function
\[ T_{\psi(P)}(\delta) := \mathbb{E}[T_{\psi(P)}(\delta + L_\eta, U)]; \text{ hence } T_n(\hat{\psi}_n) \text{ is regular.} \]

For \( Q_n \in \mathcal{P}_{P,\beta} \), write

\[
\mathbb{E}_n[\ell(\hat{\beta}_n - \beta)] - \mathbb{E}^*[\ell(T_n(\hat{\psi}_n) - \beta)] \\
= \mathbb{E}_n[\ell(\hat{\beta}_n - \beta)] - \mathbb{E}[\ell(T_{\psi(P)}(\hat{\psi}_Pg + L_\psi + L_\eta, U) - \beta)] \\
+ \mathbb{E}[\mathbb{E}[\ell(T_{\psi(P)}(\hat{\psi}_Pg + L_\psi + L_\eta, U) - \beta) - \ell(T_{\psi(P)}(\hat{\psi}_Pg + L_\psi) - \beta) | L_\psi]] \\
+ \mathbb{E}[\ell(T_{\psi(P)}(\hat{\psi}_Pg + L_\psi) - \beta) - \mathbb{E}^*[\ell(T_n(\hat{\psi}_n) - \beta)].
\]

The first difference converges to zero by Proposition 2.1 and Van der Vaart and Wellner (1996, Theorem 1.11.3); the second difference is nonnegative since the inner conditional expectation is nonnegative by a generalized Jensen’s inequality (To and Yip, 1975); the third difference converges to zero by approximability of \( T_n \), the extended continuous mapping theorem (Van der Vaart and Wellner, 1996, Theorem 1.11.1 and Problem 1.11.1), and Van der Vaart and Wellner (1996, Theorem 1.11.3). ■

2. A.1 Auxiliary Results

As weakly regular parameters are transformations of the local parameters of regular parameters, estimators of weakly regular parameters are as well transformations of estimators of the local parameters of regular parameters. The asymptotic distributions of such estimators, therefore, are granted by a continuous mapping theorem applied to regular parameters. Their transformations, however, can often be estimated by not known a priori, hence the need for a continuous mapping theorem that accounts for vanishing randomness that dwells in their estimators. We accomplish this by further extending the “extended continuous mapping theorem” in Van der Vaart and Wellner (1996, Theorem 1.11.1) to subsume random mappings. Requirement of uniform convergence on an open set is arguably strong, but it covers many empirically relevant examples we have discussed.

Addendum 2.1 (Further extended continuous mapping theorem). Let \((\mathbb{D}, d)\) and \((\mathbb{E}, e)\) be metric spaces, \(\mathbb{D}_n\) and \(\mathbb{D}_\infty\) subsets of \(\mathbb{D}\), and \(g_n : \mathbb{D}_n \rightarrow \mathbb{E}\) and \(g : \mathbb{D}_\infty \rightarrow \mathbb{E}\) functions such that if \(x_n \in \mathbb{D}_n\), \(x \in \mathbb{D}_\infty\), and \(x_n \rightarrow x\), then \(g_n(x_n) \rightarrow g(x)\). Let \(X_n\) be
maps with values in $\mathbb{D}_n$ and $X$ a separable Borel measurable map with values in $\mathbb{D}_\infty$. Let $\mathbb{G}_n$ be maps taking values as functions from $\mathbb{D}_n$ to $\mathbb{E}$ such that for every $\varepsilon > 0$, there exists an open set $D_\varepsilon \subset \mathbb{D}$ such that $\Pr(\varepsilon \in D_\varepsilon) \geq 1 - \varepsilon$ and

$$\sup_{\delta \in D_\varepsilon \cap \mathbb{D}_n} e(\mathbb{G}_n(\delta), g_n(\delta)) \xrightarrow{P*} 0 \quad \text{as} \quad n \to \infty.$$ 

Then, $X_n \Rightarrow X$ implies $\mathbb{G}_n(X_n) \Rightarrow g(X)$.

Proof. As in the proof of Van der Vaart and Wellner (1996, Theorem 1.11.1), we may assume without loss of generality that every point $x$ in $\mathbb{D}_\infty$ has a sequence $x_n \in \mathbb{D}_n$ converging to $x$ and that $g$ is continuous on $\mathbb{D}_\infty$. Let $F \subset \mathbb{E}$ be closed. Fix $\varepsilon > 0$ and let $D_\varepsilon \subset \mathbb{D}$ be the given open set. Then,

$$\Pr^*(\mathbb{G}_n(X_n) \in F) \leq \Pr^*(X_n \notin D_\varepsilon) + \Pr^*(X_n \in D_\varepsilon \text{ and } \mathbb{G}_n(X_n) \in F).$$

Since $D_\varepsilon$ is open, the portmanteau theorem implies

$$\limsup_{n \to \infty} \Pr^*(X_n \notin D_\varepsilon) \leq \Pr(X \notin D_\varepsilon) \leq \varepsilon.$$

Since $F$ is closed and $\mathbb{G}_n$ converges uniformly in outer probability to $g_n$ on $D_\varepsilon$,

$$\limsup_{n \to \infty} \Pr^*(X_n \in D_\varepsilon \text{ and } \mathbb{G}_n(X_n) \in F) = \limsup_{n \to \infty} \Pr^*(X_n \in D_\varepsilon \text{ and } g_n(X_n) \in F) \leq \limsup_{n \to \infty} \Pr^*(g_n(X_n) \in F) \leq \Pr(g(X) \in F)$$

by the extended continuous mapping theorem (Van der Vaart and Wellner 1996, Theorem 1.11.1). Letting $\varepsilon \to 0$, conclude

$$\limsup_{n \to \infty} \Pr^*(\mathbb{G}_n(X_n) \in F) \leq \Pr(g(X) \in F).$$

This completes the proof. ■

181
2.B General Weak Linear IV Models

This section analyzes the general linear IV model from Example 2.1 in which $\pi$ approaches a rank deficient matrix instead of zero. Recall

\[
\begin{cases}
    y_i = z_i'\psi_1 + u_i, & \mathbb{E}[z_iu_i] = 0, \\
x_i' = z_i'\psi_2 + v_i', & \mathbb{E}[z_i'v_i] = 0,
\end{cases}
\]

where $\beta = \psi_2^{-1}\psi_1$. We are interested in a path $Q_n$ that approaches a point of identification failure $P$ such that

\[
\begin{align*}
\psi_2(Q_n) &= \pi + \frac{\dot{\pi}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \\
\beta(Q_n) &= \beta + \frac{\dot{\beta}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \\
\psi_1(Q_n) &= \psi_2(Q_n)\beta(Q_n) = \pi\beta + \frac{\dot{\pi}\beta + \pi\dot{\beta}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right).
\end{align*}
\]

Apply the singular value decomposition to $\pi$ to write $\pi = USV'$ for a $k \times k$ orthogonal matrix $U$, a $k \times d$ nonnegative diagonal matrix $S$, and a $d \times d$ orthogonal matrix $V$. If $\beta$ is weakly regular, then the number $\ell$ of positive elements of $S$ is strictly less than $d$. The regression equations can then be written as

\[
\begin{cases}
y_i = \tilde{z}_i'\tilde{\psi}_1 + u_i, \\
\tilde{x}_i' = \tilde{z}_i'\tilde{\psi}_2 + \tilde{v}_i',
\end{cases}
\]

for $\tilde{\psi}_1 = \Lambda^{-1}U'\psi_1$, $\tilde{\psi}_2 = \Lambda^{-1}U'\psi_2 V$, $\tilde{\beta} = V'\beta$, $\tilde{z}_i = \Lambda U'z_i$, $\tilde{x}_i = V'x_i$, $\tilde{v}_i = V'v_i$, and $\Lambda^{-1}$ is a positive diagonal matrix such that the first $\ell$ diagonal entries are inverses of positive entries of $S$ and all other diagonal entries are one. Since $\psi_2(P) = \pi$ is known in the limit, we may assume without loss of generality that $\pi$ is a diagonal matrix whose first $\ell$ elements are one and all others zero.

Due to this structure of $\pi$, the first $\ell$ elements of $\beta$ are uniquely determined by $\psi_1(P) = \pi\beta$ and $\psi_2(P) = \pi$ regardless of the score. On the other hand, the remaining elements of $\beta$ must be determined by the local parameters $\hat{\psi}_{1,p}(g) = \hat{\pi}\beta + \pi\hat{\beta}$ and $\hat{\psi}_{2,p}(g) = \hat{\pi}$. For $\beta$ to be weakly regular, the remaining components of $\beta$ must be
uniquely determined by the score. Since $A\pi = 0$ for a $k \times k$ diagonal matrix $A$ whose first $\ell$ diagonal elements are zero and remaining elements one, if the last $d - \ell$ elements of $\beta$ are uniquely determined by $\dot{\gamma}_{2, P}(g) = \dot{\pi}$ and $A\dot{\gamma}_{1, P}(g) = A\dot{\pi}\beta$, then $\beta$ is weakly regular; this is the case when the lower bottom $(k - \ell) \times (d - \ell)$ matrix of $\dot{\pi}$ is of full column rank. We henceforth make this assumption.

In order to see when the first $\ell$ elements of $\beta$ become regular, we aim to represent the first $\ell$ elements of $\dot{\beta}$ as a continuous linear map of the score. Similarly as Example 2.1, the score is given by

$$g = g_{uz} - z'(\dot{\pi}\beta + \pi\dot{\beta}) \frac{dP_{uz,u}}{dP} - z'\pi \frac{dP_{uz,v}}{dP}$$

and we have

$$E_P[zu g] = E_P[zz'](\dot{\pi}\beta + \pi\dot{\beta}), \quad E_P[zv' g] = E_P[zz']\dot{\pi}.$$ 

Denote by $\beta_\ell$ and $\beta_{-\ell}$ the first $\ell$ and last $d - \ell$ elements of $\beta$ and

$$\dot{\pi} =:\begin{bmatrix} \dot{\pi}_1 & \dot{\pi}_3 \\ \dot{\pi}_2 & \dot{\pi}_4 \end{bmatrix}$$

where $\dot{\pi}_1$ is $\ell \times \ell$. Since $\beta_\ell$ does not depend on the score, we may take its value as granted. Then

$$E_P\left[z\left(u - u'\begin{bmatrix} \beta_\ell \\ 0 \end{bmatrix}\right)g\right] = E_P[zz']\left(\dot{\pi} \begin{bmatrix} 0 \\ \beta_{-\ell} \end{bmatrix} + \pi\dot{\beta}\right)$$

$$= E_P[zz']\left(\begin{bmatrix} 0 & \dot{\pi}_3 \\ 0 & \dot{\pi}_4 \end{bmatrix}\begin{bmatrix} 0 \\ \beta_{-\ell} \end{bmatrix} + \begin{bmatrix} I_\ell & 0 \\ 0 & 0 \end{bmatrix}\dot{\beta}\right).$$

Therefore, if $\dot{\pi}_3$ is zero, then with the $k \times k$ diagonal matrix $B$ whose first $\ell$ diagonal
elements are one and all others zero, one obtains

\[
B \mathbb{E}_P[zz']^{-1} \mathbb{E}_P \left[ z \left( u - v' \begin{bmatrix} \beta \ell \\ 0 \end{bmatrix} \right) g \right] = B \mathbb{E}_P[zz']^{-1} \pi \dot{\beta}.
\]

Thus, one can represent the first \( \ell \) components of \( \dot{\beta} \) as a linear functional of the score. Otherwise, one needs to know \( \beta_{-\ell} \) in order to eliminate \( \dot{\pi} \beta \), suggesting that \( \beta_{\ell} \) might not be regular.

Now we investigate the tangent spaces. With these notations, we have

\[
\beta_P(g) = \begin{bmatrix} \beta \ell \\ \hat{\pi}_4^{-1}(\hat{\pi}_4 \beta_{-\ell}) \end{bmatrix},
\]

where \( \hat{\pi}_4 \) and \( \hat{\pi}_4 \beta_{-\ell} \) are calculated from

\[
\hat{\pi} = \mathbb{E}_P[zz']^{-1} \mathbb{E}_P[zu'g], \quad \begin{bmatrix} 0 \\ \hat{\pi}_4 \beta_{-\ell} \end{bmatrix} = A \mathbb{E}_P[zz']^{-1} \mathbb{E}_P \left[ z \left( u - v' \begin{bmatrix} \beta \ell \\ 0 \end{bmatrix} \right) g \right],
\]

whence we may guess the minimal sufficient underlying parameter. Denote by \( \psi_{1,\ell} \) and \( \psi_{1,-\ell} \) the first \( \ell \) and last \( k - \ell \) components of \( \psi_1 \) and

\[
\psi_2 = \begin{bmatrix} \psi_{2,1} & \psi_{2,2} \\ \psi_{2,3} & \psi_{2,4} \end{bmatrix}
\]

for an \( \ell \times \ell \) submatrix \( \psi_{2,1} \). Then the parameters \( \psi_{1,-\ell} \) and \( \psi_{2,4} \) that induce local parameters \( \hat{\pi}_4 \beta_{-\ell} \) and \( \hat{\pi}_4 \) can be guessed to constitute the minimal sufficient underlying parameter. To compute the nuisance tangent space, recall the score formula

\[
g = g_{uwz} - z_\ell' (\hat{\pi}_1 \beta_\ell + \hat{\pi}_3 \beta_{-\ell}) \frac{dP_{uwz,u}}{dP} - z_{-\ell}' (\hat{\pi}_2 \beta_\ell + \hat{\pi}_4 \beta_{-\ell}) \frac{dP_{uwz,u}}{dP} - z_\ell' \pi \dot{\beta} \frac{dP_{uwz,u}}{dP} - (z_\ell' \hat{\pi}_1 + z_{-\ell}' \hat{\pi}_2) \frac{dP_{uwz,u,\ell}}{dP} - (z_\ell' \hat{\pi}_3 + z_{-\ell}' \hat{\pi}_4) \frac{dP_{uwz,u,-\ell}}{dP},
\]

where \( z_\ell \) and \( z_{-\ell} \) denote the first \( \ell \) and last \( k - \ell \) components of \( z \). It is clear that
adding the scores
\[ g_{uvz}, \quad z'\pi C \frac{dP_{uvz,u}}{dP}, \quad z'_\ell D \beta'_\ell \frac{dP_{uvz,u}}{dP} + z'_\ell D \frac{dP_{uvz,v,\ell}}{dP}, \quad z'_\ell F \beta'_\ell \frac{dP_{uvz,u}}{dP} + z'_\ell F \frac{dP_{uvz,v,\ell}}{dP} \]

for any \( d \times 1 \) vector \( C \), \( \ell \times \ell \) matrix \( D \), and \((k-\ell) \times \ell \) matrix \( E \) does not change the value of \( \beta_{-\ell} \); therefore, they are in \( N(\beta_P) \). Also, adding the score
\[ z'_\ell F \frac{dP_{uvz,v,-\ell}}{dP} \]

for any \( \ell \times (d-\ell) \) matrix \( F \) would change the values of \( \pi_3 \) and \( \hat{\beta} \) (depending on the value of \( \beta_{-\ell} \)) but does not change the value of \( \beta_{-\ell} \) itself; therefore, it is in \( N(\beta_P) \) as well.

\section*{2.C Partial Weak Identification}

It has been observed that weak identification and partial identification in linear IV models arise from similar sources \cite{PoskittandSkeels}. Indeed, some instances of weak identification in the literature has a flavor of partial identification. For instance, \cite{AndrewsandMikusheva} argue in their supplementary material that a simple stylized DSGE model can asymptote to a point where only four out of the six parameters are identified; \cite{Cox} finds in extremum estimation models that different asymptotics is needed for what he calls “super-weak sequences of parameters.”

In this section, we explain how these cases involve partial identification in the local expansion. Partial identification in the local expansion means that given the model score, which summarizes all information of the first-order approximation of the model, the value of the parameter of interest is not pinned down uniquely.

\textbf{Example 2.1} (Linear IV with partial weak identification, continuing from p. 165). If the first-stage coefficients converge to degeneracy faster than root-\( n \), the structural parameter \( \beta \) suffers partial identification in the limit. Let \( d = k = 2 \) and \((u,v) \sim \ldots \)
$N(0, I_3)$, and consider the embedding
\[
\pi_n = \left( \frac{1}{\sqrt{n}} + \frac{1}{n}, \frac{1}{\sqrt{n}} + \frac{1}{n} \right).
\]

The first-stage coefficients $\pi_n$ are of full rank at each $n$ and approach zero with rate $\sqrt{n}$, and $\sqrt{n}\pi$ approaches a degenerate matrix $\pi_0 := (\frac{1}{1} \frac{1}{1})$ with rate $n$. The corresponding path is given by
\[
dQ_n = \frac{1}{(2\pi)^{3/2}} \exp\left(-\frac{(y' - z'\pi_n\beta)^2 + (x' - z'\pi_n)(x' - z'\pi_n)'}{2}\right).
\]

From here, the score can be calculated as
\[
\sqrt{n} \frac{dQ_n - dP}{dP} \to g = z'\pi_0\beta y + z'\pi_0 x = z' \left( \frac{\beta_1 + \beta_2}{\beta_1 + \beta_2} \right) y + z' \left( \begin{array}{c} 1 \\ 1 \end{array} \right) x.
\]

Therefore, only the combination $\beta_1 + \beta_2$ is identified as a function of the score, but not $\beta_1$ and $\beta_2$ separately. \hfill \square

**Example 2.10 (Full information DSGE model).** Consider the simple DSGE model in [Andrews and Mikusheva (2014)](Andrews2014) and [Andrews and Mikusheva (2016a Supplementary Appendix)](Andrews2016a):\[
\begin{align*}
& b\mathbb{E}_t\pi_{t+1} + \kappa x_t - \pi_t = 0, & \text{(Phillips curve)} \\
& r_t - \mathbb{E}_t\pi_{t+1} - \rho \Delta a_t = \mathbb{E}_t x_{t+1} - x_t, & \text{(Euler equation)} \\
& b^{-1}\pi_t + u_t = r_t, & \text{(Monetary policy)}
\end{align*}
\]

where the shock processes follow \[
\begin{align*}
\Delta a_t &= \rho \Delta a_{t-1} + \varepsilon_{a,t}, \\
u_t &= \delta u_{t-1} + \varepsilon_{u,t}, \\
\begin{pmatrix}
\varepsilon_{a,t} \\
\varepsilon_{u,t}
\end{pmatrix} &\sim N \left( 0, \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_u^2 \end{bmatrix} \right) \text{ i.i.d.}
\end{align*}
\]
These equations reduce to

\[ m_{t1} := \frac{\varepsilon_{a,t}}{\sigma_a} = \frac{1}{\sigma_a} \left( \frac{1 - \rho b}{b^2 \kappa} \right) \frac{(b + \kappa - \rho b)}{\rho (\rho - \delta)} \left[ \pi_t - \rho \pi_{t-1} - \frac{\kappa}{1 - \delta b} (x_t - \rho x_{t-1}) \right], \]

\[ m_{t2} := \frac{\varepsilon_{a,t}}{\sigma_u} = \frac{1}{\sigma_u} \left( \frac{1 - \rho b}{b^2 \kappa} \right) \frac{(b + \kappa - \rho b)}{\rho (\rho - \delta)} \left[ \pi_t - \delta \pi_{t-1} - \frac{\kappa}{1 - \delta b} (x_t - \delta x_{t-1}) \right], \]

which are i.i.d. We are concerned about identification when \( \rho \) is close to \( \delta \), that is, \( \rho_T = \delta + h/\sqrt{T} \). Other parameters are left unspecified; thus, we consider \( b_T = b + h_b/\sqrt{T} \), \( \kappa_T = \kappa + h_k/\sqrt{T} \), \( \sigma_{a,T} = \sigma_a + h_a/\sqrt{T} \), and \( \sigma_{u,T} = \sigma_u + h_u/\sqrt{T} \).

Since \((m_{t1}, m_{t2})\) is \( O_P(1) \), the terms in the square brackets stay \( O_P(1/\sqrt{T}) \). In light of this, define

\[ z_t := \frac{\pi_t - \delta \pi_{t-1} - \frac{\kappa}{1 - \delta b} (x_t - \delta x_{t-1})}{\sqrt{T}}. \]

Then the asymptotic representation of the bracket terms can be given by

\[ \pi_t - \rho_T \pi_{t-1} - \frac{\kappa_T (x_t - \rho_T x_{t-1})}{1 - \delta b} = \frac{z_t}{\sqrt{T}} - \frac{1}{\sqrt{T}} \left[ \frac{h_k}{\kappa} + \frac{\delta h_b}{1 - \delta b} \right] \frac{\kappa (x_t - \delta x_{t-1})}{1 - \delta b} - \frac{h}{\sqrt{T}} \left[ \frac{\pi_{t-1} - \frac{\kappa x_{t-1}}{1 - \delta b}}{1 - \delta b} \right] + o_P \left( \frac{1}{\sqrt{T}} \right), \]

\[ \pi_t - \delta \pi_{t-1} - \frac{\kappa_T (x_t - \delta x_{t-1})}{1 - \rho_T b} = \frac{z_t}{\sqrt{T}} - \frac{1}{\sqrt{T}} \left[ \frac{h b + \delta h_b}{1 - \delta b} \right] \frac{\kappa (x_t - \delta x_{t-1})}{1 - \delta b} - \frac{1}{\sqrt{T}} \frac{h_k \kappa (x_t - \delta x_{t-1})}{1 - \delta b} + o_P \left( \frac{1}{\sqrt{T}} \right). \]

The distribution of \( z_t \) will be uniquely determined by that of \((m_{t1}, m_{t2}) \sim O_T \). Let \( P \) be the limit distribution that yields \( \rho = \delta \in (0, 1) \) (but satisfies \( 1 - \delta b \neq 0 \), \( b + \kappa - \delta b \neq 0 \), \( b \neq 0 \), \( \kappa \neq 0 \), \( \sigma_a > 0 \), \( \sigma_u > 0 \)). Denote the weak limit of \((m_{t1}, m_{t2})\) by

\[ m_1 := \frac{1}{\sigma_a} \left( \frac{1 - \delta b)^2 b + \kappa - \delta b}{b^2 \kappa} \right) \delta \left[ \frac{z}{\delta h} - \pi_{t-1} + \frac{\kappa x_{t-1}}{1 - \delta b} \right], \]

\[ m_2 := \frac{1}{\sigma_u} \left( \frac{1 - \delta b)^2 (b + \kappa - \delta b)}{b^2 \kappa} \right) \left[ \frac{z}{\delta h} - \frac{b \kappa (x - \delta x_{t-1})}{(1 - \delta b)^2} \right]. \]
Finally, define the sequence of moments by

\[
m_{t1,T} := \frac{(1 - \rho_T b_T) (1 - \delta b_T) (b_T + \kappa_T - \rho_T b_T)}{\sigma_a T b_T^2 \kappa_T \rho_T (\rho_T - \delta)} \left[ \frac{1}{\pi_t - \rho_T \pi_{t-1} - \kappa_T (x_t - \rho_T x_{t-1})} \right],
\]

\[
m_{t2,T} := \frac{(1 - \rho_T b_T) (1 - \delta b_T) (b_T + \kappa_T - \delta b_T)}{\sigma_u T b_T^2 \kappa_T (\rho_T - \delta)} \left[ \frac{1}{\pi_t - \delta \pi_{t-1} - \kappa_T (x_t - \delta x_{t-1})} \right],
\]

After tedious algebra, one can compute the score for this general path as

\[
\sqrt{T} [dQ_T(m_{t1,T}, m_{t2,T}) - dP(m_1, m_2)] \rightarrow g = g_Q + \frac{dP_1}{dP} m_1 \left[ -\frac{h_a}{\sigma_a} - \frac{bh + 2h b}{1 - \delta b} \delta \frac{h_k}{b + \kappa - \delta b} - \frac{h}{\delta} \right] + \frac{dP_1}{dP} A \left[ -\frac{1}{h} \left( \frac{h_k}{\kappa} + \frac{\delta h_b}{\delta b} \right) \kappa (x - \delta x_{t-1}) \right] + \frac{dP_2}{dP} m_2 \left[ -\frac{h_a}{\sigma_u} - \frac{bh + 2h b}{1 - \delta b} \delta \frac{h_k}{b + \kappa - \delta b} + \frac{h_b + h_k - \delta h_b}{\delta b} \right] + \frac{dP_2}{dP} B \left[ -\frac{1}{h} \left( \frac{h_k}{\kappa} + \frac{\delta h_b}{\delta b} \right) \kappa (x - \delta x_{t-1}) \right] := g_Q + \frac{dP_1}{dP} m_1 H_1 + \frac{dP_1}{dP} (x - \delta x_{t-1}) H_A + \frac{dP_2}{dP} m_2 H_2 + \frac{dP_2}{dP} (x - \delta x_{t-1}) H_B.
\]

Thus, by knowing the model score \( g \) we can recover only up to four local parameters \( H_1, H_2, H_A, \) and \( H_B \) but not all of six parameters; there is no map \( \beta_P : \hat{P}_{P,\beta} \rightarrow \mathbb{R}^6 \) that recovers six parameters. \( \square \)
Bibliography


Chapter 3

Controlling Tail Risk Measures with Estimation Error

Assessment of risk and its control play an important role in investment decision making, financial regulations, actuarial science, and operations research. In practice, accuracy of estimated risk is subject to estimation error. While estimation error can be estimated in many cases, it remains a question as to how the error thus estimated can be incorporated into actual control of the true but unobservable risk. We propose the class of risk measures called the tail risk measures that give the upper bounds below which the quantities of interest fall with probability at least as much as a pre-specified confidence level. We show that a simple rule based on the Bonferroni inequality can control a tail risk measure at a desired level even when the true risk is unknown and needs to be estimated. Most popular risk measures such as Value-at-Risk and expected shortfall are interpreted as tail risk measures. For coherent tail risk measures, the true risk of any combination of assets can be controlled by knowledge of estimated risk and estimated error of individual assets. Empirical applications illustrate how the proposed concept can be applied to practical risk control problems.

This chapter is a joint work with Hyungjune Kang.
3.1 Introduction

Assessment of risk and its control play an important role in finance, management, actuarial science, and operations research. In the context of investment decision making, Markowitz (1952) and Roy (1952) developed theory of optimal investment decisions that take into account risk. Value-at-Risk (VaR) has been adapted in Wall Street since the 1970’s. Artzner et al. (1999) proposed the concept of coherence as a desirable property of risk measures. Rockafellar and Uryasev (2000) proposed expected shortfall (ES), a coherent risk measure that is now widely used in banking regulations. ES is generalized to distortion risk measures, a class of risk measures that are coherent (Acerbi, 2002). Ahmadi-Javid (2012) propose another coherent risk measure that gives an upper bound on VaR and ES and has computational advantages. Various other risk measures are succinctly summarized in Dowd and Blake (2006).

Risk measures are often used as one of the key components in large systems. In investment decision making, there are studies on the portfolio optimization with VaR or ES constraints (Pflug, 2000; Campbell et al., 2001; Favre and Galeano, 2002; Pflug, 2000; Rockafellar and Uryasev, 2000; Ciliberti et al., 2007; Akçay and Yalçın, 2010). The equilibrium consequences of VaR constraints are also studied, both from the partial equilibrium standpoint (Basak and Shapiro, 2001; Cuoco and Liu, 2006; Cuoco et al., 2008) and from the general equilibrium standpoint (Danielsson et al., 2004). In banking regulations, the required capital of each bank is calculated based on VaR (Chen, 2014). However, since risk measures depend on the underlying unknown distributions of returns, we need to estimate them to be used in practice.

Estimation of risk measures is a widely studied area, ranging from parametric to nonparametric methods. Embrechts et al. (1997) consider estimation of VaR using extreme-value theory; Barone-Adesi et al. (1999) give filtered historical simulation to estimate VaR; Chen and Tang (2005) and Scaillet (2004) provide nonparametric

---

1 This quantity is originally named in finance conditional Value-at-Risk (Rockafellar and Uryasev, 2000). We use expected shortfall since conditional Value-at-Risk sometimes refers to a different concept in the literature (Chernozhukov and Umantsev, 2001; Wang and Zhao, 2016).
estimators of VaR and ES, respectively; Chen (2008) discuss performances of two major nonparametric estimation methods; Linton and Xiao (2013) and Hill (2015) consider nonparametric estimation of ES when the portfolio return may not have variance. There is also a large body of the literature on the estimation of VaR and ES conditional on covariates (Chernozhukov and Umantsev, 2001; Cai and Wang, 2008; Kato, 2012; Chernozhukov and Fernández-Val, 2011; Chun et al., 2012; Chernozhukov et al., 2017; Martins-Filho et al., 2018).

How crucial is the fact that risk measures are estimated? Concerns about accuracy of practically used VaR estimates have been raised from the early stages of the literature (Jorion, 1996; Hendricks, 1996; Pritsker, 1997a,b, 2006; Barone-Adesi and Giannopoulos, 2001; Berkowitz and O’Brien, 2002; Krokhmal et al., 2002; Aussenegg and Miazhynskaja, 2006). Chen (2008) notes that the effective sample size for ES at confidence level 1 − \( \alpha \) is the actual sample size times \( \alpha^2 \), stating that the estimator’s high volatility is a common challenge for statistical inference. Caccioli et al. (2017) note that because of the high dimensionality of institutional portfolios and the lack of stationarity in the long run, portfolio optimization is plagued by (relatively) small sample problems. Correspondingly, they derive “lower bounds” to estimation error in the context of ES optimization. Zhu (2008) proposes robust portfolio construction in the context of mean-variance models with estimation risk. Estimation error can thus be a nonnegligible aspect of risk control; we illustrate this point by simulations in Appendix 3.A.

McNeil et al. (2005) also warn the danger of direct interpretation of VaR using estimated quantities, noting that the estimate of the loss distribution is subject to estimation error and model risk as well as market liquidity. In the practice of banking regulations, banks are usually required to hold about 3 times VaR estimated by them in order to cope with such uncertainty (Chen, 2014). This number 3 is called the multiplier, and Stahl (1997) justified the choice of 3 retrospectively by showing that it approximately accounts for model uncertainty for VaR. The new Basel accord also uses ES as the risk measure, and Leippold and Vanini (2002) follow up on this to conclude that the multiplier of 1.5 is appropriate for dealing with model uncertainty with
ES. See also Hendricks and Hirtle (1997), Lopez (1998), and Novak (2010) regarding this matter. However, there has not been discussion on how to integrate potential estimation errors into risk management, beyond quantifying the distributions (standard errors) of the estimates.

The aim of this chapter is to develop a method to reflect the precision assessment on the control of risks that are given by probability bounds. For example, consider VaR of confidence level $1 - \alpha$: $\Pr(X \leq -\text{VaR}_\alpha) \leq \alpha$. Controlling VaR at $1 - \alpha$ thus means bounding the probability of a portfolio loss exceeding the VaR by $\alpha$. True VaR is unobservable, however, so this calls for an estimate, $\hat{\text{VaR}}_\alpha$, to implement the control in practice. The problem with estimates is that we are not ensured to have $\Pr(X \leq -\hat{\text{VaR}}_\alpha) \leq \alpha$; thus, the probability of a large loss is no longer controlled in the original sense. Suppose that a one-sided $(1 - \alpha)$-confidence interval for VaR is available, satisfying $\Pr(\text{VaR}_\alpha \geq \hat{\text{VaR}}_\alpha) \leq \alpha$. Then, by the Bonferroni inequality, we have

$$\Pr(X \leq -\hat{\text{VaR}}_\alpha) \leq \Pr(X \leq -\text{VaR}_\alpha) + \Pr(-\text{VaR}_\alpha \leq -\hat{\text{VaR}}_\alpha) = 2\alpha.$$  

Thus, $\hat{\text{VaR}}_\alpha$ “controls” true VaR with confidence level $1 - 2\alpha$. Retrospectively, if we want to control true VaR at $1 - \alpha$, we may choose $\alpha/2$ for each probability on the right-hand side.

This chapter generalizes this observation and defines a class of risk measures to which the above argument is applicable, which we call tail risk measures. Intuitively, if a risk measure prescribes the probability of some bad events to be lower than some threshold, it is susceptible to the use of the Bonferroni inequality as above. More precisely, we let $\chi(X)$ be the measure of “badness” of a random variable $X$. A tail risk measure associated with $\chi$ will give the number $c_\alpha$ such that the probability of an event in which $\chi(X)$ exceeds $c_\alpha$ is bounded by $\alpha$. Thus, if we have a valid $(1 - \alpha)$ confidence interval for this tail risk measure, we may use the upper bound of the confidence interval to control the tail risk measure with confidence level $(1 - 2\alpha)$.\footnote{Note that there are two distinct concepts referred to as “confidence levels.” One is associated with the concept of the risk measure (as in VaR with confidence level $1 - \alpha$), and the other with the}
We then show that two most popular risk measures, VaR and ES, have interpretations as tail risk measures.

Our method assumes that we have a valid (one-sided) confidence interval for the tail risk measure of interest. Many papers in the literature that provide estimation methods derive inference methods that are valid at least asymptotically; in addition to aforementioned papers such as Chen (2008), Linton and Xiao (2013), and Hill (2015), Belomestny and Krätschmer (2012) derive asymptotic normality for plug-in estimators of law-invariant coherent risk measures; Christoffersen and Gonçalves (2005) provide a resampling method that can correctly estimate the precision of VaR and ES estimates under mind conditions; Gao and Song (2008) establish asymptotic distribution of these estimates.

We demonstrate that our framework can be easily applied to portfolio optimization problems with tail risk measure constraints. For subadditive risk measures such as expected shortfall, one can employ our method to obtain the upper bounds at the individual asset level and approximate the portfolio constraints by a linear combination of the upper bounds.\(^3\) The method does not require any knowledge on the joint distribution and therefore avoids the computational burden related to the estimation of the correlation structure. Using historical data, we show that our method does not suggest extremely conservative portfolio management—the upper bound estimates are only about 10% larger than the point estimates. We deduce that a moderately conservative portfolio choice helps one control the true tail risk.

This chapter is organized as follows. Section 3.2 introduces the notion of a tail risk measure and explains the concept using examples. Section 3.3 shows the Bonferroni argument works for tail risk measures, providing a method to control the true probabilities. Section 3.4 illustrates the use of our method in investment decision making, using the stock return data obtained from Yahoo! Finance. Appendix 3.A describes the problem of estimation error in risk control by simulation studies.

\(^3\)VaR is also known to be subadditive in some special cases (Ibragimov, 2005).
3.2 Tail Risk Measures

This section defines the tail risk measure and introduces its examples. We show that the popular risk measures, VaR and ES, are examples of the tail risk measure.

We first introduce our mathematical setup. Let \((\Omega, \mathcal{F}, P)\) be a probability space that is rich enough to contain all random variables and events introduced below. Let \(X : \Omega \to \mathbb{R}\) be a random variable representing the return of a portfolio. Denote by \(F_X\) and \(Q_X\) the distribution function and quantile function of \(X\). We set \(U\) to be the uniformly distributed random variable on \((0, 1)\) such that \(Q_X(U) = X\) almost surely.

For a random variable \(Y : \Omega \to \mathbb{R}\), denote by \(\sigma(Y) \subset \mathcal{F}\) the \(\sigma\)-algebra generated by \(Y\). For event \(E \in \mathcal{F}\), denote by \(F_{X,E}\) the conditional distribution function of \(X\) conditional on the occurrence of event \(E\).

The tail risk measure is a quantity such that the probability that some characteristic of the portfolio return \(X\) exceeds it is bounded by some number. The characteristic is given by a function \(\chi\) that maps the return distribution \(F_X\) to a real number. The bound on the probability is denoted by \(\alpha\) and we call \(1 - \alpha\) the confidence level.

**Definition.** For \(\alpha \in (0, 1)\) and a map \(F_{X,E} \mapsto \chi(F_{X,E}) \in \mathbb{R}\), the tail risk measure of confidence level \(1 - \alpha\) is the infimum of \(c_{\alpha}\) that satisfies

\[
\sup_{E \in \sigma_0} \{ \Pr(\omega \in E) : \chi(F_{X,E}) \geq c_{\alpha} \} \leq \alpha
\]

for some fixed \(\sigma_0 \subset \sigma(U)\). If the supremum is attained at \(E = \emptyset\), the infimum of \(c_{\alpha}\) is defined to be \(\infty\).

Intuitively, the tail risk measure controls the probability of the occurrence of some bad event. The supremum with respect to \(E\) represents the search for the worst event over a set of events \(\sigma_0\); \(\chi\) represents the measure of badness of the return distribution \(F_{X,E}\) given the event \(E\); \(c_{\alpha}\) gives the threshold above which \(\chi\) will not exceed with probability \(1 - \alpha\).

As a regulator of the banking system, one may want to bound the probability of financial crisis at some level. As a portfolio manager, one may want to minimize the
probability of the portfolio incurring a significant loss. This notion of bounding the probability is the key idea that the tail risk measure captures.

Note that the function \( \chi \) does not depend on \( E \) or \( \alpha \), and the supremum is taken over a subset of the \( \sigma \)-algebra generated by \( U \), not \( \sigma(X) \) nor any other subset of \( \mathcal{F} \). In general, risk measures are defined as functions that map random variables to real numbers; dependence on \( E \) (the \( \sigma \)-algebra) gives too much freedom to qualify as a risk measure. Also, we want \( \alpha \) to represent some probability of a fixed concept, so not allowing \( \chi \) to depend on \( \alpha \) is also natural. The third point will be revisited at the end of this section.

First, we show that risk measures popularly used in practice are tail risk measures. In particular, we explain how \( \text{VaR} \), \( \text{ES} \), and some distortion risk measures can be written as tail risk measures. Then, we motivate the above definition by explaining why variance cannot (and should not) be interpreted as a tail risk measure.

**Example 3.1 (Value-at-Risk).** For \( X : \Omega \to \mathbb{R} \), \( \text{VaR} \) of confidence level \( 1 - \alpha \) is defined as

\[
\text{VaR}_\alpha(X) := \inf\{x \in \mathbb{R} : \Pr(X < -x) \leq \alpha\}.
\]

In words, \( \text{VaR} \) gives the upper bound on the loss of the portfolio that binds with probability at least \( 1 - \alpha \), or equivalently, the worst loss one can expect with probability \( 1 - \alpha \) or higher.

This quantity can be cast as the infimum of \( c_\alpha \) that satisfies

\[
\sup_{E \in \sigma(U)} \left\{ \Pr(\omega \in E) : -\sup \supp(F_{X,E}) \geq c_\alpha \right\} \leq \alpha,
\]

where \( \supp(F_{X,E}) \) denotes the support of distribution \( F_{X,E} \). Therefore, \( \text{VaR} \) at \( 1 - \alpha \) is a tail risk measure of confidence level \( 1 - \alpha \), holding \( \sigma_0 = \sigma(U) \) and \( \chi(F_{X,E}) = -\sup \supp(F_{X,E}) \). \( \square \)

**Example 3.2 (Expected shortfall).** \( \text{VaR} \) has some undesirable features as a risk measure; it ignores the fatness of the tail loss and does not satisfy the so-called coherence. In response to these problems, a new risk measure is proposed and gradually
adapted in practice [Acerbi and Tasche 2002].

For $X : \Omega \to \mathbb{R}$, ES of confidence level $1 - \alpha$ is defined as [Acerbi et al., 2008]

$$\text{ES}_\alpha(X) := \frac{1}{\alpha} \int_0^\alpha \text{VaR}_\gamma(X) d\gamma.$$ 

Note that VaR is (the negative of) the $\alpha$-quantile of $X$. Thus, for continuous returns, ES is equal to $\mathbb{E}[ -X \mid X \leq -\text{VaR}_\alpha ]$, the expected loss in the worst $\alpha$ events. ES is known to hold a nice property, coherence [Acerbi and Tasche 2002].

ES can be cast as the infimum of $c_\alpha$ that satisfies

$$\sup_{E \in \sigma(U)} \left\{ \Pr(\omega \in E) : \mathbb{E}[ -X \mid \omega \in E ] \geq c_\alpha \right\} \leq \alpha.$$ 

Therefore, ES at $1 - \alpha$ is a tail risk measure of confidence level $1 - \alpha$, holding $\sigma_0 = \sigma(U)$ and $\chi(F_{X,E}) = \mathbb{E}[ -X \mid E ] = \int -x dF_{X,E}$. □

Remark. If the distribution of $X$ is discontinuous at $\alpha$, that is, there does not exist $c$ that solves $\Pr(X \leq c) = \alpha$, then the event $U \in (0, \alpha)$ is not in $\sigma(X)$. For this reason, we use $\sigma(U)$ instead of $\sigma(X)$ in the definition.

**Example 3.3** (Tail Value-at-Risk). A quantity that is closely related to expected shortfall is tail Value-at-Risk (TVaR). For $X : \Omega \to \mathbb{R}$, TVaR of confidence level $1 - \alpha$ is given by

$$\text{TVaR}_\alpha(X) := \mathbb{E}[ -X \mid -X \geq \text{VaR}_\alpha(X) ].$$ 

This coincides with expected shortfall if the distribution of $X$ is continuous at $\text{VaR}_\alpha(X)$. TVaR is known not to be coherent [Acerbi 2004].

TVaR can be cast as the infimum of $c_\alpha$ that satisfies

$$\sup_{E \in \sigma(X)} \left\{ \Pr(\omega \in E) : \mathbb{E}[ -X \mid \omega \in E ] \geq c_\alpha \right\} \leq \alpha.$$ 

Note that the only difference from ES is how the supremum is taken; TVaR takes supremum over $\sigma(X)$ while ES over $\sigma(U)$. Thus, TVaR at $1 - \alpha$ is a tail risk measure of confidence level $1 - \alpha$, holding $\sigma_0 = \sigma(X)$ and $\chi(F_{X,E}) = \mathbb{E}[ -X \mid E ]$. □
Example 3.4 (Distortion risk measure). Let $K : [0, 1] \to [0, 1]$ be increasing, $K(0) = 0$, and $K(1) = 1$. The distortion risk measure for the distortion function $K$ is defined as

$$\rho_K(X) := \int_0^1 \text{VaR}_\gamma(X) dK(\gamma) = \int_0^1 -Q_X dK.$$ 

In this chapter we consider distortion functions that satisfy an additional assumption that there exists $\kappa \in (0, 1)$ such that $K(u) < 1$ for $u < \kappa$ and $K(u) = 1$ for $u \geq \kappa$. \text{VaR} is the distortion risk measure with $K(u) = 1\{u \geq \alpha\}$; ES is the distortion risk measure with $K(u) = \max\{u/\alpha, 1\}$. Therefore, both are distortion risk measures that are considered in this chapter. With this assumption, we may write $\rho_K(X)$ as the infimum of $c_\alpha$ such that $\alpha = \kappa$ and

$$\sup_{E \in \sigma(U)} \left\{ \Pr(\omega \in E) : \int_0^1 -Q_{X,E}(u) dK\left(\frac{u}{\kappa}\right) \geq c_\alpha \right\} \leq \alpha,$$

where $Q_{X,E}$ denotes the generalized inverse function of $F_{X,E}$. \hfill \Box

To motivate our definition of the supremum taken only over a subset of the $\sigma$-algebra generated by $U$, it is useful to look at examples of risk measures in the literature that are not tail risk measures.

Example 3.5 (Variance). For $X : \Omega \to \mathbb{R}$ with a second moment, variance is defined by

$$\text{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Variance is not a feature of $X$ conditional on some event, but a feature that captures how spread the entire distribution is; for example, variance penalizes positive and negative returns equally. As variance depends on the entirety of the distribution, it cannot be represented as some conditional quantity of $X$ conditional on an event that restricts a part of the distribution; hence, it is not a tail risk measure.

On the other hand, if we allow the supremum to be taken over a more general sub-$\sigma$-algebra, then one can represent variance in a similar form. Assume that $\mathcal{F}$ is
rich enough so it contains some random variable $V$ that is independent of $U$. Then, variance of $X$ is the infimum of $c_\alpha$ that satisfies
\[
\sup_{E \in \mathcal{F}(V)} \left\{ \Pr(\omega \in E) : \mathbb{E}[(X - \mathbb{E}[X | E])^2 | E] \geq c_\alpha \right\} \leq \alpha
\]
for arbitrary $\alpha \in [0, 1]$, since $\text{Var}(X | V) = \text{Var}(X)$.

However, it is not sensible to regard a decision maker who aims to control variance as one who wants to control the conditional variance on any events at confidence level 0. Put differently, the event that attains the supremum is the one in which $\Pr(V \in V(E)) = \alpha$ and hence the conditional distribution of $X$ is kept equal to its marginal. Therefore, considering such an event as the “worst” case goes astray from what it aims to mean. \qed

**Example 3.6** (Entropic Value-at-Risk). For $X : \Omega \to \mathbb{R}$ that has a finite moment-generating function $M_X(z) := \mathbb{E}[e^{zX}]$ for nonpositive values $z \leq 0$, Ahmadi-Javid (2012) considers entropic Value-at-Risk (EVaR) with confidence level $1 - \alpha$ by
\[
\text{EVaR}_\alpha(X) := \inf_{z < 0} \frac{\log(M_X(z)/\alpha)}{-z}.
\]

Since $M_X$ depends on the entirety of the distribution of $X$, it is not a tail risk measure.

Similarly as variance, however, if we allow a random variable $V$ independent of $U$, EVaR can be written as the infimum of $c_\alpha$ that satisfies
\[
\sup_{E \in \mathcal{F}(V)} \left\{ \Pr(\omega \in E) : \inf_{z < 0} \frac{\log(M_{X_E}(z)/\alpha)}{-z} \geq c_\alpha \right\},
\]
where $X_E$ follows distribution $F_E$. However, again, this does not capture what we want tail risk measures to capture. \quad \square

\footnote{Note that we measure the risk of the negative tail of $X$ while Ahmadi-Javid (2012) measures that of the positive tail, hence the sign differences.}
3.3 Controlling Tail Risk Measures

This section provides a simple rule to control tail risk measures with observable quantities. The problem of traditional approach is that the true risk is unobserved, and upon decision making, the estimated value has been used without taking into account the estimation error. We show that tail risk measures allow a very simple rule based on the Bonferroni inequality to control the true risk.

The intuition is best explained by VaR. VaR of confidence level $1 - \alpha$ satisfies

$$\Pr(X \leq -\text{VaR}_\alpha) \leq \alpha.$$ 

However, if one replaces $\text{VaR}_\alpha$ with an estimated quantity $\hat{\text{VaR}}_\alpha$, the inequality no longer holds. Suppose that one can construct a one-sided $(1 - \alpha)$-confidence interval such that

$$\Pr(\text{VaR} \in (-\infty, \hat{\text{VaR}}_\alpha]) \geq 1 - \alpha.$$ 

Then, by the Bonferroni inequality,

$$\Pr(X \leq -\hat{\text{VaR}}_\alpha) \leq \Pr(X \leq -\text{VaR}) + \Pr(-\text{VaR} \leq -\hat{\text{VaR}}) = 2\alpha.$$ 

Thus, we may control VaR of confidence level $1 - 2\alpha$ with $\hat{\text{VaR}}$. 

The following theorem shows that the same logic holds for tail risk measures defined in the previous section. If one wants to control a tail risk measure of confidence level $1 - \beta$, then one may pick $\alpha > 0$ and $s > 0$ such that $\alpha + s \leq \beta$, and compute the upper bound of a one-sided $(1 - s)$-confidence interval of an estimator for the tail risk measure of confidence level $1 - \alpha$. This allows us to control the true risk with the knowledge of a valid confidence interval, taking into account the estimation error.

**Theorem 3.1.** Let $c_\alpha$ be the true tail risk measure of confidence level $1 - \alpha$, and $\bar{c}_{\alpha,s}$ the $(1 - s)$-confidence bound of $c_\alpha$, that is, $\Pr(\bar{c}_{\alpha,s} \leq c_\alpha) \leq s$. Then for any
\(\alpha, s \in (0, 1)\), we have

\[
\sup_{E \in \sigma_0} \Pr(\omega \in E \land \chi(F_{X,E}) \geq \bar{c}_{\alpha,s}) \leq \alpha + s.
\]

In other words, controlling \(\bar{c}_{\alpha,s}\) controls \(c_{\alpha+s}\) ex ante.

**Proof.** Observe that

\[
\sup_{E \in \sigma_0} \Pr(\omega \in E \land \chi(F_{X,E}) \geq \bar{c}_{\alpha,s}) \leq \sup_{E \in \sigma_0} \left\{ \Pr(\omega \in E) : \chi(F_{X,E}) \geq c_{\alpha} \right\} + \Pr(\bar{c}_{\alpha,s} \leq c_{\alpha})
\]

\[
\leq \alpha + s.
\]

The first inequality follows from the Bonferroni inequality and the second from the properties of \(c_{\alpha}\) and \(\bar{c}_{\alpha,s}\).

**Remark.** Although for any fixed choice of \(\alpha\) and \(s\) such that \(\alpha + s \leq \beta\) the theorem is valid, one cannot “hunt” for the pair \((\alpha, s)\) that gives the lowest upper bound \(\bar{c}_{\alpha,s}\). This will change the distribution of \(\bar{c}_{\alpha,s}\) (due to randomness introduced by the confidence level hunting), and the probability bound may no longer be valid.

One key assumption of our method is the availability of a valid (one-sided) confidence interval for the tail risk measure we aim to control. There is a large body of the literature that concern estimation of risk measures, and many such papers also discuss distributions of risk measure estimators \cite{Christoffersen2005, Gao2008, Chen2008, Linton2013, Hill2015, Belomestny2012}.

### 3.4 Empirical Applications to Expected Shortfall

An investor is considering to invest in three securities, namely, the stock of the Bank of America Corp., the stock of Morgan Stanley, and the index fund of the Dow Jones Industrial Averages. She wants to control the expected shortfall of level 90\% of her portfolio below some threshold \(C\). Since ES is a coherent risk measure, it is
subadditive, meaning that the ES of a diversified portfolio does not exceed the sum of ES of each asset. Therefore, upon optimizing her portfolio, she does not need to calculate ES of every candidate portfolio, but only needs to calculate ES for each asset and combine them in her optimization.

The investor is to control the “true” ES of her portfolio by the combination of the estimated ES and their confidence intervals. Our statistical objective is thus to estimate each ES of level $1 - \tau_1$ of the three securities and their joint confidence intervals of level $1 - \tau_2$ so that $\tau_1 + \tau_2 \leq 0.1$. We choose $\tau_1 = \tau_2 = 0.05$. Denoting by $Y_t$ the daily price, the daily return is defined as $r_t := (Y_t - Y_{t-1})/Y_{t-1}$, so the target ES is given by

$$\text{ES}_\tau = -\int_0^\tau Q_{r_{T+1}|Y_T,Y_{T-1},...}(u)du,$$

where $Q_{r_{T+1}|Y_T,Y_{T-1},...}$ is the quantile function of $r_{T+1}$ given past data $Y_T,Y_{T-1},...$. For continuous distributions, this is equal to $\mathbb{E}[\{-r_{T+1} \mid r_{T+1} < -\text{VaR}_\tau, Y_T,Y_{T-1},...\}]$.

For stock returns, it has been widely documented that volatility clusters. Markets tend to experience volatile moves following immediate volatile moves, and calm moves after calm moves. Following Hansen and Lunde (2005), Brownlees et al. (2012), and Barone-Adesi et al. (1999), we model daily stock returns as generalized autoregressive conditional heteroskedasticity (GARCH) processes. Formally, daily returns $\{r_t\}$ are modeled according to GARCH(1,1):

$$r_t = \mu + \varepsilon_t = \mu + \sigma_t z_t,$$
$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2,$$

where $\{z_t\}$ are i.i.d. random variables. To ensure nonnegativity of the variance process, parameters are restricted to $\omega, \alpha, \beta \geq 0$. For stationarity, we also assume that $\alpha + \beta < 1$.

To illustrate the utility of our method, we compare ES estimates at the 90% level and the upper bound of a 95% confidence interval around ES at 95%. Since there is no full agreement on which estimation and inference methods do well on financial returns, we employ two methods to illustrate our results. First, we use historical
Figure 3-1: The daily returns of the stock of Bank of America Corp., the stock of Morgan Stanley, and the index fund of the Dow Jones Industrial Averages from January 4, 2016 to December 31, 2017.

*Historical Simulation (HS)* to estimate ES ([Barone-Adesi and Giannopoulos, 2001]) and nonparametric bootstrap to construct confidence intervals (Chapter 1). This is a reasonable method when returns are i.i.d., i.e., $\alpha = \beta = 0$, with very little distributional restrictions on $z_t$. Second, we employ *filtered historical simulation (FHS)* to estimate ES ([Barone-Adesi et al., 1999]) and the resampling technique proposed by [Christoffersen and Gonçalves, 2005] to construct confidence intervals. They allow the full GARCH structure on returns but assumes normality of $z_t$ in the maximum likelihood estimation of GARCH parameters.

We use the daily returns from January 4, 2016 to December 31, 2017, consisting of 502 business days in total. The price data are retrieved from Yahoo! Finance. Figure 3-1 shows the chart of daily returns of the three assets.

### 3.4.1 Historical Simulation

This section assumes that the returns are i.i.d., that is, $\alpha = \beta = 0$ for all three assets. As our interests are in the immediate prediction of ES at $T + 1$, the HS estimator of ES of level $\tau$ is simply given by the tail sample average:

$$\hat{ES}_{\tau,1} = -\frac{1}{[\tau T]} \sum_{t=1}^{[\tau T]} r(t),$$
Table 3.1: To control ES at 90%, we estimate ES of the daily stock returns at 95% from Feb 24, 1993 to Dec 31, 2017, and the upper bounds of the one-sided CB of level 95% by HS.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Marginal CB</th>
<th>Joint CB</th>
<th>Conservative CB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\text{ES}}_{10%,1}$</td>
<td>$\hat{\text{ES}}_{5%,1}$</td>
<td>$\hat{\text{ES}}_{5%,1}$</td>
<td>$\hat{\text{ES}}_{5%,1}$</td>
</tr>
<tr>
<td>Bank of America Corp.</td>
<td>3.16%</td>
<td>4.01%</td>
<td>4.04%</td>
<td>4.05%</td>
</tr>
<tr>
<td>Morgan Stanley</td>
<td>3.06%</td>
<td>3.95%</td>
<td>3.98%</td>
<td>3.98%</td>
</tr>
<tr>
<td>Dow Jones Industrial Average</td>
<td>1.21%</td>
<td>1.53%</td>
<td>1.54%</td>
<td>1.54%</td>
</tr>
</tbody>
</table>

*The CBs are based on bootstrap with 1,000 simulations.

where $r(t)$ is the $t$th order statistic of $\{r_t\}$. This is a reasonable nonparametric estimator given that $r_t$ are i.i.d.$^5$

For confidence intervals, we use nonparametric bootstrap to approximate the asymptotic distribution of $\hat{\text{ES}}_{\tau,1}$ (Chapter [1]). In particular, we are interested in the confidence bound (CB), the upper bound of a one-sided confidence interval. The procedure to obtain a CB is as follows:

Step 1. Draw a subsample of size $T$ from the historical returns $\{r_t\}_{t=1}^{T}$ with replacement, letting $\{r_t^*\}_{t=1}^{T}$.

Step 2. Compute $\hat{\text{ES}}_{\tau,1}^*$ using $\{r_t^*\}_{t=1}^{T}$.

Step 3. Repeat Steps 1–2 for $S$ times, where $S$ is a large number.

Step 4. Approximate the distribution of $\hat{\text{ES}}_{\tau,1}$ by that of $\{\hat{\text{ES}}_{\tau,1,s}\}_{s=1}^{S}$. In particular, construct the CB for ES$_\tau$ with confidence level $1-\tau'$ by the $(1-\tau')$th quantile of $\{\hat{\text{ES}}_{\tau,1,s}\}_{s=1}^{S}$. Denote it by $\hat{\text{ES}}_{\tau,1}^{\tau'}$.

The joint dependence among ES estimators of different assets is taken care of by subsampling from the same time indices. That is, when we construct a subsample, we construct the subsample of the entire vector of three returns, so that the joint distribution of three $\{\hat{\text{ES}}_{\tau,1,s}\}_{s=1}^{S}$ approximates the joint distribution of $\hat{\text{ES}}_{\tau,1}$.

$^5$Under dependence assumptions, [Chen (2008)] studies this as an estimator for stationary ES.
Table 3.1 gives the point estimates and CBs of ES. The values are in the units of negative daily returns in the worst 5% cases in percentage points. Column (1) gives the point estimates of ES of level 90% for each asset; for instance, the expected return of the Bank of America Corp. stock in a day in the worst 5% events is estimated to be $-3.16\%$. These numbers are subject to estimation error, and Columns (2–4) give numbers that take this into account. Column (2) shows the CB of level 95% for ES of level 95% constructed with marginal distributions; Columns (3–4) give the same CB but constructed in ways that ensure joint coverage. In particular, from the joint bootstrapped distribution $\{\hat{ES}_{5\%,1,s,BAC}, \hat{ES}_{5\%,1,s,MS}, \hat{ES}_{5\%,1,s,DJI}\}_{s=1}^S$, we pick $c$ so that

$$\begin{pmatrix} \hat{ES}_{5\%,1,s,BAC} \\ \hat{ES}_{5\%,1,s,MS} \\ \hat{ES}_{5\%,1,s,DJI} \end{pmatrix} \leq c \cdot \text{SE} \begin{pmatrix} \hat{ES}_{5\%,1,s,BAC} \\ \hat{ES}_{5\%,1,s,MS} \\ \hat{ES}_{5\%,1,s,DJI} \end{pmatrix}$$

is ensured for 95% $S$ times, where SE(·) denotes the vector of marginal standard errors; the right-hand side of the display is Column (3), which we call the joint CB. For Column (4), we take the marginal CB of level $100\% - 5\%/3 \approx 98.3\%$; then, by
the Bonferroni inequality, joint coverage is guaranteed. To avoid confusion with the Bonferroni inequality applied in Section 3.3, we call this the conservative CB, instead of more descriptive “Bonferroni CB.” It can be seen that the inflation of the bounds required for the correct risk coverage is not very big. Even the conservative CBs are only about 27–30% larger than the original point estimates. Figure 3-2 visually shows how large the CBs are compared to the point estimates.

With the results from Section 3.3, we know that the CB can control the true risk probability at 10% given that the CB is valid. Meanwhile, in reference to results in Appendix 3.A, the risk probability with point estimates of ES of level 90% may not be controlled at the intended level. Thus, the direct estimates of 90% ES in Column (1) is not appropriate to control the risk for the investor.

Going back to the investor’s problem, she can control her expected shortfall, for its subadditivity, by meeting the following criterion (McNeil et al., 2005, p. 240):

\[ 4.01\% w_{\text{BAC}} + 3.95\% w_{\text{MS}} + 1.53\% w_{\text{DJI}} \leq C \]

for \( w_{\text{BAC}}, w_{\text{MS}}, w_{\text{DJI}} \geq 0 \) and \( w_{\text{BAC}} + w_{\text{MS}} + w_{\text{DJI}} = 1 \), where \( w_i \) denotes the share of security \( i \) in her portfolio\(^6\). For example, she can maximize the expected return of her portfolio subject to this constraint (Krokhmal et al., 2002; Marianne and Sannes, 2016). To extend the setup to allow for short positions, observe that the expected shortfall of a short position is the highest 100\( \tau \)% expected return of each security. Thus, simply adding the new security \(-Y_t\) will do the job.

### 3.4.2 Filtered Historical Simulation

Figure 3-3 shows the fitted volatility processes \( \hat{\sigma}_t \) of the three assets computed by FHS, which can be interpreted as an indication of volatility clustering. Thus, to allow for the full GARCH structure (\( \alpha > 0 \) and \( \beta > 0 \)), this section applies FHS to estimate ES (Barone-Adesi et al., 1999). This assumes that \( z_t \) follows independent standard

\[^6\text{Alternatively, we may use the conservative CB, especially when it is infeasible to compute the joint CB.}\]
normal distribution and estimates the GARCH parameters by maximum likelihood. With estimated parameters, we predict $\sigma_{T+1}$ and approximate the distribution of $r_{T+1}$ by $\{\hat{\mu} + \hat{\sigma}_{T+1} \hat{z}_t\}_{t=1}^T$, where $\{\hat{z}_t\}$ are the fitted residuals. Then, compute ES by

$$\hat{\text{ES}}_{r,2} = -\frac{1}{|\tau T|} \sum_{t=1}^{[\tau T]} (\hat{\mu} + \hat{\sigma}_{T+1} \hat{z}_{(t)}),$$

where $\hat{z}_{(t)}$ is the $t$th order statistic of $\{\hat{z}_t\}$.

In order to construct confidence intervals, we follow the subsampling method proposed by Christoffersen and Gonçalves (2005). The procedure is as follows:

**Step 1.** With the historical returns $\{r_t\}_{t=1}^T$, estimate the GARCH parameters in (3.1–3.2) by maximum likelihood, yielding $(\hat{\mu}, \hat{\omega}, \hat{\alpha}, \hat{\beta})$. Fit the standardized residuals $\hat{z}_t = (r_t - \hat{\mu})/\hat{\sigma}_t$.

**Step 2.** Draw a subsample of size $T$ from $\{\hat{z}_t\}_{t=1}^T$ with replacement, letting $\{\hat{z}^*_t\}_{t=1}^T$. Use it to construct bootstrap returns $\{r^*_t\}_{t=1}^T$.

**Step 3.** Regarding the bootstrap returns $\{r^*_t\}_{t=1}^T$ as the true returns, estimate the GARCH parameters by maximum likelihood, yielding $(\hat{\mu}^*, \hat{\omega}^*, \hat{\alpha}^*, \hat{\beta}^*)$.

**Step 4.** With $(\hat{\mu}^*, \hat{\omega}^*, \hat{\alpha}^*, \hat{\beta}^*)$ and the *historical* returns $\{r_t\}_{t=1}^T$, predict volatility in the next period $\hat{\sigma}^*_{T+1}$ by applying (3.1–3.2) iteratively.
Step 5. Compute $\widehat{ES}_{\tau,2}^*$ using the distribution of $\{\hat{\mu}^* + \hat{\sigma}_{T+1}^* \hat{z}_t^*\}$ constructed with the bootstrap standardized residuals $\hat{z}_t^* = (r_t^* - \hat{\mu}^*)/\hat{\sigma}_t^*$. 

Step 6. Repeat Steps 2–5 for $S$ times, where $S$ is a large number.

Step 7. Approximate the distribution of $\widehat{ES}_{\tau,2}^*$ by that of $\{\widehat{ES}_{\tau,2,s}^*\}_{s=1}^S$ to construct confidence intervals.

Table 3.2 gives the point estimates and CB of ES. Again, the values are in terms of negative daily returns in the worst 5% cases in percentage points. Column (1) enumerates the point estimates of ES of level 90% of the three assets. The expected return of the Bank of America Corp. stock in a day in the worst 5% events is estimated to be $-2.26\%$, which is smaller in magnitude compared to Table 3.1. This is considered to be because of the volatility movement. In Figure 3-3 we see that the overall volatility in 2017 is lower than in 2016. The GARCH structure takes into account that the current volatility is on the low end of the historical data, resulting in small predicted ES as opposed to HS that assumes i.i.d. returns. Column (2) shows the 95% CB for 95% ES constructed with marginal distributions; Columns (3–4) give the same CB but ensure joint coverage as before. With FHS, the numbers that ensure the correct risk coverage is about 60–90% larger than the original point estimates. This is larger than the previous section, but still comparable to the inflation rate required for model risk (Leippold and Vanini, 2002). Figure 3-4 visualizes the magnitudes of CBs against point estimates.

These estimates allow the investor to control ES of her portfolio under the GARCH assumptions. Similarly as Section 3.4.1, her ES constraint is given by $3.75\% w_{\text{BAC}} + 3.19\% w_{\text{MS}} + 1.32\% w_{\text{DJI}} \leq C$, where $w_i$ denotes the share of security $i$ in her portfolio.

3.5 Conclusion

In this chapter, we addressed the issue of estimation errors in financial risk control. When we use estimated risk measures in the practice of risk management—be it portfolio optimization or banking regulation—we incur additional uncertainty due to
Table 3.2: To control ES at 90%, we estimate ES of the daily stock returns at 95% from Feb 24, 1993 to Dec 31, 2017, and the upper bounds of the one-sided confidence bounds (CB) of level 95% by FHS.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Marginal CB</th>
<th>Joint CB</th>
<th>Conservative CB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\text{ES}}_{10%,2}$</td>
<td>$\hat{\text{ES}}_{5%,2}$</td>
<td>$\tilde{\text{ES}}_{5%,2}$</td>
<td>$\tilde{\text{ES}}_{5%/3}$</td>
</tr>
<tr>
<td>Bank of America Corp.</td>
<td>2.26%</td>
<td>3.60%</td>
<td>3.75%</td>
<td>3.75%</td>
</tr>
<tr>
<td>Morgan Stanley</td>
<td>1.94%</td>
<td>3.04%</td>
<td>3.19%</td>
<td>3.19%</td>
</tr>
<tr>
<td>Dow Jones Industrial Average</td>
<td>0.71%</td>
<td>1.27%</td>
<td>1.32%</td>
<td>1.32%</td>
</tr>
</tbody>
</table>

* The CBs are based on bootstrap with 1,000 simulations.

Figure 3-4: ES of level 95% of the daily returns of the stock of Bank of America Corp., the stock of Morgan Stanley, and the index fund of the Dow Jones Industrial Averages from Jan 4, 2016 to Dec 31, 2017. The marginal, joint, and conservative CB of level 95% are also indicated. FHS is employed for estimation.

estimation errors. Observing that many risk measures are designed to control the probabilities of bad events, we proposed a method to control the “true but unobservable” risk probability under the assumption that a valid confidence interval is available.

Let $\chi(F)$ be a quantity that measures “badness” associated with the return distribution $F$. We defined a class of risk measures, called tail risk measures, which bound the maximum probabilities of events that entail large values of $\chi$. For example, if
\( \chi(F) \) is the negative of the expectation of \( F \), ES can be given as a quantity that bounds the probability with which the negative expectation of the return exceeds ES. We showed, most notably, that VaR and ES, arguably the two most popular risk measures in the present, are given as tail risk measures. We also gave examples of risk measures, variance and EVaR, that are not considered as tail risk measures.

Next, we established a method to control the true risk probability using the Bonferroni inequality. The idea is to construct a valid (one-sided) confidence interval of the risk measure, and use the upper bound of the confidence interval as the risk estimate. Because of the properties of the risk measure and the confidence interval, we showed that this upper bound manages to control the true probability that the risk measure was designed to control. The key to the proof was the Bonferroni inequality.

Our empirical application showed how our method can be used to control ES of a portfolio that consists of three assets. We used both HS and FHS to compute ES of daily returns of the Bank of America Corp. stock, the Morgan Stanley stock, and the index fund of the Dow Jones Industrial Average. For confidence intervals, we applied nonparametric bootstrap and the bootstrap method of Christoffersen and Gonçalves (2005), respectively. Our results showed that our technique proposed for tail risk measures require controlling ES that are about 30–90% larger than the point estimates, comparable to the multiplier needed to control model risk (Leippold and Vanini, 2002). Also, combining with the subadditivity of ES, we showed an effective way to control the true ES of a portfolio that only requires knowledge of ES of individual assets.

In addition to the risk control literature, this chapter contributes to the work on microfounding the multiplier that is used in modern banking regulations (Stahl, 1997; Leippold and Vanini, 2002). Practical risk control suffers from estimation error and model risk. The previous literature gave justification on the use of multiplier 3 on VaR control from the perspective of accounting for model risk. This chapter provides new insights on how much conservativeness is required in order to guard against possible estimation error.
Appendix

3.A Distortion of Risk Probability with Estimated VaR

The defining feature of VaR is that the probability of the return variable $X$ falling below VaR, $\Pr(X < -\text{VaR}_\alpha)$, is bounded by, often equal to, $\alpha$. We call this probability the risk probability. Once we substitute VaR with its estimator $\hat{\text{VaR}}_\alpha$, however, we do not know whether $\Pr(X < -\hat{\text{VaR}}_\alpha)$ is bounded by $\alpha$.

This section carries out simulation studies to illustrate how the substituted risk probability compares to $\alpha$. Our simulation consists of two iterations. In the first iteration, we draw $n$ return observations $X_1, \ldots, X_n$ and estimate $\text{VaR}_\alpha$. For this estimate, there is a corresponding probability by which the new draw falls below it, i.e., $\Pr(X_{n+1} < -\hat{\text{VaR}}_\alpha \mid \hat{\text{VaR}}_\alpha)$. Since $\hat{\text{VaR}}_\alpha$ may realize at any value, this probability can itself be larger or smaller than $\alpha$. In the next iteration, we repeat this exercise for $S$ times and compute the unconditional probability by which the new draw falls below the estimator, i.e., $\Pr(X_{n+1} < -\hat{\text{VaR}}_\alpha)$. For simplicity, we hereafter drop the subscript $n + 1$. Note that by the law of iterated expectations,

$$\Pr(X < -\hat{\text{VaR}}_\alpha) = \mathbb{E}[\mathbb{E}1\{X < -\hat{\text{VaR}}_\alpha \mid \hat{\text{VaR}}_\alpha]\mid \hat{\text{VaR}}_\alpha] = \mathbb{E}[\Pr(X < -\hat{\text{VaR}}_\alpha \mid \hat{\text{VaR}}_\alpha)].$$

This probability naturally depends on the estimation method we employ for VaR. Thus, we take three methods and compute the unconditional probability for each. We will see, somewhat surprisingly, that even when an estimation method does a great
job in estimating VaR, it does not necessarily make the unconditional probability close to \( \alpha \).

To make clear that the issue is not a mere byproduct of the intrinsically complicated financial data structure, we keep the simulation setup as simple as possible. First, we assume that the return variables \( X_1, \ldots, X_n \) follow independent standard normal distribution, where \( X_{(1)} \leq \cdots \leq X_{(n)} \) denote their order statistics. To reflect the small sample nature of financial data (Caccioli et al., 2017), we let \( n = 100 \) and \( \alpha = 0.01 \). To accurately assess unconditional probabilities, we carry out \( S = 10,000 \) iterations of VaR estimation. Denoting by \( \Phi \) the cdf of the standard normal distribution, the true VaR is \( -\Phi^{-1}(0.01) \approx 2.33 \).

The following three methods are employed to estimate VaR: maximum likelihood estimation (MLE), a method of Weissman (1978), and historical simulation. Roughly speaking, these three methods correspond to parametric, semiparametric, and non-parametric methods, respectively. All of them are “correctly specified;” in MLE, we maximize the independent normal likelihood; in Weissman, we assume that the return distribution is in the domain of attraction of Gumbel distributions, which indeed is so for normal returns; in historical simulation, we take the empirical quantile, which is a valid estimator for i.i.d. random variables.

MLE maximizes the normal likelihood of the data. In particular, we first estimate the two parameters of the normal distribution by \( \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \) and \( \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2 \), and then estimate VaR by

\[
\hat{\text{VaR}}_{\alpha, 1} := -\hat{\mu} - \hat{\sigma} \Phi^{-1}(0.01).
\]

Figure 3.A.1a shows the histogram of VaR estimated by MLE around the true VaR represented by the red line. Being a correctly specified parametric model, MLE estimates VaR more accurately than other methods, compared to histograms in Figures 3.A.1c and 3.A.1e. Mean squared error (MSE) of MLE estimator is 0.0378, the smallest of the three. However, in Figure 3.A.1b we see the sign of risk probability distortion. It shows the histogram of conditional risk probabilities \( \Pr(X < -\hat{\text{VaR}}_{\alpha, 1} \mid \)
\( \widehat{\text{VaR}}_{\alpha,1} \), together with the unconditional risk probability depicted by the orange line. The unconditional probability 1.21\% is much larger than \( \alpha = 1\% \). This means that, by repeatedly using MLE estimates, the chances that the loss gets worse than the estimate are larger than intended, that is, we underestimate the risk.

Weissman’s estimator uses the smallest \( k \) observations to estimate a tail quantile, relying on extreme value theory. We set \( k = n/10 = 10 \) and estimate VaR by

\[
\widehat{\text{VaR}}_{\alpha,2} := -\left( \frac{1}{k} \sum_{i=1}^{k} X(i) - X(k) \right) \log\left( \frac{k}{\alpha n} \right) - X(k).
\]

Figure 3.A.1c gives the histogram of Weissman’s VaR estimates. The performance of this method (as an estimator of VaR) is in the middle of the three methods, giving MSE of 0.0943. However, we see in Figure 3.A.1d that there is a large upward distortion of the risk probability entailed by this method; \( \Pr(X < -\widehat{\text{VaR}}_{\alpha,2}) = 0.0143 > \Pr(X < -\text{VaR}_\alpha) = 0.01 \). Again, the expected frequency that the loss goes above Weissman’s VaR is larger than the targeted probability, resulting in underestimation of risk.

Historical simulation in this context is quite simply implemented as the empirical \( \alpha \)-quantile of historical data. Thus, the estimate is the maximum loss of the historical return, i.e.,

\[
\widehat{\text{VaR}}_{\alpha,3} := -X(1).
\]

In Figure 3.A.1e we see that the histogram of historical estimates is more dispersed than the previous estimates, with MSE equal to 0.2197. Figure 3.A.1f is the histogram of conditional risk probabilities for historical simulation. Despite the VaR estimates being “worse” than the first two methods, the unconditional risk probability 0.0099 is very close to the intended probability 0.01. However, we cannot take this result at face value since the simulation assumes a very restrictive case of i.i.d. normal returns and it is known in the literature that historical simulation does not usually work well with actual financial data (Pritsker 2006).

\footnote{Weissman (1978) proposes two estimators: one based on MLE and the other on minimum variance unbiased estimation (MVUE). We use the first one.}
Finally, the source of distortion may be either the bias or variance of the estimator, or both of them. De Haan et al. (2016) discuss bias that comes from applying extreme value theory. It is noteworthy that our method works whenever we have a valid confidence interval, regardless of the original estimator suffering from large bias or large variance. Moreover, our method does not even require that we have an estimator; if we can obtain the (one-sided) confidence interval of the true VaR, we can control the true risk probability applying the results of this chapter.
(a) Histogram of VaR estimated by MLE. MSE = 0.0378.

(b) Histogram of risk probability associated with MLE.

(c) Histogram of VaR estimated by Weissman’s estimator. MSE = 0.0943.

(d) Histogram of risk probability associated with Weissman’s estimator.

(e) Histogram of VaR estimated by historical simulation. MSE = 0.2197.

(f) Histogram of risk probability associated with historical simulation.

Figure 3.A.1: Simulations of $n = 100$ normal returns with $S = 10,000$ iterations illustrate distortion of risk probability, $\Pr(X < -\hat{\text{VaR}}_\alpha) \neq \alpha$. Draws outside the range are shown as clusters at the boundaries.
Bibliography


