INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6” x 9” black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600
RICE UNIVERSITY

Practical and Effective Methods of Simulation Based Parameter Estimation for Multidimensional Data

by

Otto W. Schwalb III

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

James R. Thompson, Professor
Department of Statistics
Thesis Director

Marek Kimmel, Professor
Department of Statistics

Peter Olofsson, Lecturer
Department of Statistics

Richard Tapia, Noah Harding Professor
Computational and Applied Mathematics

Houston, Texas
April, 1999
Abstract

Practical and Effective Methods of Simulation Based Parameter Estimation for Multidimensional Data

by

Otto W. Schwalb III

In 1983, Atkinson, Bartoszynski, Brown, and Thompson proposed a method of parameter estimation referred to as “simulation based estimation”, or SIMEST. SIMEST is closely related to maximum likelihood, in that both methods deal with parameter estimation in the context of a fully parametric model. With SIMEST, however, the arduous step of obtaining the density function from a set of model axioms is avoided via simulation.

In this dissertation, we extend the ideas of SIMEST to the case of multidimensional data. A nearest neighbor based binning scheme is proposed where the observations are divided into bins determined by the “rings” of concentric ellipsoids, the “rings” being chosen to roughly approximate regions of equal probability. The ellipsoids are each allowed to have different axes, the axes for each ellipsoid being determined by the data. Some theoretical justification is developed which establishes strong consistency
for the parameter estimates obtained by this method. The theory also suggests a promising variation on the idea using many overlapping bins. Another theoretical topic related to the problems associated with global optimization in SIMEST is also treated.

We explore the usefulness of these techniques in modeling the secondary tumor generation mechanisms of cancer. In one model, it is assumed that 3-dimensional data is available: (a) the time from the detection and removal of the primary to the discovery of the first secondary tumor, (b) the volume of the primary tumor, and (c) the volume of the first secondary tumor. In a second model, it is assumed that two additional dimensions of information are available (i.e. 5-dimensional data): (d) the time from the detection of the first secondary tumor to the detection of the second secondary tumor and (e) the volume of the second secondary tumor.
Acknowledgments

I would like to acknowledge the U.S. Army Research Office (Durham) for financial support throughout my research for this dissertation, specifically U.S. Army Research Office (Durham) DAA1-104-95-1-0665 and DAAD19-99-1-0150. I would like to thank the members of my committee, each of whom I hold in the highest esteem, for their thoughtful consideration of my work and their excellent questions and comments. I would like to thank my advisor Jim Thompson, for suggesting such interesting research topics, for guidance over nearly three years at Rice, for staying in Houston during his sabbatical over the past year partly for my benefit, and for his encouragement particularly during some of the darker hours. I would also like to thank Dr. Thompson for giving me so much independence, allowing me to head in whatever theoretical directions seemed most promising to me, and never expressing disappointment when weeks sometimes turned into months and the efforts of several months sometimes resulted in nothing which we could use.

For me, this dissertation represents only the tip of an iceberg, only the most successful work making its way into the main body. The main theoretical results; however, are very mathematically satisfying to me, rigorously answering some of the questions which were originally posed to me in the generality in which they were posed.
I would like to thank my lovely wife, Melissa, for her emotional support during my six years of graduate school. Certainly there could only be happier years ahead, and hopefully the knowledge and experience I have gained during this time will pay off for both of us.

I would like to dedicate this dissertation to my father, Otto William Schwalb, and to the memory of my mother, Mary Burnett Schwalb.
Contents

Abstract ii
Acknowledgments iv

1 Introduction 1

1.1 Introduction and Notation 1
1.2 Lumpy Estimation Surfaces 4
1.3 Multivariate SIMEST 7
1.4 Main Ideas of this Dissertation 8
1.5 Overview and Discussion of Literature 9

2 Lumpy Likelihoods and Banach-valued Random Variables 17

2.1 Example of Cauchy Location Likelihood 18
2.2 Extending the Perlman [42, 1983] Formulation 23
2.3 A Simpler Formulation 29
2.4 Examples 32
2.5 Discussion and Literature 41

3 A Strong Consistency Theorem for SIMEST 43

3.1 Multivariate Binning Strategies 44
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>Concentric Ellipsoidal Binning for Multivariate SIMEST</td>
<td>47</td>
</tr>
<tr>
<td>3.3</td>
<td>Theoretical Issues Involved in Our Methods</td>
<td>48</td>
</tr>
<tr>
<td>3.4</td>
<td>Consistency Results</td>
<td>52</td>
</tr>
<tr>
<td>3.5</td>
<td>The Use of Data-Based Bins</td>
<td>72</td>
</tr>
<tr>
<td>4</td>
<td>Strong Consistency for the Cressie-Read Class</td>
<td>77</td>
</tr>
<tr>
<td>4.1</td>
<td>A Strong Consistency Theorem for the CR Class</td>
<td>77</td>
</tr>
<tr>
<td>4.2</td>
<td>Proof of Theorem 4.1</td>
<td>80</td>
</tr>
<tr>
<td>4.3</td>
<td>Simulation Based Theorem for the Cressie-Read Class</td>
<td>89</td>
</tr>
<tr>
<td>5</td>
<td>Some Illustrative Examples</td>
<td>93</td>
</tr>
<tr>
<td>5.1</td>
<td>Carcinogen Example</td>
<td>94</td>
</tr>
<tr>
<td>5.2</td>
<td>Accommodating Rejection Algorithms in Simulation</td>
<td>105</td>
</tr>
<tr>
<td>5.3</td>
<td>Cancer Model of Atkinson, Bartoszyński, Brown and Thompson</td>
<td>109</td>
</tr>
<tr>
<td>6</td>
<td>A Data-based Selection Criterion for SIMEST for Improved Parameter Estimation with Multidimensional Data</td>
<td>116</td>
</tr>
<tr>
<td>6.1</td>
<td>Cancer Model of Atkinson, Bartoszyński, Brown and Thompson</td>
<td>118</td>
</tr>
<tr>
<td>6.2</td>
<td>Estimation for 3D Data</td>
<td>121</td>
</tr>
<tr>
<td>6.3</td>
<td>Estimation for 5D Data</td>
<td>141</td>
</tr>
<tr>
<td>6.4</td>
<td>Computational Section</td>
<td>164</td>
</tr>
<tr>
<td>6.5</td>
<td>A Future Direction of Research</td>
<td>184</td>
</tr>
</tbody>
</table>
Bibliography

A Background on Banach-valued Strong Laws

A.1 Measurability and Banach-valued Random Variables ................. 195
A.2 Continuity of Mappings and Measurability ............................... 196
A.3 Strong Measurability When $\Xi$ is Separable ......................... 201
A.4 The Bochner Integral and Expected Value .............................. 204
A.5 The Strong Law for $\Xi$-valued Random Variables .................... 211

B Background on Generalized Glivenko-Cantelli Results 214

B.1 Generalized Glivenko-Cantelli Results ................................. 214
B.2 Permissible Classes and Measurability .................................. 219
B.3 Easy Ways to Establish Permissibility ................................... 221

C Literature search for Perlman papers ................................. 224

D Consistency Summary for Maximum Likelihood 227

E Goodness-of-Fit Literature .................................................. 231

E.1 Obvious Approaches to Minimum Chi-square ......................... 231
E.2 Parameter Estimation in the Goodness-of-Fit Literature ............ 232
F An Asymptotic Result for the Case Where the Number
of Bins Tends to Infinity

F.1 Technical Assumptions ................................................. 240
F.2 Outline ................................................................. 245
F.3 Weak consistency of $\hat{\theta}_n$ ..................................... 247
F.4 Asymptotic normality of $n^{-1/2}g'(\theta_0)$ .......................... 250
F.5 Proof of $\sum_{i=1}^{k} h_i p_i - \frac{1}{n} g''(\theta_0) = o_p(1)$ .......... 254
F.6 Proof of $-\frac{1}{2n} (\hat{\theta}_n - \theta_0) g''(\theta_0^*) = o_p(1)$ .......... 257
F.7 Proof of the identity $\sum_{i=1}^{k} g_i^2 p_i = - \sum_{i=1}^{k} h_i p_i$ ............... 266
F.8 $\sum_{i=1}^{k} g_i^2 p_i$ is bounded below .................................. 267
F.9 Main Results (Lemmas) ............................................... 269
F.10 A Simplification (Our Main Theorem) ............................ 281
Chapter 1

Introduction

1.1 Introduction and Notation

Basic Notation

In this dissertation, attention will be generally focused on the following situation.

Let $x_1, x_2, \ldots, x_n$ be a set of observations from a continuous density function $f(x, \theta_0)$ where the $x_j$'s are $d$-dimensional vectors. As an example, suppose that for $n$ cancer patients one has three pieces of information about each individual:

- $t =$ time from the discovery of the primary tumor until the discovery of the first secondary tumor,
- $y_0 =$ volume of the primary tumor at discovery, and
- $y_1 =$ volume of the first secondary tumor at discovery.

Here both the time and the two volumes are continuous dimensions. The dimension of such data would be three.

It is assumed that $\theta_0 \in \Theta$ where $\Theta$ is referred to as the parameter space. It is also assumed that $\Theta$ is some subset of $\mathbb{R}^s$, making $\theta$ a vector in $\mathbb{R}^s$. The main interest throughout will be in binning the continuous (multivariate) problem. The bins will
be denoted by $I_1, I_2, \ldots, I_k$, where there are $k$ bins and $n$ observations. The index $i$ will be reserved for the bins (as in $I_i$), and the index $j$ will generally be reserved for the observations (as in $x_j$).

**What is SIMEST?**

The reader who may not be familiar with SIMEST will probably appreciate a quick explanation of what is meant by SIMEST. SIMEST is an acronym for simulation based estimation, and was proposed originally by Atkinson, Bartoszyński, Brown and Thompson [1, 1983]. The basic motivation for SIMEST is depicted in Figure 1.1. One starts with a set of model axioms or a fully specified model. Even with a simple set of model axioms, it can be quite difficult to get from these axioms to $f(x, \theta)$, the density function. Of course, one must possess the density function in order to do maximum likelihood. On the other hand, from a set of axioms it is almost always an easy exercise to simulate realizations from the model for any given choice of the model parameters. Is it possible to exploit this situation somehow by finding a method of parameter estimation which requires only the simulator and does not require $f(x, \theta)$?

The answer of course is yes.

Recall minimum chi-square (MCS) estimation, where the parameter estimate $\hat{\theta}_n$ is chosen so that

$$
\hat{\theta}_n = \arg \min_{\theta \in \Theta} X^2(\theta) = \arg \min_{\theta \in \Theta} \sum_{i=1}^k \frac{(n_i - np_i(\theta))^2}{np_i(\theta)} .
$$
where \( n_i \) denotes the number of observations which fall in bin \( I_i \) and \( p_i(\theta) \) denotes \( \int_{I_i} f(x, \theta) \, dx \), this probability being fully determined by the assumptions of the model. Notice that MCS requires at least a pointwise function evaluation of \( X^2(\theta) \). But MCS still suffers from the need to know \( f(x, \theta) \), since \( p_i(\theta) = \int_{I_i} f(x, \theta) \, dx \). Where is the way out? Notice that for any choice of \( \theta \), one may obtain a ready estimate of \( p_i(\theta) \) via simulation

\[
p_i^* (\theta) := \text{proportion of simulations which fall in } I_i.
\]

As the number of simulations gets large \( p_i^* (\theta) \approx p_i(\theta) \). Thus the need to have \( f(x, \theta) \) in closed form may be avoided by the combination of \( a \) binning the observations and \( b \) simulation.
How does one do SIMEST?

There are perhaps two main approaches to MCS (hence to SIMEST) both depending on the binning approach one chooses. Since the model is fully specified, the bins may be determined in one of two ways (a) one may use the model and the data to construct the bins, or (b) one may ignore the model and use the data alone to construct the bins. In order for (a) to be feasible, one really needs to have \( f(x, \theta) \). However, if one has \( f(x, \theta) \) one would be better off with maximum likelihood than with MCS, as (of course) there is a loss of efficiency due to the binning. Within the SIMEST paradigm then, (b) is really the only feasible option. Using (b), the way SIMEST is done is depicted in Figure 1.2. One passes the data to some binning routine, in effect trading in the data for a set of \( k \) bins. One uses the (fully specified) model to obtain a black box simulator. The bins and the black box simulator are then used in conjunction with a search routine to hopefully find

\[
\hat{\theta}_n := \arg \min_{\theta \in \Theta} \sum_{i=1}^{k} \frac{(n_i - np_i(\theta))^2}{np_i(\theta)} .
\]

1.2 Lumpy Estimation Surfaces

In research related to the work of Atkinson, Bartoszyński, Brown and Thompson [1, 1983], there was much difficulty with the numerical optimization problem of finding

\[
\hat{\theta}_n := \arg \min_{\theta \in \Theta} \sum_{i=1}^{k} \frac{(n_i - np_i(\theta))^2}{np_i(\theta)} .
\]
The optimization routine was plagued by an estimation surface which seemed to be "egg carton"-like in shape with many local minimizers. In simulations, it was discovered that for large sample sizes these local minimizers disappeared, allowing the global minimizer to be quite easily found. This led the author's advisor (Thompson) to conjecture that perhaps such behavior occurs for many statistical problems. The practical implications of such a result are quite important and might be expressed as:

**CONJECTURE:** As statisticians we are rescued from the difficult problems of global optimization by large enough samples.\(^1\) Our hero is the strong law of large numbers.

\(^1\)For the non-statistician, the phrase "large enough samples" may also be interpreted to mean "given enough data."
For SIMEST, optimization difficulties are even more pronounced than for maximum likelihood (an example of such difficulties is given in Chapter 6 in our work extending the cancer model of Atkinson, Bartoszyński, Brown and Thompson [1, 1983]), hence the implications of such a result for SIMEST are extremely important from an applied standpoint. To attempt to establish this conjecture, it seemed natural to proceed in steps by solving the following problems:

**Problem 1.1** In maximum likelihood estimation for the location parameter of a Cauchy distribution, one notices that for small samples the likelihood surface exhibits many local maxima. Often these local maxima are a good distance away from the global maximum. But on any fixed compact set, say $\Gamma$, containing $\theta_0$, one notices that as the sample size increases all local maxima of the likelihood on $\Gamma$ tend to disappear leaving only the global maximum. Prove that this behavior will always occur for the Cauchy.

**Problem 1.2** Extend the ideas and proofs used for Problem 1.1 to the general case of maximum likelihood estimation for an arbitrary multivariate/multiparameter density.

**Problem 1.3** Extend the ideas and proofs used for Problem 1.2 to the case of minimum chi-square estimation (and related methods of estimation).

**Problem 1.4** Extend the ideas and proofs used for Problem 1.3 to the case of SIMEST.
Chapter 2 treats the theory for Problems 1.1 and 1.2. Some preliminary results related to Problem 1.3 were proved, but the author does not feel that any of these results are satisfactory at the present time. Thus results related to Problems 1.3 and 1.4 are not included in this dissertation.

1.3 Multivariate SIMEST

In Chapter 3, we propose the following method for multivariate SIMEST:

Concentric Ellipsoidal Binning

1. Standardize the dimensions to mean = 0, standard deviation = 1

2. Calculate some estimate of the center, or mode, call it \( \hat{\eta} \). For the models discussed in this dissertation, the simplest such \( \hat{\eta} \) was used, namely the mean.

3. Set \( i = 0 \)

   - Repeat until \( i = k - 1 \)
     
     - \( i = i + 1 \)
     
     - choose the \( m_i = i \cdot n/k \) nearest neighbors of \( \hat{\eta} \)
     
     - fit the ellipsoid of minimum volume containing these \( m_i \) neighbors

   - End repeat

4. This determines \( k - 1 \) ellipsoids, where each concentric "ring" contains roughly the proportion \( 1/k \) of the total number of observations.
While this idea seems very simple, no theoretical work has been done which can be used to directly justify such a method, although there is some work which is closely related. We state this as another problem:

**Problem 1.5** Provide as much theoretical justification as possible for multivariate simulation based estimation (SIMEST) using the proposed concentric ellipsoidal binning strategy.

Chapters 3 and 4 treat theory associated with the proposed method for multivariate SIMEST. Chapter 5 treats the application of the theory of Chapters 3 and 4 to a toy problem and a real problem. Chapter 6 treats an application of the proposed method for multivariate SIMEST to the cancer model of Atkinson, Bartoszyński, Brown, and Thompson [1, 1983].

### 1.4 Main Ideas of this Dissertation

The main ideas in this dissertation are:

1. It is intended to establish and/or provide proof of the conjecture that the problem of getting stuck at a local maximum (or minimum) of the estimation surface disappears for large samples. Specifically, the results obtained in this dissertation treat Problems 1.1 and 1.2.

2. It is intended to establish and/or provide theoretical justification for the nearest neighbor based (multivariate) binning strategy proposed for SIMEST. The theo-
retical justification given in this dissertation takes the form of strong consistency for the SIMEST parameter estimates based on multivariate data. Specifically, the results obtained in this dissertation treat Problem 1.5.

3. It is intended to apply the multivariate SIMEST approach proposed to an interesting application. The application used is the modeling of the secondary tumor generation mechanism of cancer, extending the work of Atkinson, Bartoszyński, Brown, and Thompson [1, 1983] to some multivariate situations.

1.5 Overview and Discussion of Literature

Comments on Chapter 2 and Appendix A

In the material of Chapter 2, quite a lot of the background material needed to justify the strong law for Banach-valued random variables has been included in Appendix A. Although this material is known to those who may be familiar with research topics in probability in Banach spaces, it is generally not part of a standard probability sequence. Still further, the material presented in Appendix A has been drawn from several sources none of which are quickly read. The layout of Appendix A is my own and seems the entirely logical (and rigorous) way to do it, but for some reason nothing so complete seems to be available. Perhaps it is considered trivial by some. As the reader unfamiliar with probability in Banach spaces may observe, one wonders if there is not a faster way to obtain the generalized strong laws. In fact, this is a good
criticism of Appendix A, and without an application of this theory to some Banach space besides $C(\Gamma)$,\(^2\) it would be hard to justify including Appendix A in the main development of Chapter 2. Hence the material has been placed in an appendix.

A back-of-the-envelope justification for the generalized strong law (for $C(\Gamma)$) might be obtained in a manner similar to that of Ferguson ([17, 1996, pp. 107-118]) although that material is not included here. One notices again (in the introduction of [17]) though that Ferguson owes a lot of his understanding to LeCam, which is quite clear from the Ferguson formulation. Thus the Ferguson material is really a simpler version of the more general Perlman techniques which we use in Section 2.2. With all this in mind, the reader can skip Appendix A, although the details given there are very satisfying mathematically if one can spare the time.

**What is and is not new on the global optimization problem**

The work of this dissertation began on the problem of global optimization in SIMEST (Problem 1.4). Closely related to this problem is the problem of global optimization in maximum likelihood estimation (Problem 1.2). The theory for SIMEST of course goes hand in hand with the theory of minimum chi-square (MCS) estimation (Problem 1.3). For binning the continuous problem as is done throughout this dissertation, it is intuitively obvious that as $k$ (the number of bins used for MCS) becomes large, MCS will approach the efficiency of maximum likelihood. Since more theoretical work

\(^2\) $C(\Gamma)$ denotes the space of continuous functions on a compact set $\Gamma$. 
exists for maximum likelihood, it was natural to start the investigation for Problem 1.4 with the work which has been done for maximum likelihood, i.e. start on Problem 1.2.

It turns out that Problem 1.2 is not established in the literature for the multiparameter/multivariate case. In Perlman [42, 1983], the results are established for univariate observations from a one parameter density. Although it does not seem that difficult to extend the ideas of that paper to the multiparameter/multivariate case, such an extension is not carried out in that paper (as is admitted in the “future work” section). This may be due to the fact that the multiple modes aspect of the problem in the 1983 paper [42] is not the main goal of that paper. The main goal of that paper is to answer the question: When is it the case that for large samples the only solution to the likelihood equation is the unique global maximizer of the likelihood surface? And to answer this question for the multiparameter/multivariate case would indeed become quite unpleasant, which is perhaps why Perlman never seems to return to the question in subsequent work (as evidenced by our literature review in Appendix C). Neither do any other authors return to treat the multiparameter/multivariate version of this problem (again, as evidenced by the literature review in Appendix C).

In some preliminary talks related to the work of Chapter 2 of this dissertation, it was suggested that some of these results may have already been done in the 70's. After doing much searching through the references which were suggested at that time (the results and details of this searching are not included in this dissertation), nothing
relevant was found. Although the literature search in Appendix C is exhaustive, it is only exhaustive in the sense that it aggressively follows up on any work which is related to Perlman [41, 1972] and Perlman [42, 1983]. It is of course possible that the multivariate/multiparameter version could have been established independently of Perlman [41, 42]. However, this possibility seems unlikely for the following reasons. First of all, the Perlman [41, 1972] paper is the culmination of a lot of similar work done by people at Berkeley, notably LeCam and others. Second of all, in the 1983 paper Perlman thanks several people for help and discussions on the topic at hand, among these people are Lehmann and Bahadur. Bahadur was one of the authors who had done quite a lot of work related to strong consistency of the m.l.e. which preceded the Perlman [41, 1972] paper, so he would be intimately familiar with the relevant literature. Since Lehmann is also at Berkeley (and also based on some sections in [30, 1991] which discuss the multiple mode problem), he must also be intimately familiar with the relevant literature. It is unlikely that authors of such caliber would be discussing ideas about a paper to be published in 1983 if the ideas had already been published in the 70's.

**What is and is not new on the multivariate SIMEST problem**

As pointed out in Appendix E, there has been quite a lot of work done which is related to minimum chi-square estimation in the goodness-of-fit literature. There are no fewer than two recent books concerning the goodness-of-fit literature (the Read
and Cressie book [47, 1988] and the Greenwood and Nikulin book [19, 1996]) as well as the good review paper by David Moore [35, 1978], which would seem to make the possibility that we have missed any work which is relevant as rather unlikely. Of great weight in this regard is the fact that the Pollard (1979) paper is mentioned only at the very end of a small section in the the Read and Cressie book in which the literature pertinent to the problem of data-dependent cells is discussed [47, 1988, pp. 150-51]. The thoroughness of the Read and Cressie book is remarkable, hence the fact that the “data-dependent cells” section of their literature review is so tiny lends substantial weight to our claims regarding what has and has not been done here. In the Read and Cressie book, as is the case in nearly all of the GOF literature, the following two aspects of the GOF problem are generally ignored:

- Bins are chosen depending upon the data which makes them random, and
- Typically, \( k \) will be chosen depending upon \( n \), and any reasonable strategy naturally allows \( k \to \infty \) as \( n \to \infty \).

The difficulty of rigorously treating the randomness of the bins and allowing \( k \to \infty \) at the same time is evidenced by the excellent paper of Drost [13, 1989], where with great effort some results along this line are established for the case of a location scale model for 1D data. The mere existence of the Drost [13, 1989] paper should add great weight to our claims about what has and has not been done theoretically relevant to MCS theory. To read some of the off-handed remarks in the literature about what has
been theoretically established, one would sometimes be led to believe that something like Drost [13, 1989] was a trivial application of some paper of the 50's. But such is not the case.

In Chapters 3 and 4, we treat the theory related to the strong consistency of SIMEST estimates for the case of multidimensional data. In discussions which have followed the presentation of these theoretical results in the past, there is often the inability on the part of some listeners to believe that something seemingly so trivial has not been done already. In answer to this, the Pollard (1979) paper treats strong consistency in a way which is not satisfactory from the point of view of SIMEST, namely he assumes that an estimator akin to the m.i.e. is also available. There is no doubt that in 1979 Pollard could have established the consistency results we are presenting here, but he clearly had no motivitation to do so. The problem was treated in that paper in as great a generality as the author believed necessary and the generality conceived there was not general enough to include SIMEST. Further, the Cressie and Read paper [12] was 1984, so the results we present in Chapter 4 at the very least could not have been established prior to 1984 (even ignoring the simulation aspects of the problem), since the Cressie-Read class did not exist prior to that time.

An additional objection to the theory presented in Chapters 3 and 4 is that no distributional theory is provided to accompany the strong consistency results. One would expect that obtaining the distributional results would be a straightforward Taylor series argument or something along those lines once one has already estab-
lished consistency. But (perhaps surprisingly again) this understanding is not correct (the author certainly wishes that it were), the arguments to get the distributional theory are non-trivial (as evidenced for example by some of the assumptions in Pollard [43, 1979]). For example, to use the theory in this dissertation one only needs to verify that certain classes of sets are polynomial classes.\textsuperscript{3} In order to get distributional results, these classes must additionally be what is referred to as Donsker classes (for the definition of a Donsker class, see, for example, [69, 1996]). The theory related to Donsker classes is involved with the distributional type theory in empirical processes. The distributional theory of empirical processes requires understanding such topics as the generalizations of Brownian motion to $P$-motion and similar probabilistic ideas.

It was decided that learning about such ideas would have taken the author (of this dissertation) substantially more time than the timetable for completing the dissertation at the present time would have allowed. Thus a rigorous strong consistency theory is a natural place to stop based on time constraints.

For the $k \to \infty$ problem, the author spent roughly four months in the derivations of Appendix F. At the present time, it is felt that these results are still not satisfactory. It was after completing that work that the author realized that those arguments cannot be easily adjusted to account for the randomness in the bins as was hoped. An important aspect of our treatment in Appendix F is that it does not assume a compact support for $f(x, \theta_0)$, rather $f(x, \theta_0)$ is allowed to have support on all of $\mathbb{R}$.

\textsuperscript{3}What is meant by a polynomial class will be discussed in Chapter 3.
This creates numerous difficulties, as the length of the treatment there suggests. As one illustration of these difficulties, it is of course commonly known that as $k \to \infty$, MCS approaches the efficiency of maximum likelihood. But if one attempted to find a rigorous proof of this fact for the case where $f(x, \theta_0)$ had support on all of $\mathbb{R}^d$, this author expects one would be looking a long time. I do not believe such a rigorous proof is published anywhere, or at least I have been unable to find one. The main source of the difficulty is that the arguments related to the convergence of Riemannian-type sums to the appropriate integrals are not easy to make rigorous in the case of infinite support.

As far as applications of parameter estimation for multivariate data using random bins are concerned, the work in the Moore and Stubblebine paper [33, 1981] and the Quiroz and Dudley paper [45, 1991] seems to suggest that very little has been published with regard to anything besides a multivariate normal density. In fact, the two papers [33, 45] deal only with goodness-of-fit, not with parameter estimation.
Chapter 2

Lumpy Likelihoods and Banach-valued Random Variables

In research related to the work of Atkinson, Bartoszyński, Brown and Thompson [1, 1983], there was much difficulty with the numerical optimization problem of finding

\[
\hat{\theta}_n := \arg \min_{\theta \in \Theta} \sum_{i=1}^{k} \frac{(n_i - np_i(\theta))^2}{np_i(\theta)}.
\]

The optimization routine was plagued by an estimation surface which seemed to be "egg carton"-like in shape with many local minimizers. In simulations, it was discovered that for large sample sizes these local minimizers disappeared, allowing the global minimizer to be quite easily found. This led the author's advisor (Thompson) to conjecture that perhaps such behavior occurs for many statistical problems. The practical implications of such a result are quite important and might be expressed as:

**CONJECTURE:** As statisticians we are rescued from the difficult problems of global optimization by large enough samples.\(^1\) Our hero is the strong law of large numbers.

For SIMEST, optimization difficulties are even more pronounced than for maximum likelihood (an example of such difficulties is given in Chapter 6 in our work on

\(^1\)For the non-statistician, the phrase "large enough samples" may also be interpreted to mean "given enough data."
the cancer model), hence the implications of such a result for SIMEST are extremely important from an applied standpoint.

In this chapter, we start with a prototypical example to illustrate these ideas, maximum likelihood estimation for the location parameter of a Cauchy distribution. While the Cauchy problem has been studied in several papers, it has generally never been approached from the standpoint of the conjecture above, as is done here. The generalization of the Cauchy result to the arbitrary 1-parameter, 1-dimensional data problem for maximum likelihood is given in Perlman [42, 1983]. In Section 2.2, we generalize this result to the multiparameter/multivariate case for maximum likelihood. Following this, in Section 2.3 we present perhaps a more useful (though less general) formulation of this result using the generalized strong law of large numbers, and we illustrate the usefulness of this second formulation with several examples in Section 2.4. In the interest of making this chapter relatively self-contained, many of the basics and some proofs of the relevant results regarding the generalized SLLN are included in Appendix A.

2.1 Example of Cauchy Location Likelihood

Consider the problem of estimating the location parameter of the Cauchy density via maximum likelihood. The density function for the Cauchy is

\[ f(x, \theta) = \frac{1}{\pi(1 + (x - \theta)^2)} \quad x \in (-\infty, \infty) \]
where $\theta \in (-\infty, \infty)$. The log likelihood function based on $n$ observations is

$$\log L_n(\theta) = \log \left( \prod_{j=1}^{n} \frac{1}{\pi(1 + (x_j - \theta)^2)} \right) = -\sum_{j=1}^{n} \log \left( \pi(1 + (x_j - \theta)^2) \right).$$

For small sample sizes, this likelihood function has the noteworthy property that it tends to exhibit several local maxima. We want to discuss whether these multiple maxima tend to persist or tend to disappear as we increase the sample size.

Observe that

$$E \frac{\partial}{\partial \theta} \log f(X, \theta) = \int \frac{2(x - \theta)}{(1 + (x - \theta)^2) \pi(1 + (x - \theta_0)^2)} dx = -\frac{2(\theta - \theta_0)}{4 + (\theta - \theta_0)^2}$$

and that

$$E \frac{\partial^2}{\partial \theta^2} \log f(X, \theta) = \frac{2(\theta - \theta_0)^2 - 4}{(4 + (\theta - \theta_0)^2)^2}.$$ 

For the purposes of illustration, we let $\theta_0 = 0$. A plot of both of these expectations is given in Figure 2.1.

By the ordinary strong law of large numbers for real-valued random variables,

$$\frac{1}{n} \frac{\partial}{\partial \theta} \log L_n(\theta) \to E_{\theta_0} \frac{\partial}{\partial \theta} \log f(X, \theta) = -\frac{2\theta}{4 + \theta^2} \quad a.s. \quad (2.1)$$

and

$$\frac{1}{n} \frac{\partial^2}{\partial \theta^2} \log L_n(\theta) \to E_{\theta_0} \frac{\partial^2}{\partial \theta^2} \log f(X, \theta) = \frac{2\theta^2 - 4}{(4 + \theta^2)^2} \quad a.s.$$ 

We observe that for the Cauchy, the a.s. limit given on the right hand side of (2.1) has only one zero (see Figure 2.1), and this zero is located at $\theta_0 = 0$. 

Figure 2.1

In Figure 2.2 we illustrate the behavior of a simulated sequence of realizations of $\frac{1}{n} \log L_n(\theta)$ based upon a simulated sequence of Cauchy observations with $\theta_0 = 0$. It would be nice if we could also plot the limit function (i.e. $E \log f(X, \theta)$) in this same figure; however, we cannot calculate the necessary integral to obtain it. In Figure 2.3 we plot the corresponding sequence of realizations of $\frac{1}{n} \frac{\partial}{\partial \theta} \log L_n(\theta)$. The limit function (i.e. $E \frac{\partial}{\partial \theta} \log f(X, \theta)$) can be calculated in this case, hence it is also plotted in Figure 2.3.

Suppose for the sake of the argument that we were attempting to maximize this likelihood function numerically, without plotting the likelihood surface to observe
Figure 2.2 Simulation of the convergence of the log-likelihood function to its a.s. limit. The legend gives the sample sizes used to construct each realization. Notice the local maximizers far from $\theta_0$ for the cases of $n = 3$ and $n = 5$. Also notice how they disappear for the larger samples of $n = 15$ and $n = 30$. 
Figure 2.3 Simulation of the convergence of the derivative of the log-likelihood to its a.s. limit. The legend gives the sample sizes used to construct each realization.
what is happening. In such a situation, we could easily find that our numerical procedure returns as its maximum value a $\hat{\theta}$ which corresponds to one of the lumps in the likelihood surface which are far from $\theta_0$. For example, in Figure 2.2 for the case $n = 3$, there is a local maximum of the likelihood near 15, quite far from the global maximum which appears to be near 6. Getting stuck at a local maximizer would be a real problem for the Cauchy likelihood if one could only maximize this likelihood numerically, without being able to look at a plot.

Notice, however, that by the time the sample size reaches $n = 15$, these local maximizers appear to have largely disappeared. Only the very pronounced global maximum remains. For $n = 30$, this trend continues.

2.2 Extending the Perlman [42, 1983] Formulation

In this section, we present some of the ideas from Perlman [42, 1983] which are relevant to our discussion. Also, we present some of our direct generalizations of his results to the multiparameter/multivariate case. The Perlman formulation is more general than the simpler formulation to follow in Section 2.3, although we observe that the formulation in Section 2.3 is much easier to apply.

Let $X_1, X_2, \ldots$ be an infinite sequence of $\mathbb{R}^d$-valued random variables with common density function $f(x, \theta_0)$. Assume that $\Theta$ is a subset of $\mathbb{R}^s$, possibly all of $\mathbb{R}^s$, and

---

2 Notice that if there are three or more parameters, such an approach is not unlike the approach one must take in practice; one cannot plot such a surface very easily.
that \( \theta_0 \) lies in the interior of \( \Theta \). For our purposes take \( \Gamma \) to be a compact subset of \( \Theta \) containing \( \theta_0 \).\(^3\) Denote by \( g(x, \theta) \) any real-valued function. Suppose that \( \Gamma \) and \( g(x, \theta) \) are such that \( \sup_{\Gamma} g_n(\theta) \) and \( \inf_{\Gamma} g_n(\theta) \) are measurable, where

\[
g_n(\theta) := \frac{1}{n} \sum_{j=1}^{n} g(X_j, \theta) \quad .
\]

Although the following two definitions can be stated more concisely together, it is easier to see them stated separately the first time through.

**Definition 2.1 (Perlman [42], 1983)** A function \( g(x, \theta) \) is dominated on \( \Gamma \) with respect to \( P_{\theta_0} \) if

\[
E_{\theta_0} \left[ \sup_{\Gamma} g_\nu(\theta) \right] < \infty
\]

for some integer \( \nu \geq 1 \). A function \( g(x, \theta) \) is locally dominated on \( \Gamma \) with respect to \( P_{\theta_0} \) if for each \( \theta' \in \Gamma \), there is a neighborhood \( \Gamma' \) of \( \theta' \) (i.e. \( \theta' \in \Gamma' \subseteq \Gamma \)) such that \( g(x, \theta) \) is dominated on \( \Gamma' \).

**Definition 2.2 (Perlman [42], 1983)** A function \( g(x, \theta) \) is dominated by 0 on \( \Gamma \) with respect to \( P_{\theta_0} \) if

\[
E_{\theta_0} \left[ \sup_{\Gamma} g_\nu(\theta) \right] < 0
\]

\(^3\)One may consider more general sets \( \Gamma \) under the Perlman formulation, but we do not wish to consider such cases.
for some integer \( \nu \geq 1 \). A function \( g(x, \theta) \) is **locally dominated** by 0 on \( \Gamma \) with respect to \( P_{\theta_0} \) if for each \( \theta' \in \Gamma \), there is a neighborhood \( \Gamma' \) of \( \theta' \) (i.e. \( \theta' \in \Gamma' \subseteq \Gamma \)) such that \( g(x, \theta) \) is dominated by 0 on \( \Gamma' \).

Although there is little intuition to be had here, it may be helpful to point out that there is no intuitive difference between *local dominance* and *dominance*. The *local* dominance (*local* dominance by 0) is largely a mathematical convenience. The definition of *local dominance* is a clever way of asking that an open covering of \( \Gamma \) exist with the property that on each set \( O_\gamma \) in the open covering, \( g(x, \theta) \) is dominated on \( O_\gamma \). If, in addition, \( \Gamma \) happens to be compact, we are guaranteed the existence of a finite sub-covering where \( \Gamma \subseteq \bigcup_{\gamma=1}^{m} O_\gamma \), and on each member of this finite subcovering \( g(x, \theta) \) is (still) dominated. It should be intuitively clear that \( g(x, \theta) \) is then dominated on \( \Gamma \), and in fact, this is not difficult to demonstrate. This point is summarized in the following result.

**Result 2.1 (Perlman [41], 1972)** If \( g(x, \theta) \) is locally dominated (by 0) on \( \Gamma \) and \( \Gamma \) is compact, then \( g(x, \theta) \) is dominated (by 0) on \( \Gamma \).

The motivation for having the (completely non-intuitive) condition on the expectation hold for any \( \nu \geq 1 \) in Definitions 2.1 and 2.2 can be traced back at least to Kiefer and Wolfowitz [27, 1956]. Perlman [41, 1972] discusses at length the difficulty in practice of determining the integer \( \nu \) which satisfies the expectation condition. Even for some easy examples, having the sup inside the expectation makes calculations
fairly difficult. With all these reservations in mind, the dominance ideas are very important mathematically. Consider the following results from Perlman [42, 1983].

**Lemma 2.1 (Perlman [42], 1983)** If \( g \) is dominated on \( \Gamma \), then

\[
P_0 \left\{ \lim_{n \to \infty} \sup_{\Gamma} g_n(\theta) = \lim_{n \to \infty} E_0 \sup_{\Gamma} g_n(\theta) \right\} = 1.
\]

**Lemma 2.2 (Perlman [42], 1983)** Suppose \( \Gamma \) is a compact subset of \( \Theta \) such that \( g \) is locally dominated on \( \Gamma \) and \( g(x, \cdot) \) is upper semicontinuous on \( \Gamma \) a.s. Then \( g \) is dominated on \( \Gamma \) and

\[
P_0 \left\{ \lim_{n \to \infty} \sup_{\Gamma} g_n(\theta) = \sup_{\Gamma} E_0 g(X, \theta) \right\} = 1,
\]

and \( E_0 g(X, \cdot) \) is upper semicontinuous on \( \Gamma \).\(^5\)

We use Lemma 2.1 to prove the next theorem. Lemma 2.2 was included here to illustrate that there is no way out of some type of assumption about the smoothness of \( g(x, \theta) \) if we wish to obtain strong law type results via Perlman's dominating approach.

**Theorem 2.1** Let \( X_1, X_2, \ldots \) be a sequence of random variables (or vectors) with common density \( f(x, \theta_0) \), where \( f(x, \theta_0) \) is a member of the family of probability densities \( \{ f(x, \theta) : \theta \in \Theta \} \). For each \( l = 1, \ldots, s \), suppose that for each \( \theta \in \Theta \)

---

\(^4\)This is Perlman's Lemma 2.5.

\(^5\)This is Perlman's Lemma 2.6.
$$g(x, \theta; l) := \frac{\partial}{\partial \theta_l} \log f(x, \theta)$$

exists and additionally that $\inf_{\Gamma} g_n(\theta; l)$ and $\sup_{\Gamma} g_n(\theta; l)$ are measurable, where $\Gamma$ denotes any compact subset of $\Theta$ containing $\theta_0$. If there is a finite partition $\Gamma_1, \ldots, \Gamma_m$ of $\Gamma \backslash B_0(\delta)$ such that for each $\Gamma_v$, there is an $l$ such that either

(a) $g(x, \theta; l)$ or
(b) $-y(x, \theta; l)$

is locally dominated by 0 on $\Gamma_v$, then for all $\delta$ such that $B_0(\delta)$ is properly contained in $\Gamma$,

$$P[\exists \text{ no stationary points of } L_n(\theta) \text{ in } \Gamma \backslash B_0(\delta) \text{ eventually}] = 1.$$  

Proof. Consider $\Gamma_v$. Suppose that case (a) holds (the argument is similar for case (b)). Local dominance by 0 on $\Gamma_v$ and compactness of $\Gamma_v \Rightarrow$ dominance by 0 on $\Gamma_v \Rightarrow$ dominance on $\Gamma_v$. Apply Lemma 2.1 to get that (abbreviate $g_n(\theta; l)$ to $g_n(\theta)$)

$$\lim_n \sup_{\Gamma_v} g_n(\theta) = \lim_n E \left[ \sup_{\Gamma_v} g_n(\theta) \right] \text{ a.s.}$$

$$\Rightarrow \lim_n \sup_{\Gamma_v} g_n(\theta) < 0 \text{ a.s. (2.2)}$$

The last line follows by the fact that $E \left[ \sup_{\Gamma_v} g_q(\theta) \right] < 0$ for some $q \geq 1$ via the dominance by 0 assumption and also by the fact that the expectation is non-increasing.
for $q' \geq q$. The non-increasing property of this expectation is remarkable and outlined more fully in Perlman [41, 1972]. Continuing, (2.2) implies that (by properties of $\lim$)

$$\sup_{\Gamma_v} g_n(\theta) < 0 \quad \text{a.s. eventually}$$

$$\Leftrightarrow \sup_{\Gamma_v} \frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial \theta_i} \log f(X_j, \theta) < 0 \quad \text{a.s. eventually}$$

But we can repeat this same argument for every $\Gamma_v$. Since there are only a finite number of the $\Gamma_v$'s, we observe that eventually the sample version of the gradient will always have at least one of its components bounded away from zero on $\Gamma \setminus B_0(\delta)$. Thus with probability 1, there will eventually be no stationary points of the likelihood function in this region.

Now we present another similar theorem which is slightly more general, but perhaps not importantly so.

**Theorem 2.2** Let $X_1, X_2, \ldots$ be a sequence of random variables (or vectors) with common density $f(x, \theta_0)$, where $f(x, \theta_0)$ is a member of the family of probability densities $\{f(x, \theta) : \theta \in \Theta\}$. For each $l = 1, \ldots, s$, suppose that for each $\theta \in \Theta$

$$g(x, \theta; l) := \frac{\partial}{\partial \theta_i} \log f(x, \theta)$$

exists and additionally that $\inf_{\Gamma} g_n(\theta; l)$ and $\sup_{\Gamma} g_n(\theta; l)$ are measurable, where $\Gamma$ denotes any compact subset of $\Theta$ containing $\theta_0$. If there is a finite partition $\Gamma_1, \ldots, \Gamma_m$ of $\Gamma \setminus \{\theta_0\}$ such that for each $\Gamma_v$, there is an $l$ such that either
(a) $g(x, \theta; l)$ or (b) $-g(x, \theta; l)$

is locally dominated by 0 on $\Gamma_v$, then for all $\delta$ such that $B_0(\delta)$ is properly contained in $\Gamma$,

$$P[\exists \text{ no stationary points of } L_n(\theta) \text{ in } \Gamma \setminus B_0(\delta) \text{ eventually}] = 1.$$ 

Proof. Let $\Gamma_v \equiv (\Xi_v \cap \Gamma) \setminus B_0(\delta)$, $v = 1, \ldots, m$. Then $\Gamma_1, \ldots, \Gamma_m$ is a partition of $\Gamma \setminus B_0(\delta)$. Since $\Gamma \setminus B_0(\delta)$ is compact, without loss of generality take each of the $\Gamma_v$'s to be compact also. But now the conditions in the previous theorem are satisfied, so we can apply that theorem to get the desired result. □□□

2.3 A Simpler Formulation

The formulation of the results in the present section is based on an application of strong laws of large numbers for Banach-valued random variables. These SLLN's are also referred to as generalized strong laws. For the reader not familiar with these generalized strong laws, we have included an extensive explication of the relevant definitions and ideas in Appendix A. The layout of Appendix A is our own, but is merely logical. We were unable to find anything similar in completeness to the explication we have given in Appendix A despite much reading, but perhaps the
material is considered trivial to people in the field. Theorems 2.3 and 2.4 are presented without proof, since the proofs are fairly simple with the machinery of Appendix A in hand.

**Theorem 2.3** Let \( \{f(x, \theta) : \theta \in \Theta\} \) be a parametric family of probability densities. Let \( X_1, X_2, \ldots \) be iid \((\mathbb{R}^d\text{-valued})\) random variables with common density function \( f(x, \theta_0) \), where \( \theta_0 \in \Theta \). Let \( \Gamma \) denote any compact subset of \( \Theta \) which is properly contained in \( \Theta \) and which contains \( \theta_0 \). If \( g : \mathbb{R}^d \to \Xi = C(\Gamma) \) is Borel measurable and \( \int \sup_{\Gamma} |g(x)| f(x, \theta_0) dx < \infty \) then

\[
\sup_{\Gamma} \left| \frac{1}{n} \sum_{j=1}^{n} g(X_j) - Eg(X) \right| \to 0 \quad \text{a.s.} \tag{2.3}
\]

**Theorem 2.4** Let \( \{f(x, \theta) : \theta \in \Theta\} \) be a parametric family of probability densities. Let \( X_1, X_2, \ldots \) be iid \((\mathbb{R}^d\text{-valued})\) random variables with common density function \( f(x, \theta_0) \), where \( \theta_0 \in \Theta \). Let \( \Gamma \) denote any compact subset of \( \Theta \) which is properly contained in \( \Theta \) and which contains \( \theta_0 \). If \( g : \mathbb{R}^d \to \Xi = C(\Gamma) \) is Borel measurable and \( \int \sup_{\Gamma} |g(x)| f(x, \theta_0) dx < \infty \) then (2.3) holds. If in addition \( Eg(X) = 0 \) only at \( \theta = \theta_0 \), then

\[
P\left( \frac{1}{n} \sum_{j=1}^{n} g(X_j) \neq 0 \text{ in } \Gamma \setminus B_0(\delta) \text{ eventually} \right) = 1 \tag{2.4}
\]

where \( B_0(\delta) = \{\theta : ||\theta - \theta_0|| < \delta\} \) and \( B_0(\delta) \) is properly contained in \( \Gamma \).
As an application of this theorem, we consider the case where

\[ g(x, \theta; l) = \frac{\partial}{\partial \theta_l} \log f(x, \theta) , \quad l = 1, \ldots, s . \]

With such a choice for \( g \), (2.4) says that for large enough samples, we will eventually find no local maxima of the likelihood surface in \( \Gamma \setminus B_0(\delta) \). Indeed it is enough to verify the condition about the 0's of \( Eg(X, \theta; l) \) for only one choice of \( l \) (assuming the conditions of Theorem 2.3 are satisfied for this same choice of \( l \)), since the gradient of \( n^{-1} \log L_n(\theta) \) fails to be 0 as soon as one of its components fails to be 0.\(^6\) We formalize this discussion in Corollary 2.1. Corollary 2.1 is not the most general statement possible (see Section 2.2), but it perhaps strikes a good balance between generality and utility.

**Corollary 2.1** Let \( X_1, X_2, \ldots \) be a sequence of random variables (or vectors) with common density \( f(x, \theta_0) \), where \( f(x, \theta_0) \) is a member of the family of probability densities \( \{ f(x, \theta) : \theta \in \Theta \} \). Let \( \Gamma \) denote any compact subset of \( \Theta \) which is properly contained in \( \Theta \) and which contains \( \theta_0 \). For each \( l = 1, \ldots, s \), suppose that for each \( \theta \),

\[ g(x, \theta; l) := \frac{\partial}{\partial \theta_l} \log f(x, \theta) \]

exists. Suppose that for some choice of \( l \), \( g(x, \theta; l) \) is a measurable function of \( x \) for each choice of \( \theta \), \( g(x, \theta; l) \) is continuous\(^7\) in \( \theta \) for each choice of \( x \),

\(^6\)These remarks are in the lim inf sense, of course.

\(^7\)Continuity can be relaxed somewhat, if desired.
\[ \int \sup_{\Gamma} |g(x, \theta; l)| f(x, \theta_0) dx < \infty, \]

and \( Eg(X, \theta; l) = 0 \) only at \( \theta = \theta_0 \). Then

\[ P \left[ \frac{1}{n} \sum_{j=1}^{n} g_l(X_j) \neq 0 \text{ in } \Gamma \setminus B_0(\delta) \text{ eventually} \right] = 1 \]

where \( B_0(\delta) = \{ \theta : \| \theta - \theta_0 \| < \delta \} \) and \( B_0(\delta) \) is properly contained in \( \Gamma \). In particular, there will eventually be no stationary points of the likelihood in \( \Gamma \setminus B_0(\delta) \).

2.4 Examples

\[ \text{Example 2.1} \]

In the Cauchy example (discussed in Section 2.1), let

\[ g(x) := \frac{\partial}{\partial \theta} \log f(x, \theta) = \frac{2(x - \theta)}{1 + (x - \theta)^2}. \]

Observe that \( g(x) \) is bounded on \( \mathbb{R} \times \mathbb{R} \), say by \( K \). Thus

\[ \int \sup_{\Gamma} |g(x)| f_0(x) dx \leq K \int f_0(x) dx = K < \infty \]

for any choice of \( \Gamma \). Since \( g(x) \) is continuous in \( x \) and \( \theta \), it satisfies the measurability requirements of Corollary 2.1. We showed in Section 2.1 that

\[ Eg(X) = -\frac{2(\theta - \theta_0)}{4 + (\theta - \theta_0)^2} \]
which is 0 if and only if \( \theta = \theta_0 \). Apply Corollary 2.1 to get that eventually there will be no stationary points of the likelihood in \( \Gamma \setminus B_0(\delta) \). This was illustrated in Figures 2.2 and 2.3.

Applying Corollary 2.1 to the Cauchy example was deceptively easy. Generally, we should not expect to be so lucky as to have \( \partial / \partial \theta \log f(x, \theta) \) bounded on \( \mathcal{R} \times \mathcal{R} \) or even on \( \Gamma \times \mathcal{R} \). The next example illustrates this point.

\[ \text{Example 2.2} \]

Consider the case of \( X \sim N(0, \theta) \),

\[
f(x, \theta) = \frac{1}{\sqrt{2\pi \sqrt{\theta}}} \exp \left( -\frac{1}{2} \frac{x^2}{\theta} \right) .
\]

As before, let

\[
g(x) := \frac{\partial}{\partial \theta} \log f(x, \theta) = \frac{\partial}{\partial \theta} \left( -\log \sqrt{2\pi} - \log \sqrt{\theta} - \frac{1}{2} \frac{x^2}{\theta} \right) = \frac{1}{2} \left( \frac{x^2 - \theta}{\theta^2} \right).
\]

Notice that \( g(x) \) is unbounded on \( \Gamma \times \mathcal{R} \) for any choice of \( \Gamma \); hence it is the weighting effect of \( f_0(x) \) which must come to the rescue. Observe that \( |x^2 - \theta| \leq x^2 + \theta \), so

\[
\sup_{\Gamma} |g(x)| \leq \sup_{\Gamma} \frac{1}{2\theta^2} \sup_{\Gamma} (x^2 + \theta) = c_1 (x^2 + c_2) 
\]

(2.5)

for some constants \( c_1 \) and \( c_2 \). The sup's are attained in (2.5) because \( \Gamma \) is compact and is properly contained in \( \Theta = (0, \infty) \). Thus

\[
\int_{\Gamma} \sup_{\Gamma} |g(x)| f_0(x) dx \leq c_1 \int (x^2 + c_2) f_0(x) dx = c_1 (E X^2 + c_2) = c_1 (\theta_0 + c_2) < \infty .
\]
Also,

\[ \int g(x)f_0(x)dx = \int \frac{1}{2\theta^2}(x^2 - \theta)f_0(x)dx = \frac{1}{2\theta^2}(\theta_0 - \theta) \]

which is 0 if \( \theta = \theta_0 \) and non-zero otherwise. Again, \( g(x) \) is continuous in \( x \) and \( \theta \), so the measurability requirements of Corollary 2.1 are satisfied. Apply Corollary 2.1.

Our last example is somewhat more involved as it deals with a less trivial problem - maximum likelihood estimation for the two parameter Weibull. Although still idealized, the Weibull is not an exponential family and additionally there is no sufficient statistic for \( \theta = (\rho, \alpha) \) (cf. Cox and Oakes [9, 1984, p. 41]), hence maximum likelihood estimation is reasonable. Further, the likelihood equations must be solved numerically (cf. Cox and Oakes [9, p. 42]). Observe that this two parameter example takes us beyond Perlman [42, 1983].

\[ \Box \text{ Example 2.3} \]

Suppose \( X \sim \text{Weibull}(\rho, \alpha) \), i.e.

\[ f(x) = \alpha \rho \rho^\alpha - 1 \exp \{-\rho x\} \quad , \quad \rho > 0 \quad , \quad \alpha > 0 \]

Observe that

\[ \log f(x) = \log \alpha + \log \rho + (\alpha - 1) \log \rho x - \rho x^\alpha \]

Let

\[ g_1(x) = \frac{\partial}{\partial \alpha} \log f(x) = \frac{1}{\alpha} + \log \rho x - \rho x^\alpha \log \rho x \]
\[ g_2(x) = \frac{\partial}{\partial \rho} \log f(x) = \frac{\alpha}{\rho} - \alpha x \rho x^{\alpha - 1} \cdot \]

For this problem, we calculate \(\int g_i(x) f_0(x) dx, i = 1, 2\), first, since some of this integrals will be used again later.

\[ \int g_2(x) f_0(x) dx = \int \frac{\alpha}{\rho} f_0(x) dx - \int \alpha x \rho x^{\alpha - 1} f_0(x) dx \]

\[ = \frac{\alpha}{\rho} - \int \alpha x \rho x^{\alpha - 1} \exp \{-(\rho_0 x)^{\alpha_0}\} \alpha_0 (\rho_0 x)^{\alpha_0 - 1} \rho_0 dx \]

and substitute \(u = (\rho_0 x)^{\alpha_0}\) to get

\[ = \frac{\alpha}{\rho} - \alpha \rho^{\alpha - 1} \int_0^\infty \left( \frac{1}{\rho_0} u^{1/\alpha_0} \right)^\alpha \exp(-u) du \]

\[ = \frac{\alpha}{\rho} - \alpha \rho^{\alpha - 1} \frac{1}{\rho_0^\alpha} \int_0^\infty u^{\alpha/\alpha_0} \exp(-u) du = \frac{\alpha}{\rho} - \frac{\alpha \rho^{\alpha - 1}}{\rho_0^\alpha} \Gamma \left( \frac{\alpha}{\alpha_0} + 1 \right) \quad (2.6) \]

which is clearly 0 for \((\alpha, \rho) = (\alpha_0, \rho_0)\) and non-zero (apparently, but don't have proof) otherwise. Also

\[ \int g_1(x) f_0(x) dx = \int \left( \frac{1}{\alpha} + \log \rho x - (\rho x)^\alpha \log \rho x \right) f_0(x) dx \]

\[ = \frac{1}{\alpha} + \int \log \rho x f_0(x) dx - \int (\rho x)^\alpha \log \rho x f_0(x) dx, \quad (2.7) \]

and
\[ \int \log \rho x f_0(x) \, dx = \log \rho + \int \log x f_0(x) \, dx = \log \rho + \int \log \left( \frac{1}{\rho_0} u^{1/\alpha_0} \right) \exp(-u) \, du \]

\[ = \log \rho + \log \frac{1}{\rho_0} + \frac{1}{\alpha_0} \int_0^\infty \log u \exp(-u) \, du = \log \frac{\rho}{\rho_0} - \frac{1}{\alpha_0} \gamma \]

where \( \gamma \) denotes Euler's constant. For the last piece of (2.7),

\[ \int (\rho x)^\alpha \log \rho x f_0(x) \, dx = \int (\rho x)^\alpha (\log \rho + \log x) f_0(x) \, dx \]

\[ = \log \rho \int (\rho x)^\alpha f_0(x) \, dx + \int (\rho x)^\alpha \log x f_0(x) \, dx \]

\[ = \log \rho \left( \frac{\rho}{\rho_0} \right)^\alpha \Gamma \left( \frac{\alpha}{\alpha_0} + 1 \right) + \int (\rho x)^\alpha \log x f_0(x) \, dx \quad (2.8) \]

by the same integration step as we did at (2.6). Treating the remaining integral in (2.8),

\[ \int (\rho x)^\alpha \log x f_0(x) \, dx = \rho^\alpha \int x^\alpha \log x f_0(x) \, dx \]

\[ = \rho^\alpha \int \left( \frac{1}{\rho_0} u^{1/\alpha_0} \right)^\alpha \log \left( \frac{1}{\rho_0} u^{1/\alpha_0} \right) \exp(-u) \, du \]

\[ = \rho^\alpha \int \left( \frac{1}{\rho_0} u^{1/\alpha_0} \right)^\alpha \log \left( \frac{1}{\rho_0} \right) \exp(-u) \, du + \rho^\alpha \int \left( \frac{1}{\rho_0} u^{1/\alpha_0} \right)^\alpha \frac{1}{\alpha_0} \log u \exp(-u) \, du \]
\[
\begin{align*}
&= \left( \frac{\rho}{\rho_0} \right)^{\alpha} \log \left( \frac{1}{\rho_0} \right) \int u^{\alpha/\alpha_0} \exp(-u) \, du + \left( \frac{\rho}{\rho_0} \right)^{\alpha} \frac{1}{\alpha_0} \int_{0}^{\infty} u^{\alpha/\alpha_0} \log u \exp(-u) \, du \\
&= \left( \frac{\rho}{\rho_0} \right)^{\alpha} \log \left( \frac{1}{\rho_0} \right) \Gamma \left( \frac{\alpha}{\alpha_0} + 1 \right) + \left( \frac{\rho}{\rho_0} \right)^{\alpha} \frac{1}{\alpha_0} S \left( \frac{\alpha}{\alpha_0} \right)
\end{align*}
\]  

where \( S(a) := \Gamma(a) \alpha \Psi(a) + \Gamma(a) \). Here \( \Psi(\cdot) \) represents the digamma function and the integral on the right in (2.9) was done in Maple. The digamma function has many properties, one of these being that \( \Psi(1) = -\gamma \). So \( S(1) = 1 - \gamma \). Also notice that \( \Gamma(a) a = \Gamma(a + 1) \) by properties of the \( \Gamma(\cdot) \) function. For convenience, let \( \rho_* = \rho/\rho_0 \), \( \alpha_* = \alpha/\alpha_0 \). Thus

\[
\int \rho^\alpha \log(\rho x) f_0(x) \, dx = \log \rho \Gamma \left( \alpha_* + 1 \right) \rho_*^\alpha + \rho_*^\alpha \log \frac{1}{\rho_0} \Gamma \left( \alpha_* + 1 \right) + \rho_*^\alpha \frac{1}{\alpha_0} S \left( \alpha_* \right)
\]

(2.10)

So finally,

\[
\int g_1(x) f_0(x) \, dx = \frac{1}{\alpha} + \log \rho_* - \frac{1}{\alpha_0} \gamma - \left\{ \log (\rho_*) \rho_*^\alpha \Gamma \left( \alpha_* + 1 \right) + \rho_*^\alpha \frac{1}{\alpha_0} S \left( \alpha_* \right) \right\}
\]

\[
= \frac{1}{\alpha} (1 - \alpha_* \gamma) + \log \rho_* - \log (\rho_*) \rho_*^\alpha \Gamma \left( \alpha_* + 1 \right) - \rho_*^\alpha \frac{1}{\alpha_0} S \left( \alpha_* \right)
\]

which is 0 for \( (\alpha, \rho) = (\alpha_0, \rho_0) \) and (again, apparently) is non-zero otherwise.

The reader will observe that (despite all the integration) all we have completed so far is \( \int g_1(x) f_0(x) \, dx \) and \( \int g_2(x) f_0(x) \, dx \). This is unfortunate, since we have come
to the more interesting part of the example. We now check that the dominating
conditions
\[
\int \sup_\Gamma |g_i(x)| f_0(x) \, dx < \infty, \quad i = 1, 2
\]
are satisfied (where \( \Gamma \) is any proper subset of \( \Theta \) containing \( \theta_0 \)). We start with \( g_2(\cdot) \) since it is easier. We abbreviate \( \sup_\Gamma \) with \( \sup \) since no other supremums are being taken.

\[
\int \sup_\Gamma |g_2(x)| f_0(x) \, dx = \int \sup_\Gamma \left| \frac{\alpha}{\rho} - \alpha x(\rho x)^{\alpha-1} \right| f_0(x) \, dx
\]

\[
\leq \int \sup_\Gamma \left| \frac{\alpha}{\rho} \right| f_0(x) \, dx + \int \sup_\Gamma \left| \alpha x(\rho x)^{\alpha-1} \right| f_0(x) \, dx \quad (2.11)
\]
The integral on the left in (2.11) is some constant, say \( c_1 \), because \( \sup |\alpha/\rho| \) is attained on \( \Gamma \) by compactness. The integral on the right in (2.11) is

\[
= \int \sup_\Gamma |\alpha \rho^{\alpha-1} x^{\alpha}| f_0(x) \, dx \leq \int \sup_\Gamma |\alpha \rho^{\alpha-1}| \sup_\Gamma |x^{\alpha}| f_0(x) \, dx = \int c_2 x^{\alpha_3} f_0(x) \, dx
\]
for some \( \alpha_3 \in \Gamma \). But

\[
\int c_2 x^{\alpha_3} f_0(x) \, dx = E[X^{\alpha_3}] = \rho_0^{-3} \Gamma \left( \frac{3}{\alpha_0} + 1 \right)
\]

because

\[
E[X^r] = \rho_0^{-r} \Gamma \left( \frac{r}{\alpha_0} + 1 \right), \quad r = 1, 2, \ldots
\]
in the case of the Weibull (cf. Kendall and Stuart [61]). In particular, all the moments
are finite. This takes care of \( \int \sup_\Gamma |g_2(x)| f_0(x) \, dx \). Now
\[
\int \sup |g_i(x)| f_0(x) \, dx = \int \sup_{\Gamma} \left| \frac{1}{\alpha} + \log \rho x - (\rho x)^\alpha \log \rho x \right| f_0(x) \, dx
\]

\[
\leq c_3 + \int \sup |\log \rho x| f_0(x) \, dx + \int \sup |(\rho x)^\alpha \log \rho x| f_0(x) \, dx \tag{2.12}
\]

We begin by treating the first integral in (2.12). We want to maximize $|\log \rho x|$ (over $\Gamma$) for each choice of $x$. So fix $x$ at $x_0$ and maximize $|\log \rho x_0|$ with respect to $\rho$. Well

\[
|\log \rho x_0| = \begin{cases} 
\log \rho x_0 & \text{if } \rho x_0 \geq 1 \\
-\log \rho x_0 & \text{if } \rho x_0 < 1
\end{cases}
\]

Observe that $\rho_1 < \rho_2$ implies that $\log \rho_1 x < \log \rho_2 x$ and that $-\log \rho_1 x > -\log \rho_2 x$.

So to make $\log \rho x$ as large as possible on $A(x) := [\rho : \rho x_0 \geq 1]$, we want $\rho$ as large as possible and to make $-\log \rho x$ as large as possible on $B(x) := [\rho : \rho x_0 < 1]$, we want $\rho$ as small as possible. Let $\rho_1^* = \max_{\Gamma} \rho$ and $\rho_2^* = \min_{\Gamma} \rho$. Notice that we can partition $\mathbb{R}^+$ into three regions, $(0, x_1]$, $[x_1, x_2]$ and $[x_2, \infty)$, where (denote these regions by I, II, and III, respectively)

I \quad (0, x_1] \quad A(x) = \emptyset, \quad B(x) \neq \emptyset

II \quad [x_1, x_2] \quad A(x) \neq \emptyset, \quad B(x) \neq \emptyset

III \quad [x_2, \infty) \quad A(x) \neq \emptyset, \quad B(x) = \emptyset

Also notice that $x_1 > 0$ and $x_2 < \infty$, since otherwise we violate either compactness of $\Gamma$ or the assumption that $\Gamma$ is properly contained in $\Theta$. Notice that
\[
| \log \rho x | = \begin{cases} 
- \log \rho x \leq - \log \rho^*_x & \text{on I} \\
\max\{- \log \rho x, \log \rho x\} \leq \max\{- \log \rho^*_x, \log \rho^*_x\} & \text{on II} \\
\log \rho x \leq \log \rho^*_x & \text{on III} 
\end{cases}
\]

so that

\[
\int_{\Gamma} \sup_x | \log \rho x | f_0(x) dx \leq 
\]

\[
\int_0^{x_1} - \log \rho^*_x f_0(x) dx + \int_{x_1}^{x_2} \max\{- \log \rho^*_x, \log \rho^*_x\} f_0(x) dx + \int_{x_2}^{\infty} \log \rho^*_x f_0(x) dx
\]

(2.13)

Let \( h_2(x) = - \log \rho^*_x \) and \( h_1(x) = \log \rho^*_x \). Since \( h_1 \) and \( h_2 \) are monotone in \( x \), there are only 4 possible situations on \([x_1, x_2]\). Either (a) \( h_2 < h_1 \), (b) \( h_1 < h_2 \), (c) \( h_2 < h_1 \) on \([x_1, a]\) and \( h_2 > h_1 \) on \([a, x_2]\) for some \( a \), or (d) \( h_1 < h_2 \) on \([x_1, a]\) and \( h_1 > h_2 \) on \([a, x_2]\) for some \( a \). So in any case, clearly all three integrals in (2.13) are finite so long as

\[
\int_0^{\infty} \log \rho x f_0(x) dx
\]

(2.14)

is finite for any choice of \( \rho_3 \). But we already showed that (2.14) is

\[
\log \frac{\rho}{\rho_0} - \frac{1}{\alpha_0 \gamma}
\]

which is finite. Thus \( \int_{\Gamma} \sup_x | \log \rho x | f_0(x) dx < \infty \). The arguments for
\[
\int_0^\infty \sup_{\Gamma} |(\rho x)^\alpha \log \rho x| f_0(x) dx
\]  
(2.15)

are similar. Observe that (2.15) is

\[
\leq c_1 \int_0^\infty x^{\alpha_1} \sup_{\Gamma} |\log \rho x| f_0(x) dx
\]

for some \(c_1\) and \(\alpha_1\) and (as above) it is sufficient to show that for any choice of \(\rho_3\)

\[
c_1 \int_0^\infty x^{\alpha_1} \log \rho_3 x f_0(x) dx
\]

(2.16)
is finite. But (2.16) is finite by (2.10). Lastly, it is again clear that \(g_1(x)\) and \(g_2(x)\) are continuous in all their arguments, so the measurability requirements of Corollary 2.1 are satisfied. Apply Corollary 2.1.

2.5 Discussion and Literature

The Cauchy example is well studied; the uniform convergence of the likelihood is demonstrated in several papers and certainly falls under the treatment in Perlman [42, 1983]. Generally, however, the relevance of the Cauchy example to the conjecture given in the introduction to this chapter is not noticed, nor is the practical relevance of being rescued from the difficult problems of global optimization by large sample sizes realized.

Our results represent a generalization of Perlman [42, 1983] to the multivariate, multiparameter case. I have devoted a substantial amount of time to searching the
literature for any subsequent work which may be related to Perlman [42, 1983] and Perlman [41, 1972]. A discussion and summary of this search is presented in Appendix C. The Perlman [42, 1983] paper is never subsequently referenced except perhaps once. Subsequent references to Perlman [41, 1972] are all related to generalizing the sufficient conditions for strong consistency given there; these references are discussed in Appendix D.

While the “dominance” and “local dominance” ideas in Perlman [41, 1972] allow for the greatest generality, I would argue that much of the relevance of such generality is lost on the types of problems considered in Perlman [42, 1983] and those in this chapter. If we are comfortable with the assumption that

\[ g_l(x) = \frac{\partial}{\partial \theta_l} \log f(x, \theta) , \quad l = 1, \ldots, s \]  

(2.17)

are continuous for each choice of \( x \) and measurable for each choice of \( \theta \), no more generality is needed than the ideas of the generalized SLLN. Removing ourselves from the quest for as much generality as possible, we are easily led to a more practical approach which requires simpler assumptions, as shown in Section 2.3. No matter how the problem is formulated, the fact remains that if \( E g_l(X) \) is not an integral we can evaluate, little can be said at all.
Chapter 3

A Strong Consistency Theorem for SIMEST

In this chapter, we propose a nearest neighbor based binning strategy for SIMEST using multidimensional data. While the proposed idea is simple, it happens that very little theoretical work has been done which could be quickly cited or easily used to rigorously justify such a method. Using the theoretical techniques of generalized Glivenko-Cantelli results (similar to applications by Pollard [43, 1979], Pollard [44, 1984], and Pakes and Pollard [40, 1989]), in this chapter we establish strong consistency for the SIMEST parameter estimates which result from the proposed method. The reader will observe that the multivariate SIMEST context is fairly remote from the applications envisioned by Pollard [43, 44, 1979, 1984] and Pakes and Pollard [40, 1989], which in some measure should explain why none of the consistency theory presented in those references can be considered satisfactory within the multivariate SIMEST context.

In the hope of making this chapter self-contained, much of the basics and relevant background concerning the generalized Glivenko-Cantelli results are included in Appendix B, much of which is adapted from Pollard [44, 1984]. For the reader interested in measurability issues, a great deal more on the topic of analytic sets can be found in Chapter 8 of Cohn [7, 1980], which is an excellent book.
3.1 Multivariate Binning Strategies

For univariate data, almost any reasonable binning strategy is likely to be successful (see Thompson and Tapia [64, 1990, p. 147]). For multivariate data, the situation becomes far more complicated. Consider, for example, a proposal to use an equi-spaced Cartesian mesh for higher dimensional data. Suppose one has 1000 observations, and one considers a mesh of 10 bins per dimension. Then the binning situation deteriorates rapidly with increasing dimension as indicated in Table 3.1.

Dividing the number of observations by the number of bins, one finds roughly 100 observations per bin for 1D, but only 1 observation per bin for 3D. Even for a modest dimensionality of 5D or 7D, the number of observations per bin becomes entirely unacceptable (note also that 1000 observations is a large data set by any practical standard). For multidimensional data, a regular mesh of bins is quickly overwhelmed by the number of empty bins as the dimensionality of the data increases. This behavior also reflects the NDE (non-parametric density estimation) guidelines which suggest that the amount of data required for NDE becomes enormous as the dimen-

<table>
<thead>
<tr>
<th>dimension of data</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>observations per bin</td>
<td>100</td>
<td>.1</td>
<td>.01</td>
<td>.0001</td>
</tr>
</tbody>
</table>

Table 3.1
sionality increases. This is one of the "curses of dimensionality". It is our position that for parameter estimation, additional dimensions of data can only be a blessing; estimation performance can only improve or at worst stay the same. Hence any binning method for which additional dimensions of data are perceived as a curse must be regarded as suspect.

With such ideas in mind, it seems difficult to persist in the naive thinking that an equi-spaced Cartesian mesh might perform satisfactorily in higher dimensions. Clearly it is not acceptable. The next obvious improvement to the fixed mesh approach is an adaptive mesh approach. One starts with the fixed mesh, but combines many of the bins in order to overcome the problem of having too many empty bins. Such an adaptive approach will offer some improvement, but presents non-trivial theoretical problems when attempting to justify the method. The bins in any adaptive approach become random variables, and this randomness of the bins is often neglected in theoretical treatments. Even in fixed-mesh treatments, the fact that the mesh of bins used is positioned in some data-based location is usually ignored.

Many multivariate data sets are observed to be roughly ellipsoidal, or transformations can nearly make them so (this fact is observed by Debrah Nolan in [39, 1991], in a paper which was recommended by C.R. Rao). Multivariate densities with strange structures such as multiple modes generally result from contamination of the density of interest with some other density, or by the mixing of several different densities in some fashion. For the types of data we are interested in (e.g. for the cancer model
in Chapter 6), we expect that there should be little or no contamination in the data, hence many of the densities of interest to us will be transformable to shapes which are roughly ellipsoidal.¹

A natural approach to exploratory NDE in higher dimensions is described in Chapter 6 of Thompson and Tapia [64, 1990]. This approach involves two phases. In the first phase, one attempts to determine the number and location of the modes of the density. The modes are the areas of interest, regions of activity likely to be separated from one another by vast regions of empty space. In the second phase, one attempts to carry out local estimation of the characteristics of each of these modes (working with each mode separately). This approach to high dimensional data is aimed directly at overcoming the problem of the vastness of the empty space in high dimensions. Research related to phase one of this approach has led to great successes (see Elliott [15, 1995]), where at the heart of the approach in Elliott [15] is the nearest neighbor based mean update algorithm (MUA) of Fwu, Tapia, and Thompson [18, 1981].

¹It has been pointed out by a committee member that generally mixing, contamination, and even some dependence seem to be characteristics of much data in cancer modeling. Hence we do not suggest that the remarks here apply to cancer modeling in general. Nevertheless, there are many problems for which these remarks do apply, the model considered in Chapter 6 being one example.
3.2 Concentric Ellipsoidal Binning for Multivariate SIMEST

Based on the success of the MUA as well as the likelihood that many densities of interest will be roughly ellipsoidal, the following binning approach would seem to be a promising idea.

**Concentric Ellipsoidal Binning**

1. Standardize the dimensions to $\text{mean} = 0$, $\text{standard deviation} = 1$

2. Calculate some estimate of the center, or mode, call it $\hat{\eta}$. For the models discussed in this dissertation, the simplest such $\hat{\eta}$ was used, namely the mean.

3. Set $i = 0$
   
   $\cdot$ Repeat until $i = k - 1$
   
   $- i = i + 1$
   
   $- \text{choose the } m_i = i \cdot n/k \text{ nearest neighbors of } \hat{\eta}$
   
   $- \text{fit the ellipsoid of minimum volume containing these } m_i \text{ neighbors}$
   
   $\cdot$ End repeat

4. This determines $k - 1$ ellipsoids, where each concentric "ring" contains roughly the proportion $1/k$ of the total number of observations.

Of course, other bin shapes might reasonably be suggested. Ellipsoids have the advantage that (a) they are the first shape which comes to mind after spheres, (b) they are
computationally tractable, (c) they are widely familiar in the statistical community, and (d) they extend easily to dimensions beyond three.

This proposal for a multivariate SIMEST methodology could hardly be simpler or more intuitively reasonable. Surprisingly, despite the vast amount of research in the goodness-of-fit area, no theory currently available can be cited as justification that such an approach will work, although there is some work which is closely related.

3.3 Theoretical Issues Involved in Our Methods

Recall from the Introduction (Chapter 1), that the SIMEST parameter estimates are determined via:

\[
\hat{\theta}_n := \arg \min_{\theta \in \Theta} \sum_{i=1}^{k} \frac{(N_i - np_i^*(\theta))^2}{np_i^*(\theta)}
\]  

(3.1)

which is merely the simulated equivalent to minimum chi-square (MCS).\(^2\) Notice that this criterion almost appears dimension free. Of course it is not. Here \(\dim(\Theta) = s\) (i.e. \(\theta\) is some \(\mathbb{R}^s\) vector), and the bins \(I_1, I_2, \ldots, I_k\) constitute a partition of \(\mathbb{R}^d\).

Consistency is the first basic requirement one asks for in any statistical estimation procedure. We would like to be able to say that the SIMEST parameter estimates which result from the method proposed in Section 3.2 are consistent. Although there

---

\(^2\) Other criterion functions can also be considered. See Chapter 4 for example. The discussion here is clearer if attention is restricted to Pearson's \(X^2\). Note, however, that the theory of the present chapter establishes results for the simulated multinomial maximum likelihood estimator (SMML). The same results for simulated MCS are given in Chapter 4. This is due to the fact that the SMML proof was developed first.
is a vast literature related to goodness-of-fit theory, unfortunately nearly all of it does not apply to a situation such as ours. We will have to establish consistency directly.

The first theoretical difficulty of our proposed method is that choosing the ellipsoidal regions in a nearest neighbor fashion as described causes the bins to be random variables. In particular, the bin counts are no longer multinomial. The second difficulty caused by our proposed method is that the criterion (3.1) is simulation-based. An example of the particularly vexing behavior that this brings about is that even when

$$\hat{\theta}_n := \arg \min_{\theta \in \Theta} \sum_{i=1}^{k} \frac{(N_i - np_i(\theta))^2}{np_i(\theta)} \tag{3.2}$$

is continuous, usually (3.1) will not be.

As justification, in part, for the claim that our theoretical results are new, it is appropriate at this time to discuss all of the relevant literature.

**Multivariate data-based binning**

The Glivenko-Cantelli results discussed in Appendix B evolved from ideas in the Vapnik and Červonenkis [70, 1971] paper (cf. Pollard [44, 1984]). Hence nothing like the results we present for concentric ellipsoidal bins can predate 1971. Further, theoretical results for random bins in $\mathbb{R}^d$ of non-rectangular shape were not established until Pollard's [43, 1979] very important paper (this is evidenced by the fact that

---

3 Some additional justification will also be given in a separate Section, Section 3.5

Our strong consistency results for multinomial maximum likelihood estimation (MML) given in this chapter are different from those of Pollard [43, 1979]. Pollard treats consistency in a different way. Specifically, he assumes that one starts with the fully parametric MLE (or an estimator with equivalent properties). This manner of treating consistency is not satisfactory from the point of view of SIMEST, as the whole point of SIMEST is avoiding the need to have the fully parametric likelihood in closed form. The proofs of our MML theorems (Theorems 3.1 and 3.2) are modeled completely on Rao's [46, 1965] proof (Rao's proof is for the case of non-random bins). Further, our proof of strong consistency for the Cressie-Read class in Chapter 4 is clearly beyond the ambit of anything prior to 1984 (as the Cressie-Read class was not established until 1984). Pollard [43, 1979] left many possible variations and generalizations of his work for another time. Our direct proofs of consistency for MML are a variation on Pollard [43, 1979], and our direct proofs of consistency for the Cressie-Read class are generalizations of Pollard [43, 1979]. These are new variations and generalizations even without considering the fact that we have additionally treated the simulation-based aspects of the problem (in Theorem 3.3). The simulation-based aspects of SMML are not treated in Pollard [43, 1979].

To reiterate, the important binning aspects that distinguish our results from the great vastness of the GOF literature is that (a) our bins can be chosen in a heavily
data-based manner (data-based bins which are chosen independently of the model assumptions and/or the model parameter estimates distinguish us from most of the random bins papers in the GOF literature) and (b) our bins can be constructed from non-rectangular shapes (non-rectangular bins distinguish our work from all work prior to 1979).

Simulation based criterion function

The simulation based aspects of our results are adapted from Pakes and Pollard [40, 1989]. However, they consider the simulation aspects of a problem where the bins are non-random (cf. Section 3.5). So they treat, for example, weak consistency for simulated analogues to MCS, MMCS, MML for the fixed-bins case.4

Combined aspects

We need strong consistency where the bins are multivariate and random, and the estimation surface is simulation based. As described above, Pollard [44, 1979] treats only the randomness in the multivariate bins, and Pakes and Pollard [40, 1989] treats only the simulation-based component of the fixed-bins problem. Our results provide a different treatment of consistency for the multivariate, random bins case than that given in Pollard [44, 1979]. In addition, we provide results which simultaneously treat (i) the multivariate, random bins and (ii) the error due to simulation.

---
4These abbreviations stand for Minimum chi square (MCS), Modified minimum chi square (MMCS), and Multinomial Maximum Likelihood (MML), respectively. The treatment of MML is only oblique.
3.4 Consistency Results

**Theorem 3.1** Let \( \{P_\theta : \theta \in \Theta\} \) be a family of probability measures on \( S \), with sigma-finite dominating measure \( \mu \), where the support of the family does not depend upon \( \theta \). Suppose \( X_1, X_2, \ldots \) is a sequence of iid observations with common distribution \( P_{\theta_0} \) (where \( P_{\theta_0} \in \{P_\theta : \theta \in \Theta\} \)). Denote by \( S \) a \( \sigma \)-algebra of \( \mu \)-measurable subsets of \( S \) (e.g. \( S = \mathbb{R}^d \), \( S = \mathcal{B}^d \), \( \mu = \) Lebesgue measure on \( \mathbb{R}^d \)). Let \( \mathcal{C} \) be a subclass of \( S \) which is permissible and has polynomial discrimination. Denote by \( I_i \) the \( i \)th bin of a partition of \( S \) into \( k \) sets from \( \mathcal{C} \). Let \( p_i(\theta) = P_\theta(I_i) \) and \( p_i^0(\theta) = P_{\theta_0}(I_i^0) \).\(^5\) Suppose that the following conditions are satisfied

1. Each \( I_i \) may be random so long as it is consistent for some \( I_i^0 \) (where \( I_i^0 \) denotes the region of \( S \) one attempts to estimate with \( I_i \)) in the sense that

\[
\mu(I_i \triangle I_i^0) \to 0, \quad i = 1, \ldots, k \quad a.s. \quad P_{\theta_0}
\]

as \( n \to \infty \), where \( \triangle \) is the symmetric difference operator, i.e. \( A \triangle B = (A \cap B^c) \cup (B \cap A^c) \) for \( \mu \)-measurable sets \( A \) and \( B \).\(^6\)

2. \( p_i^0(\theta) \neq p_i^0(\theta_0) \) for some \( i \) whenever \( \theta \neq \theta_0 \) (this is the weak identifiability condition of Rao), and

---

\(^5\)\( I_i^0 \) denotes the region of \( S \) one attempts to estimate with \( I_i \).

\(^6\)Actually, a little more than this may be needed, the argument for Lemma 3.1 below is not entirely correct as written without the additional assumption that \( f(x, \theta) \) is bounded above on any region \( A \times \Gamma \) where \( A \) is compact in \( \mathbb{R}^d \) and \( \Gamma \) is compact in \( \Theta \). For example, if the \( X_i \)'s are from a continuous density, it would be enough to additionally require only that \( f(x, \theta) \) is continuous in \( x \) and \( \theta \). This oversight was noticed by the author when it was too late (due to deadlines) to carefully reconsider these details and the necessary adjustments to the assumptions of this theorem. Such corrections will be made elsewhere.
3. \( p_i(\theta), \ i = 1, \ldots, k, \) admit first order partial derivatives,

then there is a (strongly) consistent root of the (multinomial) likelihood equations

\[
\frac{\partial}{\partial \theta_l} \sum_i \frac{N_i}{n} \log p_i(\theta) , \ l = 1, \ldots, s ,
\]  

(3.3)

where \( N_i \) denotes the number of observations which fall in \( I_i. \)

Proof. Pick \( \delta > 0. \) We will demonstrate that for large enough \( n, \) we can guarantee with probability one that there will be a root of (3.3) within the sphere \( G = \{ \theta : \| \theta - \theta_0 \| \leq \delta \}. \) Over \( G \) consider the function

\[
\sum_i p_i^0(\theta_0) \log \frac{p_i(\theta_0)}{p_i^0(\theta_0)} .
\]  

(3.4)

By assumption 3, (3.4) is continuous on \( G. \) Let \( G^* = \{ \theta : \| \theta - \theta_0 \| = \delta \}. \) The weak indentifiability condition, compactness of \( G^* \), the information inequality,\(^7\) and the continuity of the \( p_i^0(\theta) \)'s ensures that

\[
\inf_{\| \theta - \theta_0 \| = \delta} \sum_i p_i^0(\theta_0) \log \frac{p_i(\theta_0)}{p_i^0(\theta)} \geq \epsilon_1 > 0
\]  

(3.5)

for some \( \epsilon_1 > 0. \) Choose \( \epsilon_2 \in (0, \epsilon_1). \) Choose \( \eta \) small enough that

\(^7\)The “information inequality” referred to is that

\[
\sum_i p_i^0(\theta_0) \log \frac{p_i(\theta_0)}{p_i^0(\theta)} \geq 0
\]

with equality if and only if \( p_i^0(\theta_0) = p_i^0(\theta) \) for all \( i. \) The weak identifiability assumption ensures that the inequality is strict for any \( \theta \) chosen from \( G^*. \) Then continuity of the \( p_i^0(\theta) \)'s and the compactness of \( G^* \) ensure that a smallest value is attained.
0 < \epsilon_2 \leq \sum_i p_i^\circ(\theta_0) \log \frac{p_i^\circ(\theta_0)}{p_i^\circ(\theta)} - \eta \sum_i \log \frac{p_i^\circ(\theta_0)}{p_i^\circ(\theta)} , \quad \forall \theta \in G^* , \quad (3.6)

To see that such a choice of \eta is possible, observe that on G^*

\[ \sum_i \log \frac{p_i^\circ(\theta_0)}{p_i^\circ(\theta)} \leq c_1 \]

by continuity, where c_1 may be positive or negative. Pick \eta > 0. Then

\[ \eta \sum_i \log \frac{p_i^\circ(\theta_0)}{p_i^\circ(\theta)} \leq \eta c_1 \]

\[ \Rightarrow \sum_i p_i^\circ(\theta_0) \log \frac{p_i^\circ(\theta_0)}{p_i^\circ(\theta)} - \eta \sum_i \log \frac{p_i^\circ(\theta_0)}{p_i^\circ(\theta)} \geq \epsilon_1 - \eta c_1 . \quad (3.7) \]

Notice that (3.7) is true for each \eta > 0. Since \epsilon_2 \in (0, \epsilon_1), we may choose an \eta so that \epsilon_1 - \eta c_1 \geq \epsilon_2. Thus the choice of \eta for (3.6) is possible.

Since C is a permissible class with polynomial discrimination (assumption 1), we may apply Theorem B.2 to obtain

\[ \sup_{I_i \in C} |P_n I_i - PI_i| \to 0 \quad \text{a.s.} \quad (3.8) \]

or, to put (3.8) in a more familiar form,

\[ \sup_{I_i \in C} \left| \frac{1}{n} N_i - p_i(\theta_0) \right| \to 0 \quad \text{a.s.} \]

Thus there is n_1 such that

\[ p_i(\theta_0) - \eta \leq \frac{N_i}{n} \leq p_i(\theta_0) + \eta \quad \text{a.s.,} \quad \forall n \geq n_1 , \forall I_i \in C . \]
It follows that (almost surely $\forall n \geq n_1$ and for any partition of $S$ into $C$ sets)

$$\sum (p_i(\theta_0) - \eta) \log \frac{p_i(\theta_0)}{p_i(\theta)} \leq \sum \frac{N_i}{n} \log \frac{p_i(\theta_0)}{p_i(\theta)} \leq \sum (p_i(\theta) + \eta) \log \frac{p_i(\theta_0)}{p_i(\theta)} .$$  \hspace{1cm} (3.9)

We will return to (3.9) later.

**Lemma 3.1** Let $D_n^i := I_i \triangle I_i^0$. Then $\sup_{\theta \in \Theta} |p_i(\theta) - p_i^0(\theta)| \to 0$.

**Proof.** Write $I_{in}$ instead of $I_i$ with the additional subscript of "n" being added to emphasize the dependence of $I_i$ on $n$. Let

$$A_{1n}^i = \{x : \chi(x; I_{in}) = 1, \chi(x; I_i^0) = 0\}$$

and

$$A_{2n}^i = \{x : \chi(x; I_{in}) = 0, \chi(x; I_i^0) = 1\}$$

where $\chi(x; A)$ denotes the characteristic function for the set $A$ (i.e. $\chi(x; A)$ is 1 on $A$ and 0 off of $A$).\textsuperscript{8} Observe that

$$p_i(\theta) - p_i^0(\theta) = \int f(x, \theta)\chi(x; I_{in}) - f(x, \theta)\chi(x; I_i^0) \mu(dx) ,$$

since the integrand is 0 on $(A_{1n}^i \cup A_{2n}^i)^c$, $f(x, \theta)$ on $A_{1n}^i$, and $-f(x, \theta)$ on $A_{2n}^i$. Thus

$$|p_i(\theta) - p_i^0(\theta)| = \left| \int_{A_{1n}^i} f(x, \theta) \mu(dx) - \int_{A_{2n}^i} f(x, \theta) \mu(dx) \right|$$

\textsuperscript{8}The clash of the choice of the bin notation $I_i$ with the usual notation for indicator functions is unfortunate. However, this clash does not arise very often. Apologies to the reader.
\[
\leq \int_{A_{1n}} f(x, \theta) \mu(dx) + \int_{A_{2n}} f(x, \theta) \mu(dx) = \int_{A_{1n} \cup A_{2n}} f(x, \theta) \mu(dx)
\]
\[
= \int_{D_n^i} f(x, \theta) \mu(dx) = P_\theta(D_n^i)
\]

Hence

\[
\sup_{\theta \in \Theta} |p_i(\theta) - p_\theta^i(\theta)| \leq \sup_{\theta \in \Theta} P_\theta(D_n^i) \leq \mu(D_n^i) \to 0
\]

by assumption 1.

Now let

\[
r_f(\theta, n) = \sum (p_i(\theta_0) - \eta) \log \frac{p_i(\theta_0)}{p_i(\theta)}
\]

and let

\[
l_f(\theta) = \sum (p_\theta^i(\theta_0) - \eta) \log \frac{p_\theta^i(\theta_0)}{p_\theta^i(\theta)}
\]

where the notations \(r_f\) and \(l_f\) are used to serve as reminders for "random function" and "limit function", respectively. We intend to show (see Lemma 3.5) that

\[
\sup_{\theta \in \Theta} |r_f(\theta, n) - l_f(\theta)| \to 0,
\]

but we require some more lemmas first.

**Lemma 3.2** If \(\Gamma\) is compact, then \(\sup_{\theta \in \Theta} |p_i(\theta) - p_\theta^i(\theta)| \to 0\) a.s. ensures that

\[
\sup_{\Gamma} \left| \frac{p_i(\theta)}{p_\theta^i(\theta)} - 1 \right| \to 0 \quad a.s.
\]
Proof. Since our theorem assumes that the support does not depend on $\theta$, it follows that $p_i(\theta)$ must be bounded below on any compact subset $\Gamma$ of $\Theta$. Thus

$$\sup_{\Gamma} |\frac{p_i(\theta)}{p_i^\circ(\theta)} - 1| \leq \sup_{\Gamma} \frac{1}{p_i^\circ(\theta)} \sup_{\Gamma} |p_i(\theta) - p_i^\circ(\theta)| \leq c_\Gamma \sup_{\Gamma} |p_i(\theta) - p_i^\circ(\theta)| \to 0 \text{ a.s.}$$

where $c_\Gamma$ is a finite constant. □

**Lemma 3.3**

$$\limsup_{\theta} \log \frac{p_i(\theta)}{p_i^\circ(\theta)} \leq 0 \quad \text{and} \quad \liminf_{\theta} \log \frac{p_i(\theta)}{p_i^\circ(\theta)} \geq 0,$$

where $\inf_{\theta}$ and $\sup_{\theta}$ are taken over $\Gamma$, and $\lim$ and $\lim$ are taken with respect to $n$.

Proof. By Lemma 3.2, for each $\epsilon > 0$, $\exists n(\epsilon)$ such that

$$1 - \epsilon \leq \frac{p_i(\theta)}{p_i^\circ(\theta)} \leq 1 + \epsilon, \quad \forall \theta \in \Gamma, n \geq n(\epsilon).$$

Pick $\eta$. Choose $\epsilon_\eta$ such that $-\eta \leq \log(1 - \epsilon_\eta)$. Choose $n(\epsilon_\eta)$ such that

$$1 - \epsilon_\eta \leq \frac{p_i(\theta)}{p_i^\circ(\theta)}, \quad \forall \theta \in \Gamma, \ n \geq n(\epsilon_\eta),$$

$$\implies -\eta \leq \log \frac{p_i(\theta)}{p_i^\circ(\theta)}, \quad \forall \theta \in \Gamma, \ n \geq n(\epsilon_\eta),$$

$$\implies -\eta \leq \inf_{\theta} \log \frac{p_i(\theta)}{p_i^\circ(\theta)}, \quad n \geq n(\epsilon_\eta),$$

$$\implies -\eta \leq \liminf_{\theta} \log \frac{p_i(\theta)}{p_i^\circ(\theta)}.$$


But $\eta$ was arbitrary, so

$$\liminf_{\theta} \log \frac{p_i(\theta)}{p_i^0(\theta)} \geq 0.$$ 

Similarly,

$$\frac{p_i(\theta)}{p_i^0(\theta)} \leq 1 + \epsilon_\eta , \forall \theta \in \Gamma , \ n \geq n(\epsilon_\eta)$$

$$\implies \log \frac{p_i(\theta)}{p_i^0(\theta)} \leq \log(1 + \epsilon_\eta) \leq \eta , \forall \theta \in \Gamma , \ n \geq n(\epsilon_\eta)$$

$$\implies \limsup_{\theta} \log \frac{p_i(\theta)}{p_i^0(\theta)} \leq \eta$$

$\eta$ arbitrary

$$\limsup_{\theta} \log \frac{p_i(\theta)}{p_i^0(\theta)} \leq 0.$$ 

$\square$

Lemma 3.4

$$\limsup_{n \to \infty} \sup_{\theta} \left| \log \frac{p_i(\theta)}{p_i^0(\theta)} \right| = 0$$

Proof. Observe that

$$0 \leq \left| \log \frac{p_i(\theta)}{p_i^0(\theta)} \right| \leq \max \left\{ \log \frac{p_i(\theta)}{p_i^0(\theta)}, -\log \frac{p_i(\theta)}{p_i^0(\theta)} \right\}$$

$$\implies 0 \leq \sup_{\theta} \left| \log \frac{p_i(\theta)}{p_i^0(\theta)} \right| \leq \max \left\{ \sup_{\theta} \log \frac{p_i(\theta)}{p_i^0(\theta)}, \sup_{\theta} -\log \frac{p_i(\theta)}{p_i^0(\theta)} \right\}$$
\[ \Rightarrow \quad 0 \leq \limsup_{\theta} \left| \log \frac{p_i(\theta)}{p_0^i(\theta)} \right| \leq \max \left\{ \limsup_{\theta} \log \frac{p_i(\theta)}{p_0^i(\theta)}, \limsup_{\theta} - \log \frac{p_i(\theta)}{p_0^i(\theta)} \right\} \leq 0 \]

by Lemma 3.3. Since

\[ 0 \leq \limsup_{\theta} \left| \log \frac{p_i(\theta)}{p_0^i(\theta)} \right| \leq \limsup_{\theta} \left| \log \frac{p_i(\theta)}{p_0^i(\theta)} \right| \]

we get that

\[ \limsup_{\theta} \left| \log \frac{p_i(\theta)}{p_0^i(\theta)} \right| = 0 \]

as well. In particular, the limit exists.

Lemma 3.5 \( \sup_{\theta \in G} |r_f(\theta, n) - l_f(\theta)| \to 0 \)

Proof. By assumption, the support of \( P_\theta \) does not depend upon \( \theta \) and \( G^* \) is compact, so by Lemmas 3.2, 3.3, and 3.4, we get

\[ \sup_{\theta \in G^*} |\log p_i(\theta) - \log p_0^i(\theta)| \to 0 \quad \text{as} \quad n \to \infty \tag{3.10} \]

and

\[ \sup_{\theta \in G^*} \left| (p_i(\theta_0) - \eta) \log p_i(\theta) - (p_0^i(\theta_0) - \eta) \log p_0^i(\theta) \right| \leq \]

\[ \sup_{\theta \in G^*} \left| (p_i(\theta_0) - \eta) \log p_i(\theta) - (p_0(\theta_0) - \eta) \log p_0^i(\theta) \right| \tag{3.11} \]
\[ + \sup_{\theta \in G^*} |(p_i(\theta_0) - \eta) \log p_i^0(\theta) - (p_i^0(\theta_0) - \eta) \log p_i^0(\theta)| . \quad (3.12) \]

And by (3.10), (3.11) is

\[ = (p_i(\theta_0) - \eta) \sup_{\theta \in G^*} |\log p_i(\theta) - \log p_i^0(\theta)| \to (p_i^0(\theta_0) - \eta) \cdot 0 = 0 . \]

And (3.12) is

\[ = \sup_{\theta \in G^*} |\log p_i^0(\theta) (p_i(\theta_0) - \eta - (p_i^0(\theta_0) - \eta))| = |p_i(\theta_0) - p_i^0(\theta_0)| \sup_{\theta \in G^*} |\log p_i^0(\theta)| \to 0 \]

because \( \sup_{\theta \in G^*} |\log p_i^0(\theta)| < \infty \) since \( G^* \) is compact.\(^9\) Thus\(^10\)

\[ \sup_{\theta \in G^*} \left| \sum (p_i(\theta_0) - \eta) \log p_i(\theta) - \sum (p_i^0(\theta_0) - \eta) \log p_i^0(\theta) \right| \to 0 . \quad (3.13) \]

Also,

\[ \sum (p_i(\theta_0) - \eta) \log p_i(\theta_0) \to \sum (p_i^0(\theta_0) - \eta) \log p_i^0(\theta_0) . \quad (3.14) \]

Hence

\[ \sup_{\theta \in G^*} |r_f(\theta, n) - l_f(\theta)| = \sup_{\theta \in G^*} \left| \sum (p_i(\theta_0) - \eta) \log \frac{p_i(\theta_0)}{p_i(\theta)} - \sum (p_i^0(\theta_0) - \eta) \log \frac{p_i^0(\theta_0)}{p_i^0(\theta)} \right| \]

\[ \leq \left| \sum (p_i(\theta_0) - \eta) \log p_i(\theta_0) - \sum (p_i^0(\theta_0) - \eta) \log p_i^0(\theta_0) \right| \quad (3.15) \]

\(^9\)And no bin probabilities are 0 because of the support requirement.

\(^{10}\)Since \( k \) is fixed.
\[ + \sup_{\theta \in G^*} \left| \sum (p_i(\theta_0) - \eta) \log p_i(\theta) - \sum (p_i^\eta(\theta_0) - \eta) \log p_i^\eta(\theta) \right| \] (3.16)

Where (3.15) \( \to 0 \) by (3.14), and (3.16) \( \to 0 \) by (3.13).

So by Lemma 3.5, there is an \( n_2 \) such that

\[ |r_f(\theta, n) - l_f(\theta)| \leq \frac{2}{3} \epsilon_2 , \quad \forall \, n \geq n_2 , \quad \forall \, \theta \in G^* . \] (3.17)

So by (3.6) and (3.17) it follows that

\[ 0 < \frac{1}{3} \epsilon_2 \leq r_f(\theta, n) , \quad \forall \, n \geq n_2 , \quad \forall \, \theta \in G^* . \] (3.18)

To see that (3.18) follows as claimed, we use a quick proof by contradiction. Assume (3.18) is false, but that (3.6) and (3.17) hold. Then

\[ \frac{1}{3} \epsilon_2 > r_f(\theta, n) \quad \text{some } n \geq n_2 , \quad \text{some } \theta \in G^* \]

\[ \implies r_f(\theta, n) - |l_f(\theta)| < \frac{1}{3} \epsilon_2 - |l_f(\theta)| = \frac{1}{3} \epsilon_2 - l_f(\theta) \leq -\frac{2}{3} \epsilon \] (3.19)

where the last inequality follows by (3.6). Then (3.19) is

\[ \iff r_f(\theta, n) - l_f(\theta) < -\frac{2}{3} \epsilon_2 , \quad \text{some } n \geq n_2 , \quad \text{some } \theta \in G^* . \] (3.20)

But this contradicts (3.17). Hence (3.18) must follow as claimed.
Continuing with the rest of the proof, from (3.18) and (3.9),

\[ 0 < \frac{1}{3} \epsilon_2 \leq \sum \frac{N_i}{n} \log \frac{p_i(\theta_0)}{p_i(\theta)}, \quad \forall \ n \geq n_3 := \max\{n_1, n_2\}, \quad \forall \ \theta \in G^* \]  

(3.21)

Since \( \sum \frac{N_i}{n} \log p_i(\theta) \) is continuous it must attain a local maximum on \( G \), call it \( \tilde{\theta} \) (see Figure 3.1). By assumption \( \beta, \sum \frac{N_i}{n} \frac{\partial}{\partial \theta_l} \log p_i(\theta), \ l = 1, \ldots, s, \) exist and necessarily must be 0 at \( \tilde{\theta} \). But \( \delta \) was arbitrary. Thus \( \tilde{\theta} \) is consistent for \( \theta_0 \). ☐ ☐ ☐

It is natural at this point to present another similar theorem, due to the fact that its proof is largely the same as that of the previous theorem. The previous theorem required the existence of partial derivatives of \( p_i(\theta) \). The next theorem relaxes the need for derivatives to only continuity, but adds the requirement that one must restrict
attention to some compact region $\Gamma \in \Theta$ where it is believed that $\theta_0$ must surely lie. Such a restriction is clearly acceptable from an applied standpoint.

**Theorem 3.2** Let $\{P_\theta : \theta \in \Theta\}$ be a family of probability measures on $S$, with sigma-finite dominating measure $\mu$, where the support of the family does not depend upon $\theta$. Suppose $X_1, X_2, \ldots$ is a sequence of iid observations with common distribution $P_{\theta_0}$ (where $P_{\theta_0} \in \{P_\theta : \theta \in \Theta\}$). Denote by $\mathcal{S}$ a $\sigma$-algebra of $\mu$-measurable subsets of $S$ (e.g. $S = \mathbb{R}^d$, $S = \mathbb{B}^d$, $\mu = \text{Lebesgue measure on } \mathbb{R}^d$). Let $\mathcal{C}$ be a subclass of $\mathcal{S}$ which is permissible and has polynomial discrimination. Denote by $I_i$ the $i$th bin of a partition of $S$ into $k$ sets from $\mathcal{C}$. Let $p_i(\theta) = P_\theta(I_i)$ and $p_i^\circ(\theta) = P_\theta(I_i^\circ)$.\(^{11}\) Suppose that the following conditions are satisfied

1. Each $I_i$ may be random so long as it is consistent for some $I_i^\circ$ (where $I_i^\circ$ denotes the region of $S$ one attempts to estimate with $I_i$) in the sense that

$$\mu(I_i \bigtriangleup I_i^\circ) \longrightarrow 0, \quad i = 1, \ldots, k \quad \text{a.s. } P_{\theta_0}$$

as $n \to \infty$, where $\bigtriangleup$ is the symmetric difference operator, i.e. $A \bigtriangleup B = (A \cap B^c) \cup (B \cap A^c)$ for $\mu$-measurable sets $A$ and $B$,

2. $p_i^\circ(\theta) \neq p_i^\circ(\theta_0)$ for some $i$ whenever $\theta \neq \theta_0$ (this is the weak identifiability condition of Rao), and

\(^{11}\) $I_i^\circ$ denotes the region of $S$ one attempts to estimate with $I_i$. 

3. \( p_i(\theta), i = 1, \ldots, k, \) are continuous functions of \( \theta, \)

then if \( \Gamma \) is any compact region of \( \Theta \) such that \( \theta_0 \in \Gamma \subset \Theta, \) then the global maximizer of

\[
\sum_i \frac{N_i}{n} \log p_i(\theta)
\]

over \( \Gamma \) is strongly consistent. Furthermore, this global maximizer must exist for large enough samples. Here \( N_i \) denotes the number of observations which fall in \( I_i. \)

**Proof.** As mentioned, the proof here is similar to that of Theorem 3.1, very little changes. Let \( G^+ := \{ \theta : \| \theta - \theta_0 \| \geq \delta \} \cap \Gamma. \) \( G^+ \) is compact. The weak identifiability condition, compactness of \( G^+ \), the information inequality,\(^3\) and the continuity of the \( p_i^0(\theta) \)'s ensures that

\[
\inf_{G^+} \sum_i p_i^0(\theta_0) \log \frac{p_i^0(\theta_0)}{p_i^0(\theta)} \geq \epsilon_1 > 0
\]

which replaces (3.5). The rest of the arguments go through without change replacing \( G^* \) with \( G^+ \) as needed. The reader may observe that the essential property of \( G^* \) used in the preceding proof was compactness. The only new twists come at the end. Specifically (3.21) becomes

\(^{3}\)The “information inequality” referred to is that

\[
\sum_i p_i^0(\theta_0) \log \frac{p_i^0(\theta_0)}{p_i^0(\theta)} \geq 0
\]

with equality if and only if \( p_i^0(\theta_0) = p_i^0(\theta) \) for all \( i. \) The weak identifiability assumption ensures that the inequality is strict for any \( \theta \) chosen from \( G^+. \) Then continuity of the \( p_i^0(\theta) \)'s and the compactness of \( G^+ \) ensure that a smallest value is attained.
\[ 0 < \frac{1}{3} \epsilon_2 \leq \sum_i \frac{N_i}{n} \log \frac{p_i(\theta_0)}{p_i(\theta)} , \quad \forall n \geq n_3 := \max\{n_1, n_2\} , \quad \theta \in G^+ \text{ a.s.} \quad (3.22) \]

Let \( m_t(\theta) := \sum_i \frac{N_i}{n} \log p_i(\theta) \). By (3.22), \( m_t(\theta_0) \) is at least \( \frac{1}{3} \epsilon_2 \) larger than \( m_t(\theta) \) for all \( \theta \in G^+ \). As in the proof of Theorem 3.1, \( m_t(\theta) \) must attain a local maximum on \( G \), call it \( \tilde{\theta} \). If \( \tilde{\theta} \) is the only local maximum in \( G \), then it is the global maximizer of \( m_t(\theta) \) over \( \Gamma \). If there is more than one local maximum in \( G \), let \( \hat{\theta} \) be the "largest one" (i.e. \( m_t(\hat{\theta}) \) is larger than \( m_t(\tilde{\theta}) \) for the other \( \tilde{\theta} \)'s in \( G \)). Then \( \hat{\theta} \) is the global maximum of \( m_t(\theta) \) over \( \Gamma \). In any case, the global maximizer of \( m_t(\theta) \) over \( \Gamma \) must exist a.s. for any \( n \geq n_3 \). Additionally, it must fall within a distance \( \delta \) of \( \theta_0 \). But \( \delta \) was arbitrary, hence \( \hat{\theta} \) is consistent. 

Since the conclusion of Theorem 3.2 is more appealing from an applied standpoint, it seems natural to prove a corresponding version of Theorem 3.2 for the SIMEST situation next.\(^{13}\) In order to do this, we first need to discuss a (mathematically) formal way of thinking about our simulation procedure.

We will assume throughout that our simulation algorithm can be appropriately formalized as follows. Let \( U \) be a (simulated) random variable taking values in some

---

\(^{13}\) As a matter of fact, there is no satisfactory SIMEST analog to Theorem 3.1, hence we will not revisit it again. The difficulty with Theorem 3.1 is that there is no reason to expect that the SIMEST surface

\[ m_t^*(\theta) := \sum \frac{N_i}{n} \log p_i^*(\theta) \]

\( (p_i^*(\theta) \) are the SIMEST estimates of \( p_i(\theta) \)) will be continuous, much less differentiable.
set \( \mathcal{U} \). Let \( S_\theta : \mathcal{U} \to S \subseteq \mathbb{R}^d \) be a mapping from \( \mathcal{U} \) to \( S \) and suppose that the simulations for the postulated model are constructed by the composition \( S_\theta(U) \). Let \( Q(\cdot) \) denote the probability measure on \( \mathcal{U} \) determined by the simulation mechanism. Suppose that the class of subsets of \( \mathcal{U} \) given by

\[
\mathcal{D}_i = \{ [S_\theta \in I_i] : \theta \in \Theta \}
\]

is permissible and has polynomial discrimination. Under such assumptions, we may apply the generalized Glivenko-Cantelli result (Theorem B.2) to conclude that

\[
\sup_{\mathcal{D}_i} |Q_B D - Q D| \to 0 \quad \text{a.s.} \tag{3.23}
\]

Assuming for the sake of the formal arguments that

\[
Q(S_\theta(U) \in I_i) := Q(\omega_{\text{sim}} : S_\theta(U(\omega_{\text{sim}})) \in I_i) = p_i(\theta)
\]

(which is clearly intuitive), we may rewrite (3.23) as

\[
\sup_{\theta \in \Theta} |p^*_i(\theta) - p_i(\theta)| \to 0 \quad \text{a.s.}
\]

where the convergence occurs as \( B \to \infty \).

**Theorem 3.3** Let \( \{P_\theta : \theta \in \Theta\} \) be a family of probability measures on \( S \), with sigma-finite dominating measure \( \mu \), where the support of the family does not depend upon \( \theta \). Suppose \( X_1, X_2, \ldots \) is a sequence of iid observations with common distribution \( P_{\theta_0} \) (where \( P_{\theta_0} \in \{P_\theta : \theta \in \Theta\} \)). Denote by \( S \) a \( \sigma \)-algebra of \( \mu \)-measurable subsets of \( S \)
(e.g. \( S = \mathbb{R}^d \), \( S = B^d \), \( \mu = \text{Lebesgue measure on} \ \mathbb{R}^d \)). Let \( C \) be a subclass of \( S \) which is permissible and has polynomial discrimination. Denote by \( I_i \) the \( i \)th bin of a partition of \( S \) into \( k \) sets from \( C \). Let \( p_i(\theta) = P_\theta(I_i) \) and \( p_i^*(\theta) = P_\theta(I_i^*) \).\(^{14}\) Suppose that the following conditions are satisfied

1. Each \( I_i \) may be random so long as it is consistent for some \( I_i^* \) (where \( I_i^* \) denotes the region of \( S \) one attempts to estimate with \( I_i \)) in the sense that

\[
\mu(I_i \triangle I_i^*) \to 0, \quad i = 1, \ldots, k \quad \text{a.s.} \ P_{\theta_0}
\]

as \( n \to \infty \), where \( \triangle \) is the symmetric difference operator, i.e. \( A \triangle B = (A \cap B^c) \cup (B \cap A^c) \) for \( \mu \)-measurable sets \( A \) and \( B \),

2. \( p_i^*(\theta) \neq p_i^*(\theta_0) \) for some \( i \) whenever \( \theta \neq \theta_0 \) (this is the weak identifiability condition of Rao),

3. \( p_i(\theta), i = 1, \ldots, k \), are continuous functions of \( \theta \), and

4. Let \( U \) be a (simulated) random variable (\( U \) could be a vector) taking values in some set \( U \). Let \( S_\theta : U \to S \subseteq \mathbb{R}^d \) and suppose that simulations for the model are constructed by the composition \( S_\theta(U) \). Suppose that (for each \( i = 1, \ldots, k \))

the class of subsets of \( U \) given by

\[
\mathcal{D} := \{ [S_\theta \in I_i] : \theta \in \Theta \}
\]

\(^{14}\) \( I_i^* \) denotes the region of \( S \) one attempts to estimate with \( I_i \).
is permissible and has polynomial discrimination,

then if $\Gamma$ is any compact region of $\Theta$ such that $\theta_0 \in \Gamma \subset \Theta$, then the global maximizer of

$$\sum_i \frac{N_i}{n} \log p_i^*(\theta)$$

over $\Gamma$ is strongly consistent. Furthermore, this global maximizer must exist for large enough samples.

Proof. Choose $\delta > 0$. To begin with, we will assume that the sample size $n$ (recall that the sample size is $n$, whereas the number of simulations is $B$) is large enough that the global maximizer over $\Gamma$ of the multinomial likelihood, $m_i(\theta)$, exists and falls within a ball of radius $\delta$ around $\theta_0$. We intend to show that in such a situation, it is possible to simulate enough to find a global maximizer over $\Gamma$ of

$$m_i^s(\theta) := \sum_i \frac{N_i}{n} \log p_i^s(\theta)$$

within $\delta$ of $\theta_0$ as well. Since $\delta$ is arbitrary, this will give us the desired result.

From the assumptions (and the above discussion), we conclude right away that

$$\sup_{\theta \in \Theta} |p_i^s(\theta) - p_i(\theta)| \to 0 \quad \text{a.s.}$$

The rest of our proof will be more clear if we break it up into several lemmas. These proofs are essentially identical (following a change in notation) to the proofs
of Lemmas 3.2, 3.3, and 3.4, but for our own benefit and perhaps for the benefit of the interested reader we reproduce them again, making the notational adjustments.

**Lemma 3.6** If $\Gamma$ is compact, then $\sup_{\theta \in \Theta} |p_i^*(\theta) - p_i(\theta)| \to 0$ a.s. ensures that

$$\sup_{\Gamma} \left| \frac{p_i^*(\theta)}{p_i(\theta)} - 1 \right| \to 0 \quad \text{a.s.}$$

**Proof.** Since our theorem assumes that the support does not depend on $\theta$, it follows that $p_i(\theta)$ must be bounded below on any compact subset $\Gamma$ of $\Theta$ (recall that $I_i$ is now fixed). Thus

$$\sup_{\Gamma} \left| \frac{p_i^*(\theta)}{p_i(\theta)} - 1 \right| \leq \sup_{\Gamma} \frac{1}{p_i(\theta)} \sup_{\Gamma} |p_i^*(\theta) - p_i(\theta)| \leq c_\Gamma \sup_{\Gamma} |p_i^*(\theta) - p_i(\theta)| \to 0 \quad \text{a.s.}$$

where $c_\Gamma$ is a finite constant. $\square$

**Lemma 3.7**

$$\limsup_{\theta} \log \frac{p_i^*(\theta)}{p_i(\theta)} \leq 0 \quad \text{and} \quad \liminf_{\theta} \log \frac{p_i^*(\theta)}{p_i(\theta)} \geq 0 ,$$

where $\inf_{\theta}$ and $\sup_{\theta}$ are taken over $\Gamma$, and $\lim$ and $\lim$ are taken with respect to $B$.

**Proof.** By Lemma 3.6, for each $\epsilon > 0$, $\exists B(\epsilon)$ such that

$$1 - \epsilon \leq \frac{p_i^*(\theta)}{p_i(\theta)} \leq 1 + \epsilon , \quad \forall \theta \in \Gamma, B \geq B(\epsilon) .$$

Pick $\eta$. Choose $\epsilon_\eta$ such that $-\eta \leq \log(1 - \epsilon_\eta)$. Choose $B(\epsilon_\eta)$ such that
\[
1 - \epsilon_\eta \leq \frac{p_i^*(\theta)}{p_i(\theta)}, \quad \forall \theta \in \Gamma, \quad B \geq B(\epsilon_\eta)
\]

\[
\implies -\eta \leq \log \frac{p_i^*(\theta)}{p_i(\theta)}, \quad \forall \theta \in \Gamma, \quad B \geq B(\epsilon_\eta)
\]

\[
\implies -\eta \leq \inf_{\theta} \log \frac{p_i^*(\theta)}{p_i(\theta)}, \quad B \geq B(\epsilon_\eta)
\]

\[
\implies -\eta \leq \liminf_{\theta} \log \frac{p_i^*(\theta)}{p_i(\theta)}.
\]

But \( \eta \) was arbitrary, so

\[
\liminf_{\theta} \log \frac{p_i^*(\theta)}{p_i(\theta)} \geq 0.
\]

Similarly,

\[
\frac{p_i^*(\theta)}{p_i(\theta)} \leq 1 + \epsilon_\eta, \quad \forall \theta \in \Gamma, \quad B \geq B(\epsilon_\eta)
\]

\[
\implies \log \frac{p_i^*(\theta)}{p_i(\theta)} \leq \log(1 + \epsilon_\eta) \leq \eta, \quad \forall \theta \in \Gamma, \quad B \geq B(\epsilon_\eta)
\]

\[
\implies \limsup_{\theta} \log \frac{p_i^*(\theta)}{p_i(\theta)} \leq \eta
\]

\( \eta \) arbitrary \( \implies \limsup_{\theta} \log \frac{p_i^*(\theta)}{p_i(\theta)} \leq 0 \).
Lemma 3.8

\[ \limsup_{B \to \infty} \left| \log \frac{p_i^*(\theta)}{p_i(\theta)} \right| = 0 \]

Proof. Observe that

\[ 0 \leq \left| \log \frac{p_i^*(\theta)}{p_i(\theta)} \right| \leq \max \left\{ \log \frac{p_i^*(\theta)}{p_i(\theta)}, -\log \frac{p_i^*(\theta)}{p_i(\theta)} \right\} \]

\[ \Rightarrow \quad 0 \leq \sup_{\theta} \left| \log \frac{p_i^*(\theta)}{p_i(\theta)} \right| \leq \max \left\{ \sup_{\theta} \log \frac{p_i^*(\theta)}{p_i(\theta)}, \sup_{\theta} -\log \frac{p_i^*(\theta)}{p_i(\theta)} \right\} \]

\[ \Rightarrow \quad 0 \leq \limsup_{\theta} \left| \log \frac{p_i^*(\theta)}{p_i(\theta)} \right| \leq \max \left\{ \limsup_{\theta} \log \frac{p_i^*(\theta)}{p_i(\theta)}, \limsup_{\theta} -\log \frac{p_i^*(\theta)}{p_i(\theta)} \right\} \leq 0 \]

by Lemma 3.7. Since

\[ 0 \leq \limsup_{\theta} \left| \log \frac{p_i^*(\theta)}{p_i(\theta)} \right| \leq \limsup_{\theta} \left| \log \frac{p_i^*(\theta)}{p_i(\theta)} \right| \]

we get that

\[ \limsup_{\theta} \left| \log \frac{p_i^*(\theta)}{p_i(\theta)} \right| = 0 \]

as well. In particular, the limit exists. \(\square\)

From Lemma 3.8 it follows right away that

\[ \sup_{\Gamma} \left| \sum_i \frac{N_i}{n} \log p_i^*(\theta) - \sum_i \frac{N_i}{n} \log p_i(\theta) \right| \xrightarrow{B \to \infty} 0 \quad \text{a.s.} \]
or equivalently

$$\sup_{\Gamma} |m_i^s(\theta) - m_i(\theta)| \xrightarrow{B \to \infty} 0 \quad \text{a.s.}$$

Hence we may take $B$ large enough to get $m_i^s(\theta)$ as uniformly close to $m_i(\theta)$ on $\Gamma$ as we like. Then it follows that $m_i^s(\theta)$ must eventually (i.e. for $B$ large enough) exhibit a global maximum (over $\Gamma$) somewhere on $\{\theta : ||\theta - \theta_0|| < \delta\}$.

More formally, recall the choice we made of $\epsilon_2$ in the proof of Theorem 3.2. Choose $B(\epsilon_2)$ so that $m_i^s(\theta)$ is uniformly close to $m_i(\theta)$ on $\Gamma$ to within say $1/9\epsilon_2$. At the same time, make sure $B(\epsilon_2)$ is large enough to have $m_i^s(\theta_0)$ within $1/9\epsilon_2$ of $m_s(\theta_0)$. Then $m_i^s(\theta_0)$ will be at least $1/9\epsilon_2$ larger than $m_i^s(\theta)$ for $\theta \in G^+$ (recall that $G^+ := \{\theta : ||\theta - \theta_0|| \geq \delta\} \cap \Gamma$). From there, the rest of the argument in the proof of Theorem 3.2 is the same.

\[ \square \square \square \]

### 3.5 The Use of Data-Based Bins

At this point, we notice that our direct proof of consistency for SIMEST (in the case of multivariate random bins) is very different from the consistency result(s) in Pakes and Pollard [40, 1989]. The essential (theoretical) difference is also a practical one. Unlike the Pakes and Pollard paper, we do not restrict ourselves to bins selected \emph{a priori} the data. As noted earlier, for multivariate random variables, the "empty space phenomenon" drives us to use data based binning.
Our focus will be on Theorem 3.1 of Pakes and Pollard [40, 1989]. To distinguish this theorem from our own theorems, we will refer to the Pakes and Pollard Theorem as Theorem \( C \) (\( C \) being used to remind one of "consistency").

Random-bins do not fit 1989 formulation

To demonstrate that Theorem \( C \) will not fit our situation,\(^{15}\) consider for simplicity the random bins version of the estimator first considered on page 1047 of Pakes and Pollard [40, 1989]

\[
G_n^{(1)}(\theta) := \left( \frac{N_1}{n} - p_1(\theta), \ldots, \frac{N_k}{n} - p_k(\theta) \right)^T
\]

where \( N_i \) denotes the number of observations which fall in \( I_i \), the \( i \)th bin. The \( I_i \)'s are random (data-based) and are chosen from the class \( \mathcal{C} \), where \( \mathcal{C} \) is a VC class of subsets of \( \mathbb{R}^d \).\(^{16}\) Consider the estimator\(^{17}\)

\[
\hat{\theta}_n^{(1)} = \arg \min_{\theta \in \Theta} \| G_n^{(1)}(\theta) \| = \arg \min_{\theta \in \Theta} \left( \sum_i \left( \frac{N_i}{n} - p_i(\theta) \right)^2 \right)^{1/2}.
\]

Clearly this estimator satisfies condition (i) of Theorem \( C \). But what about condition (ii)? Since \( \mathcal{C} \) is a VC class, \(^{18}\)

\(^{15}\)For clarity, our situation is simulated multinomial maximum likelihood estimation in the case where the (multivariate) bins are data-dependent.

\(^{16}\)We are ignoring the simulation component of the SIMEST problem in this example.

\(^{17}\)Here as in Pakes and Pollard [40, 1989], \( \| \cdot \| \) denotes the Euclidean norm.

\(^{18}\)In what follows \( \mathcal{G} \) will denote the collection of all partitions of \( \mathbb{R}^d \) into \( \mathcal{C} \) sets. \( \gamma \) will denote a member of the class \( \mathcal{G} \), i.e. a particular partition of \( \mathbb{R}^d \) into \( \mathcal{C} \) sets. This notation follows Pollard [43, 1979].
\[
\sup_{i \in C} \left| \frac{N_i}{n} - p_i(\theta_0) \right| \rightarrow 0 \quad \text{a.s.}
\]

\[\Rightarrow \sup_{\gamma \in \mathcal{C}} \left( \sum_i \left( \frac{N_i}{n} - p_i(\theta_0) \right)^2 \right)^{1/2} \rightarrow 0 \quad \text{a.s.} \iff \sup_{\gamma \in \mathcal{C}} \| g_n^{(1)}(\theta_0) \| \rightarrow 0 \quad \text{a.s.}
\]

\[\iff \sup_{\gamma \in \mathcal{C}} \| g_n^{(1)}(\theta_0) \| = o(1) \quad \text{a.s.} \iff \sup_{\gamma \in \mathcal{C}} g_n^{(1)}(\theta_0) = o(1) \quad \text{a.s.}
\]

where the last equality is interpreted elementwise. This clearly does not fit the framework of Theorem C.

Still further along this line of thought, it is perhaps instructive to reformulate Theorem C so that it does work for the situation just described. This will emphasize the fact that Theorem C as it is given in the 1989 paper is inappropriate for such a random-bin scenario.

**Theorem 3.4 (Theorem C - Pakes and Pollard [40, 1989] (adjusted for random bins))** Under the following conditions \( \hat{\theta}_n \overset{p}{\rightarrow} \theta_0. \)

1. \( \| g_n(\hat{\theta}_n) \| \leq o_p(1) + \inf_{\theta \in \Theta} \| g_n(\theta) \| \)

2. \( \sup_{\gamma \in \mathcal{C}} \| g_n(\theta) \| = o_p(1) \)

---

\(^{19}\)Note that the almost sure version of this theorem is even easier to prove.

\(^{20}\)Note to me: started out here with \( \sup_{\gamma \in \mathcal{C}} g_n(\theta) = o_p(1) \), but don’t think that makes as much sense. Maybe this is not important.
3. $h_n(\delta, G) := \inf_{\gamma \in G} \inf_{\|\theta - \theta_0\| > \delta} \|g_n(\theta)\|$ is bounded away from 0 in probability.\textsuperscript{21}

Proof. Fix $\epsilon, \delta > 0$. Condition 3 means there is an $M > 0$ and an $n^{(1)}_M$ such that

$$P(A^n_1(M)^c) < \epsilon \quad , \quad \forall \ n \geq n^{(1)}_M$$

(3.24)

where $A^n_1(M) := [h_n(\delta, G) \geq M]$. Note that (3.24) says that for large enough $n$, there is small probability that $h_n(\delta, G)$ will be smaller than $M$. As in Pakes and Pollard [40, 1989],

$$\|g_n(\hat{\theta}_n)\| \leq o_p(1) + \|g_n(\theta_0)\| \leq o_p(1) + \sup_{\gamma \in G} \|g_n(\theta_0)\| = o_p(1)$$

where the last equality follows from Condition 2. Thus $P(\|g_n(\hat{\theta}_n)\| < M) \to 1$.\textsuperscript{22} So for large enough $n$, we get

$$P\left(\|g_n(\hat{\theta}_n)\| < M \leq \inf_{\gamma \in G} \inf_{\|\theta - \theta_0\| > \delta} \|g_n(\theta)\|\right) \geq 1 - 2\epsilon$$

(3.25)

and so (since the event in (3.25) implies the event in (3.26)) we get that for $n$ large enough

$$P\left(\hat{\theta}_n \in \{\theta : \|\theta - \theta_0\| \leq \delta\}\right) \geq 1 - 2\epsilon$$

(3.26)

\textsuperscript{21}That is, for large enough $n$, there is high probability that $h_n(\delta, G)$ will be bounded away from 0.

\textsuperscript{22}Let $A^n_2(M) := [\|g_n(\hat{\theta}_n)\| < M]$. By properties of probabilities,

$$P[A^n_2(M) \cap A^n_1(M)] = P[A^n_2(M) \cap (A^n_1(M)^c)^c] \geq P(A^n_2(M)) - P(A^n_1(M)^c) \geq 1 - 2\epsilon$$.
\[ \Leftrightarrow \quad P\left(\|\hat{\theta}_n - \theta_0\| \leq \delta \right) \geq 1 - 2\epsilon \]

The purpose of reformulating this theorem is merely to show that the random bins setup cannot be subsumed into Theorem C of the Pakes and Pollard [40, 1989] paper.

For example, although \( \sup_{\gamma \in G}(\cdot) \) is often abbreviated to \( \| \cdot \| \) in Pollard [44, 1984], if we used a similar abbreviation, say \( \inf_{\gamma \in G}(\cdot) = \| \cdot \|_G \), at Condition 3 we would still require two norms, e.g.

\[ \left\| \inf_{\|\theta - \theta_0\|_G > \delta} \|g_n^{(1)}(\theta)\|_e \right\|_G \]

where \( \| \cdot \|_e \) represents the Euclidean norm in \( \mathbb{R}^d \).\(^{23}\) Even further, we have not even added on the simulation component here; Theorem 3.4 could become no less complicated. At any rate, this is enough to demonstrate that adding the random bins component to this problem does not fall under the ambit of the Pakes and Pollard [40, 1989] paper.

---

\(^{23}\) Of course \( \inf_{\gamma \in G} \) is never denoted by \( \| \cdot \| \) in Pollard [44, 1984], but we could turn Condition 3 around so that there was a sup instead of an inf. We leave it as is, the point is clear.
Chapter 4

Strong Consistency for the Cressie-Read Class

In this short chapter we state and prove Theorems 4.1 and 4.2. These Theorems extend Theorems 3.2 and Theorem 3.3 (respectively) to the case where the criterion function may be any member of the Cressie-Read class (4.1). In Section 4.1 we present the Cressie-Read class of criterions, discuss some notation, and state Theorem 4.1. In Section 4.2 we prove Theorem 4.1. In Section 4.3 we state and prove Theorem 4.2. This chapter is largely technical and has been isolated only for the sake of clarity.

4.1 A Strong Consistency Theorem for the CR Class

While our consistency results for the multinomial likelihood criterion are an important step, we would like to extend them further. In hindsight, we observe that the critical properties used in the proofs of Theorems 3.1, 3.2, and 3.3 were

$$\sup_{\theta} \left| \frac{N_i}{n} - p_i(\theta_0) \right| \to 0 \quad \text{a.s.}$$

and

$$\sup_{\theta} \left| p_i(\theta) - p_i^0(\theta) \right| \to 0 \quad \text{a.s.}$$

along with the weak identifiability assumption and the continuity of $p_i(\theta)$. It is of immediate interest to free ourselves from being tied to the multinomial likelihood
criterion. We now do exactly this by establishing (in Theorem 4.2) strong consistency for the Cressie-Read class (CR class) given by

\[ C(\lambda, \theta) := \frac{2}{\lambda(\lambda + 1)} \sum_i N_i \left( \left( \frac{N_i/n}{p_i(\theta)} \right)^\lambda - 1 \right) \]  

(4.1)

where \( \lambda \) is some fixed number (generally Read and Cressie take \( \lambda > -1 \)). Define

\[ CR_n := \sum_i \frac{N_i}{n} \left( \left( \frac{N_i/n}{p_i(\theta)} \right)^\lambda - 1 \right) \]

so that \( 2 (\lambda(\lambda + 1))^{-1} nCR_n = C(\lambda, \theta) \). To get our bearings in the CR world, observe that we will take our parameter estimates as

\[ \hat{\theta}_n = \arg \min_{\theta} \frac{2}{\lambda(\lambda + 1)} \sum_i N_i \left( \left( \frac{N_i/n}{p_i(\theta)} \right)^\lambda - 1 \right) \]

\[ = \arg \min_{\theta} \sum_i \frac{N_i}{n} \left( \left( \frac{N_i/n}{p_i(\theta)} \right)^\lambda - 1 \right) \]

(4.2)

and that \( C(1, \theta) \) corresponds to the Pearson criterion function. Also, \( \lim_{\lambda \to 0} C(\lambda, \theta) \)
corresponds to the likelihood ratio (LR) criterion; however, it seems more rigorous to have a separate theorem for the \( \log(\cdot) \) case (as we have in Chapter 3).

It is the purpose of this chapter to show that for any fixed choice of \( \lambda \), \( \hat{\theta}_n \) defined in (4.2) is strongly consistent for \( \theta_0 \). As in Chapter 2, what makes our results new is the fact that we treat the case where simultaneously one uses (i) a simulation-based criterion function and (ii) a heavily data-based (and multivariate) binning strategy. Define
\[ CR^c_\pm(\eta) := \sum_i (p^c_i(\theta_0) \pm \eta) \left( \left( \frac{p^c_i(\theta_0) \pm \eta}{p^c_i(\theta)} \right)^\lambda - 1 \right) \]

\[ CR_\pm(\eta) := \sum_i (p_i(\theta_0) \pm \eta) \left( \left( \frac{p_i(\theta_0) \pm \eta}{p_i(\theta)} \right)^\lambda - 1 \right) . \]

For clarity, by the ± notation used, it is meant for example that

\[ CR^-_c(\eta) = \sum_i (p^c_i(\theta_0) - \eta) \left( \left( \frac{p^c_i(\theta_0) - \eta}{p^c_i(\theta)} \right)^\lambda - 1 \right) , \]

and similarly for \( CR^+_c(\eta) \). As before, let \( G = \{ \theta : ||\theta - \theta_0|| \leq \delta \} \), \( G^+ = G \cap \Gamma \) where \( \Gamma \) is compact. Consider the following theorem for the case where we have the criterion function available in closed form (this parallels Theorem 3.2).

**Theorem 4.1** Let \( \{ P_\theta : \theta \in \Theta \} \) be a family of probability measures on \( S \), with sigma-finite dominating measure \( \mu \), where the support of the family does not depend upon \( \theta \). Suppose \( X_1, X_2, \ldots \) is a sequence of iid observations with common distribution \( P_{\theta_0} \) (where \( P_{\theta_0} \in \{ P_\theta : \theta \in \Theta \} \)). Denote by \( S \) a \( \sigma \)-algebra of \( \mu \)-measurable subsets of \( S \) (e.g. \( S = \mathbb{R}^d \), \( S = \mathcal{B}^d \), \( \mu = \) Lebesgue measure on \( \mathbb{R}^d \)). Let \( C \) be a subclass of \( S \) which is permissible and has polynomial discrimination. Denote by \( I_i \) the ith bin of a partition of \( S \) into \( k \) sets from \( C \). Let \( p_i(\theta) = P_\theta(I_i) \) and \( p^c_i(\theta) = P_\theta(I^c_i) \).

\(^1I^c_i\) denotes the region of \( S \) one attempts to estimate with \( I_i \).
1. Each $I_i$ may be random so long as it is consistent for some $I_i^0$ (where $I_i^0$ denotes the region of $S$ one attempts to estimate with $I_i$) in the sense that

$$\mu(I_i \triangle I_i^0) \rightarrow 0 \quad \text{a.s. } P_{\theta_0}, \quad i = 1, \ldots, k$$

as $n \to \infty$, where $\triangle$ is the symmetric difference operator, i.e. $A \triangle B = (A \cap B^c) \cup (B \cap A^c)$ for $\mu$-measurable sets $A$ and $B$,

2. $p_i^0(\theta) \neq p_i^0(\theta_0)$ for some $i$ whenever $\theta \neq \theta_0$ (this is the weak identifiability condition of Rao), and

3. $p_i(\theta), i = 1, \ldots, k$, are continuous functions of $\theta$,

then if $\Gamma$ is any compact region of $\Theta$ such that $\theta_0 \in \Gamma \subset \Theta$, then the global minimizer of

$$\sum_i \frac{N_i}{n} \left( \frac{N_i/n}{p_i(\theta)} \right)^\lambda - 1$$

over $\Gamma$ is strongly consistent. Furthermore, this global minimizer must exist for large enough samples.

4.2 Proof of Theorem 4.1

We state the following lemmas which will greatly simplify the proof of Theorem 4.1, deferring their proofs until later.
Lemma 4.1 $\sup_G CR_n|_{\theta=\theta_0} \to 0$ a.s. where $\gamma$ denotes a partition of $S$ into $k$ sets from $C$ and $G$ is the collection of all such partitions.

Lemma 4.2 For each choice of $\eta$, there exists $n_0$ such that

[Case 1] $(\lambda > 0) : CR_-(\eta) \leq CR_n \leq CR_+(\eta), \forall n \geq n_0, \forall \theta \in \Theta, \forall \gamma \in G$

[Case 2] $(\lambda \in (-1,0)) : CR_-(\eta) \geq CR_n \geq CR_+(\eta), \forall n \geq n_0, \forall \theta \in \Theta, \forall \gamma \in G$

where $\gamma$ denotes a partition of $S$ into $k$ sets from $C$ and $G$ is the collection of all such partitions.

Lemma 4.3 Uniformly on $G^+$, $|CR_-(\eta) - CR_+^0(\eta)| \to 0$

Lemma 4.4 $\forall \epsilon, \exists \eta$ small enough that

$$\sup_{G^+} |CR_-(\eta) - CR^0(\eta)| \leq \epsilon$$

or equivalently one may conclude that

$$|CR_-(\eta) - CR^0(\eta)| \leq \epsilon \text{ on } G^+.$$

Proof of Theorem 4.1 ($\lambda > 0$ case). \footnote{The proof for the case $\lambda \in (-1,0)$ is very similar, but we omit it for brevity.} Pick $\delta > 0$. By the weak identifiability assumption, the compactness of $G^+$, continuity of the $p_\epsilon^0(\theta)$'s, and an analogue to the information inequality\footnote{For the Cressie-Read class,} there exists $\epsilon$ such that
\[
\inf_{G^+} CR^\circ(0) \geq \varepsilon \quad . 
\] (4.3)

By Lemma 4.4, we may choose an \( \eta_1 \) small enough that

\[
|CR^\circ_-(\eta_1) - CR^\circ(0)| \leq \varepsilon/2 \quad \text{on } G^+ 
\]

\[\implies CR^\circ_-(\eta_1) \geq \varepsilon/2 \quad \text{on } G^+ \quad . \]

By Lemma 4.2, there exists \( n_0 \) such that

\[
CR_-(\eta) \leq CR_n \leq CR_+(\eta) \quad , \quad \forall n \geq n_0 \ , \ \forall \theta \in \Theta \ , \ \forall \gamma \in \mathcal{G} \quad . 
\]

By Lemma 4.3 \( |CR_-(\eta_1) - CR^\circ_-(\eta_1)| \to 0 \) uniformly on \( G^+ \), so choose \( n_1 \) such that

\[
|CR_-(\eta_1) - CR^\circ_-(\eta_1)| \leq \varepsilon/4 \quad \text{on } G^+ \ , \ \forall n \geq n_1 
\]

\[\implies \varepsilon/4 \leq CR^\circ_-(\eta_1) - \varepsilon/4 \leq CR_-(\eta_1) \leq CR_n \quad , \quad \forall n \geq n_2 \ , \ \theta \in G^+ \ , \ \gamma \in \mathcal{G} \]

\[\implies CR_n \geq \varepsilon/4 > 0 \quad , \quad \forall n \geq n_2 \ , \ \theta \in G^+ \ , \ \gamma \in \mathcal{G} \]

where \( n_2 := \max\{n_0, n_1\} \).

\[
\sum_{i} p^\circ_i(\theta_0) \left( \left( \frac{p^\circ_i(\theta_0)}{p^\circ_i(\theta)} \right)^{\lambda} - 1 \right) \geq 0 \quad \iff \quad CR^\circ(0) \geq 0
\]

with equality if and only if \( p^\circ_i(\theta_0) = p^\circ_i(\theta) \) for all \( i \). This result is mentioned in [47, 1988, p. 2].
By Lemma 4.1, there exists $n_3$ such that $|\frac{\theta - \theta_0}{\theta_0} < \varepsilon/8$, $\forall n \geq n_3$, hence $CR_n$ must attain a global minimum on $\Gamma$ somewhere on $G$ ($\forall n \geq n_4 := \max\{n_2, n_3\}$) because $CR_n$ is continuous in $\theta$. So $\hat{\theta}_n \in G, \forall n \geq n_4$. But $\delta$ was arbitrary.

\[
\sup_{c} \left| \left( \frac{N_i}{n} \right)^{\lambda} - 1 \right| \to 0 \quad \text{a.s.}
\]

Proof of Lemma 4.1. First observe that

Why? Since $|p_i(\theta_0) - p_i^\theta(\theta_0)| \to 0$ a.s. it follows that $|p_i(\theta_0)^{\lambda} - p_i^\theta(\theta_0)^{\lambda}| \to 0$ a.s. because $g(x) = x^{\lambda}$ is a continuous function. Choose $\varepsilon$ small enough that $p_i^\theta(\theta_0)^{\lambda} - \varepsilon > 0$. Then there is $n_\varepsilon$ such that

\[
p_i^\theta(\theta_0)^{\lambda} - \varepsilon \leq p_i(\theta_0)^{\lambda} \leq p_i^\theta(\theta_0)^{\lambda} + \varepsilon, \quad \forall n \geq n_\varepsilon
\]

from which it follows

\[
\frac{1}{(p_i^\theta(\theta_0)^{\lambda} + \varepsilon)} \left| \left( \frac{N_i}{n} - p_i(\theta_0) \right)^{\lambda} \right| \leq \frac{1}{p_i^\theta(\theta_0)^{\lambda}} \left| \left( \frac{N_i}{n} \right)^{\lambda} - p_i(\theta_0)^{\lambda} \right|
\]

\[
\leq \frac{1}{(p_i^\theta(\theta_0)^{\lambda} - \varepsilon)} \left| \left( \frac{N_i}{n} \right)^{\lambda} - p_i(\theta_0)^{\lambda} \right|
\]

which ensures that the middle term goes to 0 a.s. since $(p_i^\theta(\theta_0)^{\lambda} \pm \varepsilon)^{-1}$ are both constant and

\[
\sup_{N_i \in \mathbb{C}} \left| \frac{N_i}{n} - p_i(\theta_0) \right| \to 0 \quad \text{a.s.}
\]

so that
\[ \sup_{i \in c} \left| \left( \frac{N_i}{n} \right)^\lambda - p_i(\theta_0) \right| \to 0 \quad \text{a.s.} \]

as well (again by continuity of \( g(x) = x^\lambda \)). So

\[ \sup_{i \in c} \left| \left( \frac{N_i/n}{p_i(\theta_0)} \right)^\lambda - 1 \right| \to 0 \quad \text{a.s.} \]

From (4.5) it follows that \( \sup_{i \in c} N_i/n \) is eventually bounded, hence

\[ \sup_{i \in c} \frac{N_i}{n} \left| \left( \frac{N_i/n}{p_i(\theta_0)} \right)^\lambda - 1 \right| \to 0 \quad \text{a.s.} \]

and so \( \sup_{\gamma \in \mathcal{G}} CR_n \mid_{\theta = \theta_0} \to 0 \text{ a.s.} \) where \( \gamma \) denotes a partition of \( S \) into \( k \) sets from \( \mathcal{C} \) and \( \mathcal{G} \) is the collection of all such partitions. \( \square \)

**Proof of Lemma 4.2 (Case of \( \lambda > 0 \)).** As before, by the conditions on \( \mathcal{C} \) we may apply Theorem B.2 (generalized Glivenko-Cantelli) to obtain for any choice of \( \eta \)

\[ p_i(\theta_0) - \eta \leq \frac{N_i}{n} \leq p_i(\theta_0) + \eta \quad \forall n \geq n_0 \quad \forall \gamma \in \mathcal{G} \quad . \]  \( (4.6) \)

As a notational convenience, let \( A := \{(n, \theta, \gamma) : n \geq n_0, \theta \in \Theta, \gamma \in \mathcal{G}\} \). For small enough \( \eta \) (4.6) implies that

\[ \frac{N_i}{n} \left( \left( \frac{p_i(\theta_0) - \eta}{p_i(\theta)} \right)^\lambda - 1 \right) \leq \frac{N_i}{n} \left( \left( \frac{N_i/n}{p_i(\theta)} \right)^\lambda - 1 \right) \leq \frac{N_i}{n} \left( \left( \frac{p_i(\theta_0) + \eta}{p_i(\theta)} \right)^\lambda - 1 \right) \quad \text{on} \ A \]

and applying (4.6) again, we get

---

4The proof for the case \( \lambda \in (-1,0) \) is the same, the inequalities merely reverse by the choice of \( \lambda \).
\[
(p_i(\theta_0) - \eta) \left( \left( \frac{p_i(\theta_0) - \eta}{p_i(\theta)} \right)^\lambda - 1 \right) \leq \frac{N_i}{n} \left( \left( \frac{N_i/n}{p_i(\theta)} \right)^\lambda - 1 \right)
\]

\[
\leq (p_i(\theta_0) + \eta) \left( \left( \frac{p_i(\theta_0) + \eta}{p_i(\theta)} \right)^\lambda - 1 \right)
\text{ on } A
\]

and the result follows. \qed

Proof of Lemma 4.3. Without loss of generality consider \( CR_-(\eta) \). As before (see Lemma 3.1), the condition \( \mu(I_i \Delta I_i^o) \to 0 \) a.s. \( P_{\theta_0} \) is enough to ensure that

\[
\sup_{G^+} |p_i(\theta) - p_i^o(\theta)| \to 0 \quad . \tag{4.7}
\]

Let

\[
a_- = (p_i(\theta_0) - \eta) \left( \left( \frac{p_i(\theta_0) - \eta}{p_i(\theta)} \right)^\lambda - 1 \right)
\]

and

\[
a_-^o = (p_i^o(\theta_0) - \eta) \left( \left( \frac{p_i^o(\theta_0) - \eta}{p_i^o(\theta)} \right)^\lambda - 1 \right)
\]

and

\[
a'_- = (p_i(\theta_0) - \eta) \left( \left( \frac{p_i^o(\theta_0) - \eta}{p_i^o(\theta)} \right)^\lambda - 1 \right)
\]

Then \( \sup_{G^+} |a_- - a_-^o| \leq \sup_{G^+} |a_- - a'_-| + \sup_{G^+} |a'_- - a_-^o| \), and then

\[
|a_- - a'_-| = |p_i(\theta_0) - \eta| \left| \left( \frac{p_i(\theta_0) - \eta}{p_i(\theta)} \right)^\lambda - \left( \frac{p_i^o(\theta_0) - \eta}{p_i^o(\theta)} \right)^\lambda \right|
\]
\[ = |p_i(\theta_0) - \eta| \frac{1}{p_i^\lambda(\theta)} \left( \left( \frac{p_i^\lambda(\theta)}{p_i(\theta)} \right)^\lambda (p_i(\theta_0) - \eta)^\lambda - (p_i^\lambda(\theta_0) - \eta)^\lambda \right) \]

\[ = |p_i(\theta_0) - \eta| \frac{1}{p_i^\lambda(\theta)} |a_1 a_2 - a_3| \]  \hspace{1cm} (4.8)

where \(a_1, a_2,\) and \(a_3\) are defined as

\[ a_1 := \left( \frac{p_i^\lambda(\theta)}{p_i(\theta)} \right)^\lambda, \quad a_2 := (p_i(\theta_0) - \eta)^\lambda, \quad a_3 := (p_i^\lambda(\theta_0) - \eta)^\lambda. \]

So (4.8) is

\[ \leq \frac{|p_i(\theta_0) - \eta|}{p_i^\lambda(\theta)} (a_1 |a_2 - a_3| + a_3 |a_1 - 1|) . \]

Then

\[ \sup_{G^+} |a_- - a'_-| \leq |p_i(\theta_0) - \eta| \sup_{G^+} \frac{1}{p_i^\lambda(\theta)} \left( \sup_{G^+} a_1 |a_2 - a_3| + \sup_{G^+} a_3 |a_1 - 1| \right) \]

\[ = |p_i(\theta_0) - \eta| \sup_{G^+} \frac{1}{p_i^\lambda(\theta)} \left( |a_2 - a_3| \sup_{G^+} a_1 + a_3 \sup_{G^+} |a_1 - 1| \right) \]

which will go to zero if we can show that \(\sup_{G^+} a_1 < \infty, \sup_{G^+} p_i^\lambda(\theta)^{-\lambda} < \infty,\) and \(\sup_{G^+} |a_1 - 1| \to 0,\) since it is obvious that \(|a_2 - a_3| \to 0.\) Now

\[ \sup_{G^+} |a_1 - 1| = \sup_{G^+} \left| \left( \frac{p_i^\lambda(\theta)}{p_i(\theta)} \right)^\lambda - 1 \right| \leq c_6 \sup_{G^+} |p_i^\lambda(\theta) - p_i(\theta)^\lambda| \to 0 \]  \hspace{1cm} (4.9)

by (4.7). This also ensures that \(\sup_{G^+} a_1 < \infty\) since
\[
\sup_{G^+} a_1 = \sup_{G^+} \left( \frac{p_1^\circ(\theta)}{p_1(\theta)} \right)^\lambda \to 1
\]

by (4.9). Lastly, \(\sup_{G^+} p_i^\circ(\theta)^{-\lambda} < \infty\) because \(p_i^\circ(\theta)\) is continuous in \(\theta\), \(G^+\) is compact, and the support does not depend upon \(\theta\). Thus \(\sup_{G^+} |a_- - a'_-| \to 0\) a.s.

Also
\[
|a'_- - a_-^o| = \left| (p_i(\theta_0) - \eta) \left( \left( \frac{p_i^\circ(\theta_0) - \eta}{p_i^\circ(\theta)} \right)^\lambda - 1 \right) - (p_i^\circ(\theta_0) - \eta) \left( \left( \frac{p_i^\circ(\theta_0) - \eta}{p_i^\circ(\theta)} \right)^\lambda - 1 \right) \right|
\]
\[
= |a_4|^1\cdot|a_5 - a_6|
\]

where \(a_4\), \(a_5\), and \(a_6\) are defined in the obvious fashion.\(^5\)

\[
\implies \sup_{G^+} |a'_- - a_-^o| = |a_5 - a_6| \sup_{G^+} |a_4| \leq |a_5 - a_6| c'_6 \to 0
\]

where \(c'_6 < \infty\) because
\[
\sup_{G^+} |a_4| \leq (p_i^\circ(\theta_0) - \eta)^\lambda \sup_{G^+} \frac{1}{p_i^\circ(\theta)^\lambda} + 1 \leq (p_i^\circ(\theta_0) - \eta)^\lambda c_6 + 1 < \infty
\]

\(\square\)

**Proof of Lemma 4.4.** Let
\[
b_{1i} := \left( \frac{p_i^\circ(\theta_0) - \eta}{p_i(\theta_0)} \right)^\lambda - 1 , \quad b_{2i} := \left( \frac{p_i^\circ(\theta_0)}{p_i(\theta_0)} \right)^\lambda - 1
\]

\(^5\)Maybe this is not so obvious. Here \(a_5 := (p_i(\theta_0) - \eta)\) and \(a_6 := (p_i^\circ(\theta_0) - \eta)\).
\[ b_{3i} := p_i^\circ(\theta_0) - \eta \quad \text{and} \quad b_{4i} := p_i^\circ(\theta_0) \]

Observe that

\[ \sup_{G^+} |CR^\circ(\eta) - CR^\circ(0)| = \sup_{G^+} \left| \sum_i b_{1i}b_{3i} - \sum_i b_{2i}b_{4i} \right| \leq \sum_i \sup_{G^+} |b_{1i}b_{3i} - b_{2i}b_{4i}| \quad (4.10) \]

which goes to 0 if we show that each \( \sup_{G^+} |b_{1i}b_{3i} - b_{2i}b_{4i}| \to 0 \). Dropping the \( i \) subscripts, observe that

\[ \sup_{G^+} |b_1 - b_2| = \sup_{G^+} \frac{1}{p_i^\circ(\theta)^\lambda} |(p_i^\circ(\theta_0) - \eta)^\lambda - p_i^\circ(\theta_0)^\lambda| = |(p_i^\circ(\theta_0) - \eta)^\lambda - p_i^\circ(\theta_0)^\lambda| \sup_{G^+} \frac{1}{p_i^\circ(\theta)^\lambda} \]

where \( |(p_i^\circ(\theta_0) - \eta)^\lambda - p_i^\circ(\theta_0)^\lambda| \) goes to 0 as \( \eta \searrow 0 \) by ordinary limits (calculus) and \( \sup_{G^+} p_i^\circ(\theta)^{-\lambda} \) is finite (this finite constant previously was referred to as \( c_\lambda \)) as before.

Thus

\[ \sup_{G^+} |b_1 - b_2| \to 0 \quad \text{as} \quad \eta \searrow 0 \quad (4.11) \]

Next

\[ \sup_{G^+} |b_1b_3 - b_2b_4| \leq \sup_{G^+} |b_1||b_3 - b_4| + \sup_{G^+} b_4|b_1 - b_2| = \eta \sup_{G^+} |b_1| + b_4 \sup_{G^+} |b_1 - b_2| \]

and the first term in this sum goes to 0 with \( \eta \), since \( \sup_{G^+} |b_1| < \infty \). The second term in this sum goes to zero due to (4.11). The lemma now follows from (4.10). \( \square \)
### 4.3 Simulation Based Theorem for the Cressie-Read Class

Finally, as before we need to connect the strong consistency result for the closed form (i.e. Theorem 4.1) with a strong consistency result for the simulation-based case. Again, we observe that the important property we used previously was that

\[
\sup_{\Gamma} |p_i^e(\theta) - p_i(\theta)| \to 0 \quad \text{a.s.} \quad (4.12)
\]

This uniformity result was guaranteed by the assumption that \( D \), a particular class of sets, was a VC class. Consider the following lemma.

**Lemma 4.5** If (4.12) holds, then

\[
\sup_{\Gamma} \left| c_\lambda \sum_i \frac{N_i}{n} \left( \left( \frac{N_i/n}{p_i^e(\theta)} \right)^{\lambda} - 1 \right) - c_\lambda \sum_i \frac{N_i}{n} \left( \left( \frac{N_i/n}{p_i(\theta)} \right)^{\lambda} - 1 \right) \right| \to 0
\]

\[\Leftrightarrow \sup_{\Gamma} |c_\lambda CR_n^e(\theta) - c_\lambda CR_n(\theta)| \to 0\]

where \( c_\lambda = 2(\lambda(\lambda + 1))^{-1} \).

**Proof.** Choose \( \epsilon \) small enough that \( p_i(\theta)^\lambda - \epsilon > 0 \). By (4.12), there exists \( B_\epsilon \) such that

\[
p_i(\theta)^\lambda - \epsilon \leq p_i^e(\theta)^\lambda \leq p_i(\theta)^\lambda + \epsilon, \quad \forall B \geq B_\epsilon, \forall \theta \in \Gamma
\]

\[
\Rightarrow \frac{1}{(p_i(\theta)^\lambda + \epsilon) p_i(\theta)^\lambda} \leq \frac{1}{p_i^e(\theta)^\lambda p_i(\theta)^\lambda} \leq \frac{1}{(p_i(\theta)^\lambda - \epsilon) p_i(\theta)^\lambda} \quad \forall B \geq B_\epsilon, \forall \theta \in \Gamma
\]
\[
\sup_{\Gamma} \frac{1}{p_i^*(\theta)^\lambda p_i(\theta)^\lambda} \leq \sup_{\Gamma} \frac{1}{(p_i(\theta)^\lambda - \epsilon) p_i(\theta)^\lambda}, \quad \forall B \geq B_\epsilon \tag{4.13}
\]

and the term on the right is finite since \(p_i(\theta)\) is continuous and \(\Gamma\) is compact.

Thus

\[
\sup_{\Gamma} \left| \frac{1}{p_i^*(\theta)^\lambda} - \frac{1}{p_i(\theta)^\lambda} \right| \leq \sup_{\Gamma} \frac{1}{p_i^*(\theta)^\lambda p_i(\theta)^\lambda} \sup_{\Gamma} |p_i^*(\theta) - p_i(\theta)^\lambda| \\
\leq c_{\Gamma} \sup_{\Gamma} |p_i^*(\theta)^\lambda - p_i(\theta)^\lambda| \quad \text{by (4.13).}
\]

\[
\rightarrow 0 \quad \text{by (4.12).}
\]

It is a simple argument from here to show that

\[
\sup_{\Gamma} |c_\lambda CR^*_n(\theta) - c_\lambda CR_n(\theta)| \rightarrow 0
\]

\(\Box\)

Notice here (as before) we are conditioning on a particular realization of the \(I_i\)'s, after which we let \(B \rightarrow \infty\) (recall that \(B\) is the number of simulations). Assuming \(n\) is large enough to ensure that the closed form minimizer falls in \(G\), we will always be able to simulate enough to find a minimizer of \(CR^*_n(\theta)\) in \(G\) as well, provided only that \(D\) be a VC class. We summarize all of this in the following theorem for the simulation-based estimators constructed from the Cressie-Read class.
Theorem 4.2 Let \( \{P_\theta : \theta \in \Theta\} \) be a family of probability measures on \( S \), with sigma-finite dominating measure \( \mu \), where the support of the family does not depend upon \( \theta \). Suppose \( X_1, X_2, \ldots \) is a sequence of iid observations with common distribution \( P_{\theta_0} \) (where \( P_{\theta_0} \in \{P_\theta : \theta \in \Theta\} \)). Denote by \( S \) a \( \sigma \)-algebra of \( \mu \)-measurable subsets of \( S \) (e.g. \( S = \mathbb{R}^d \), \( S = \mathbb{B}^d \), \( \mu = \text{Lebesgue measure on } \mathbb{R}^d \)). Let \( C \) be a subclass of \( S \) which is permissible and has polynomial discrimination. Denote by \( I_i \) the \( i \)th bin of a partition of \( S \) into \( k \) sets from \( C \). Let \( p_i(\theta) = P_\theta(I_i) \) and \( p_i^0(\theta) = P_{\theta_0}(I_i^0) \).\(^6\) Suppose that the following conditions are satisfied

1. Each \( I_i \) may be random so long as it is consistent for some \( I_i^0 \) (where \( I_i^0 \) denotes the region of \( S \) one attempts to estimate with \( I_i \)) in the sense that

\[
\mu(I_i \triangle I_i^0) \to 0, \quad i = 1, \ldots, k \quad \text{a.s. } P_{\theta_0}
\]

as \( n \to \infty \), where \( \triangle \) is the symmetric difference operator, i.e. \( A \triangle B = (A \cap B^c) \cup (B \cap A^c) \) for \( \mu \)-measurable sets \( A \) and \( B \);

2. \( p_i^0(\theta) \neq p_i^0(\theta_0) \) for some \( i \) whenever \( \theta \neq \theta_0 \) (this is the weak identifiability condition of Rao);

3. \( p_i(\theta), i = 1, \ldots, k, \) are continuous functions of \( \theta \), and

4. Let \( U \) be a (simulated) random variable (\( U \) could be a vector) taking values in some set \( \mathcal{U} \). Let \( S_\theta : \mathcal{U} \to S \subseteq \mathbb{R}^d \) and suppose that simulations for the model

\(^6I_i^0 \) denotes the region of \( S \) one attempts to estimate with \( I_i \).
are constructed by the composition $S_\theta(U)$. Suppose that (for each $i = 1, \ldots, k$) the class of subsets of $\mathcal{U}$ given by

$$
\mathcal{D} := \{ [S_\theta \in I_i] : \theta \in \Theta \}
$$

is permissible and has polynomial discrimination,

then if $\Gamma$ is any compact region of $\Theta$ such that $\theta_0 \in \Gamma \subset \Theta$, then the global minimizer of

$$
\sum_i \frac{N_i}{n} \left( \left( \frac{N_i/n}{p_i(\theta)} \right)^\lambda - 1 \right)
$$

over $\Gamma$ is strongly consistent. Furthermore, this global minimizer must exist for large enough samples.
Chapter 5

Some Illustrative Examples

In this chapter, we present two examples. The first example (Section 5.1) is a simplified version of a parameter estimation problem in carcinogen experiments. This example is only a toy, but it illustrates the following important ideas:

1. There is a genuine need in applications for the flexibility of funny densities (e.g. continuous along one dimension and discrete along another). Hence although our strong consistency theorems (of Chapters 2 and 3) may appear overly formal, the formality is justified by applications.

2. We take advantage of the simplicity of the postulated model to demonstrate how one verifies that the appropriate class(es) induced by the simulation algorithm is (are) polynomial. The detail given here is mainly for the benefit of the reader. Such detail will be omitted in subsequent examples.

The second example (Section 5.3) is the cancer model of Atkinson, Bartoszyński, Brown and Thompson [1, 1983]. This example is our first example of an important application of SIMEST. It requires only a matter of hours to verify that the assumptions needed for strong consistency of the SIMEST parameter estimates are satisfied. In particular, the VC class ideas are well-suited to axiomatic model specification. If
the number of non-linearities and comparisons involved in the simulation algorithm is not too complex, we are nearly assured of success.

In the cancer model, the simulation algorithm SM requires a rejection step. This creates some difficulties for the VC class ideas. In Section 5.2, we suggest one way to overcome these difficulties. There are important practical implications of the VC class ideas (and the ideas in Section 5.2) to SIMEST.

5.1 Carcinogen Example

In this example, we will illustrate how to verify the assumptions of our strong consistency theorems. In particular, we will motivate the need for the flexibility of densities that are not necessarily continuous along each dimension. Also, we will show how to verify that a particular class of sets is polynomial, where the class to be considered is related to the simulation mechanism.

In order to determine whether or not a given chemical is a carcinogen, experiments are performed on animals. The usual technique is to have three dose groups and one control group. The animals are given doses over the course of their lifetimes. The animals will die either due to a tumor or due to some competing cause. A hypothetical set of observations from the experiment (for a single group) is illustrated in Figure 5.1.
It is desired to compare the tumor occurrence rates of the dose groups to the tumor occurrence rate of the control group.\footnote{We do not proceed with this example far enough to compare the tumor rates between groups. We will just consider parameter estimation for a single group.} As a simplifying assumption, suppose that the tumors are rapidly fatal so that the time of death due to tumor roughly corresponds to the time of tumor occurrence. As another simplifying assumption, suppose that there is no misclassification error in determining the cause of death.\footnote{This is usually not true, in fact.}

The observations will take the form \((y_1, z_1), (y_2, z_2), \ldots, (y_n, z_n)\), where the \((y_j, z_j)\) pairs are realizations of the random variables

\[
Y = \min \{ C, T \} \quad , \quad Z = \begin{cases} 
0 \text{ if tumor} \\
1 \text{ if competing cause}
\end{cases}
\]

Here \(C = \text{time of death due to competing cause}\) and \(T = \text{time of death due to tumor}\). In analyzing these types of problems, the assumption is usually made that the competing risk and the tumor risk are independent. Of course this is an unrealistic assumption. We would like to focus our attention on one plausible method of relaxing this assumption.

Observe that \(Y\) is continuous, while \(Z\) is discrete. Thus the joint distribution will be a little unusual. For the case where \(z \in [0, 1)\) observe that

\[
F(y, z) = P[Y \leq y, Z \leq z] = P[Y \leq y, Z = 0] = P[Y \leq y, T < C]
\]

\[
= P[Y \leq y | T < C] P[T < C] = P[\min\{T, C\} \leq y | T < C] P[T < C]
\]
deaths due to tumor

□ deaths due to competing cause

Figure 5.1

\[ P[T < y]P[T < C] = F_T(y)p_{tc} \]  \hspace{1cm} (5.1)

where

\[ p_{tc} = \int \int_{[t<\infty]} f_{T,C}(t,c) \, dt \, dc . \]

Similarly, for the case where \( z \in [1, \infty) \),

\[ F(y, z) = P[Y \leq y, Z \leq z] = P[Y \leq y, Z = 0] + P[Y \leq y, Z = 1] \]

\[ = F_T(y)p_{tc} + F_C(y)(1 - p_{tc}) \]  \hspace{1cm} (5.2)

This leads us to guess that the joint density of \( Y \) and \( Z \) must take the form

\[ f(y, z) = f_T(y)p_{tc}I_{[0]}(z) + f_C(y)(1 - p_{tc})I_{[1]}(z) , \quad y > 0 , \quad z = 0, 1 \]  \hspace{1cm} (5.3)
It is easy to verify that this is the correct density. Let $\mu$ be Lebesgue measure and $\nu$ be counting measure. Let $\pi$ denote the product measure constructed from $\mu$ and $\nu$. Since both measures are sigma finite and $f(y, z)$ is nonnegative, we apply Fubini's theorem to justify iterating the integrals, i.e.

$$\int f(y, z)\pi(dy, dz) = \int \int f(y, z)\mu(dy)\nu(dz)$$

To see that $f(y, z)$ integrates to 1,

$$\int \int f(y, z)\mu(dy)\nu(dz) = \int \left( I_{[0,1]}(z)p_{tc} \int f_T(y)\mu(dy) + I_{[1,\infty)}(z)(1 - p_{tc}) \int f_C(y)\mu(dy) \right) \nu(dz)$$

$$= \int I_{[0,1]}(z)p_{tc} + I_{[1,\infty)}(z)(1 - p_{tc})\nu(dz) = p_{tc} + (1 - p_{tc}) = 1$$

And to check that this density leads to the distribution $F(y, z)$ given in (5.1) and (5.2) above, consider

$$\int_0^y \int_0^x \left\{ f_T(x)p_{tc}I_{[0,1]}(s) + f_C(x)(1 - p_{tc})I_{[1,\infty)}(s) \right\} \mu(ds)\nu(dx) \quad (5.4)$$

if $z \in [0, 1)$ then (5.4) becomes

$$\int_0^y f_T(x)p_{tc}\mu(dx) = p_{tc}F_T(y),$$

and if $z \in [1, \infty)$ then (5.4) becomes

$$\int_0^y f_T(x)p_{tc} + f_C(x)(1 - p_{tc})\mu(dx) = p_{tc}F_T(y) + (1 - p_{tc})F_C(y)$$
matching (5.1) and (5.2) as desired.

To see how easy the first condition of Theorem 3.3 is to satisfy, let \( S = \mathbb{R}^+ \otimes \{0, 1\} \) and again let \( \pi = \nu \times \mu \) where \( \nu \) is counting measure and \( \mu \) is Lebesgue measure on \( \mathbb{R} \). Let \( S \) be the sigma-algebra generated by measurable rectangles of \( S \), where by measurable rectangles we mean sets of the form \( A \otimes B \) where \( A \in \mathbb{R}^+ \), \( B \in \{0, 1\} \). For any measurable rectangle \( A \otimes B \),

\[
P_\theta(A \otimes B) := P_\theta(Y \in A, Z \in B)
\]

is defined by the model (the particular choice of a model is not important at this point). By the usual arguments, \( P_\theta(\cdot) \) extends uniquely to a complete measure on \( S \).

Consider the class \( C \) of subsets of \( S \) which partition \( \mathbb{R}^+ \otimes \{0, 1\} \) into exactly \( k \) regions (none of which are empty, say). By the usual arguments, \( C \) easily has polynomial discrimination. Let \( I^0_1, I^0_2, \ldots, I^0_k \) be a partition of \( \mathbb{R}^+ \otimes \{0, 1\} \) such that \( P_{\theta_0}(I^0_i) = k^{-1} \). As long as the distributions of \( C \) and \( T \) are continuous this is always exactly possible. Given \( n \) observations, choose \( I_1, \ldots, I_k \) so that the number of observations in \( I_i \) is roughly \( k^{-1} \). It is clearly evident that as \( n \to \infty \) we will find that

\[
\pi(I_i \triangle I^0_i) \to 0 \quad \text{a.s.} \quad P_{\theta_0}.
\]

The partitioning described here is depicted in Figure 5.2.

---

3 The condition referred to here is that \( \pi(I_i \triangle I^0_i) \to 0 \) a.s. \( P_{\theta_0} \). The confusion of notation here (\( \mu \) is being used differently in this example than the \( \mu \) that is used in the first condition of Theorem 3.3) is unfortunate. The \( \pi \) of this example corresponds with the \( \mu \) in the statement of Theorem 3.3. Apologies to the reader.

4 We use \( \otimes \) to denote the Cartesian product here.
For the purposes of illustration, consider the following mechanism for modelling some dependence between \( T \) and \( C \). Suppose that \( T \) is Weibull, i.e.

\[
F(t) = 1 - \exp(-\rho t^\alpha)
\]

and that conditional on \( t \), \( C \) is exponential with mean \( \lambda t \), i.e.

\[
f(c \mid t) = \frac{1}{\lambda t} \exp \left( -\frac{c}{\lambda t} \right) .
\]

This is a rather natural way to introduce positive association between \( C \) and \( T \). Our goal is to estimate \( \alpha, \rho, \) and \( \lambda \), given the observations \((y_1, z_1), \ldots, (y_n, z_n)\). Recalling from (5.3) the joint density of \( Y \) and \( Z \), we observe that we must have \( f(t), f(c) \) and
$p_{tc}$ in order to carry out maximum likelihood estimation. But we observe that the marginal density $f(c)$ is not available since

$$f(c) = \int_0^\infty f(c, t) \, dt = \int_0^\infty f(c \mid t) f(t) \, dt$$

$$= \int_0^\infty \frac{1}{\lambda t} \exp\left(-\frac{c}{\lambda t}\right) \alpha \rho(\rho t)^{\alpha-1} \exp(-\rho t^\alpha) \, dt = \frac{\rho}{\lambda} \int_0^\infty u^{-1/\alpha} e^{-u} \exp\left(-\frac{\rho}{\lambda} cu^{-1/\alpha}\right) \, du$$

where the last equality follows by substitution of $u = (\rho t)^\alpha$. This last integral appears to be an integral we can not evaluate in closed form. We arrive at similar difficulties in attempting to calculate $p_{tc}$. Thus even though we can write down the likelihood, it appears that pointwise evaluations of it may have to be made numerically.

To illustrate the use of SIMEST for this problem, the simulation algorithm for the model specified here is simple:

1. Generate $u_1$ and $u_2$ uniform on $(0, 1)$. Let $u = (u_1, u_2)$.

2. Let $t = \rho^{-1} \left(- \log(1 - u_1)\right)^{1/\alpha} := g_\theta(u)$,

3. Let $c = -\lambda t \log(1 - u_2) = -\lambda \rho^{-1} \left(- \log(1 - u_1)\right)^{1/\alpha} \log(1 - u_2) := h_\theta(u)$,

4. Let $y = \min(c, t)$. Let $z = 0$ if $y = t$, $z = 1$ otherwise.

We consider the regions $\{z = 0\}$ and $\{z = 1\}$ separately, combining the results at the end. For each $\theta$, let

$$U_{0i}(\theta) = \{u : g_\theta(u) \leq h_\theta(u) \text{ and } g_\theta(u) \in I_i\}$$
\[ U_{11}(\theta) = [u : g_\theta(u) \geq h_\theta(u) \text{ and } h_\theta(u) \in I_i] \quad . \]

So for a fixed choice of \( \theta \), \( U_{0i}(\theta) \) is the set of those \( u \in U \otimes U \) for which \( z(u) = 0 \) and \( t(u) \in I_i \).\(^5\) This is equivalent to saying that \( U_{0i}(\theta) \) represents the \( U \otimes U \) set for which \( y(u) \) falls in \( I_i \) and \( y(u) = t(u) \). Now let

\[ D_{0i} = \{ U_{0i}(\theta) : \theta \in \Theta \} \quad , \quad D_{1i} = \{ U_{1i}(\theta) : \theta \in \Theta \} \quad . \]

Considering \( D_{0i} \) as an example, each choice of \( \theta \) determines a partition of \( U \otimes U \) into two sets (the set where \( y(u) = t(u) \) falls in \( I_i \) and the set where \( y(u) = t(u) \) does not fall in \( I_i \)). As \( \theta \) varies over \( \Theta \), \( D_{0i} \) represents the collection of all such partitions. We will be able to apply the Glivenko-Cantelli results provided \( D_{0i} \) is not too flexible. This is what we now show.

We may treat \( U_{0i}(\theta) \) by first treating \( \{ [u : g_\theta(u) \leq h_\theta(u)] : \theta \in \Theta \} \) and then treating \( \{ [u : g_\theta(u) \in I_i] : \theta \in \Theta \} \). We proceed first with the case of \( \{ [u : g_\theta(u) \leq h_\theta(u)] : \theta \in \Theta \} \). Observe that

\[ [g_\theta(u) \leq h_\theta(u)] \quad \Leftrightarrow \quad [0 \leq -\lambda \log(1 - u_2)] \quad . \quad (5.5) \]

While a little unnecessary at this point, we introduce some mapping arguments here since they will speed up what follows. Consider the map

\(^5\)Here \( U \) denotes the set \((0,1)\) and \( \otimes \) denotes the Cartesian product.
\[ T_1 : u \mapsto \log(1 - u_2) \]

from \( \mathcal{U} \otimes \mathcal{U} \) to \( \mathcal{U}' \). Let \( g_\lambda(u') = -\lambda u' \) and \( \mathcal{G} = \{g_\lambda : \lambda > 0\} \). Observe that

\[ \{g \geq 0 : g \in \mathcal{G}\} = \{g_\lambda \geq 0 : \lambda > 0\} = \{[u' : g_\lambda(u') \geq 0] : \lambda > 0\} \]

is a VC class of subsets of \( \mathcal{U}' \) by Result B.1 (i.e. \( \{g_\lambda : \lambda > 0\} \) is a finite dimensional vector space of real functions). By Result B.3 for mappings we get that

\[ \{T_1^{-1}[u' : -\lambda u' \geq 0] : \lambda > 0\} = \{[u : g_\lambda(T_1(u)) \geq 0] : \lambda > 0\} \]

is a VC class of subsets of \( \mathcal{U} \otimes \mathcal{U} \). But observe that

\[ [u : g_\theta(u) \leq h_\theta(u)] \subseteq [u : g_\lambda(T_1(u)) \geq 0] \quad . \tag{5.6} \]

To prove (5.6), suppose \( u \) is in the left hand side of (5.6). Then

\[ g_\lambda(T_1(u)) = g_\lambda(\log(1 - u_2)) = -\lambda \log(1 - u_2) \geq 0 \]

where the inequality follows from (5.5). Thus \( u \) is in the right hand side of (5.6), so (5.6) follows. Since subclasses of VC classes are again VC (cf. Pakes and Pollard [40, 1989]), it follows that

\[ \{[u : g_\theta(u) \leq h_\theta(u)] : \theta \in \Theta\} \]

is VC in \( \mathcal{U} \otimes \mathcal{U} \).
Now for the case of \([g_\theta(u) \in I_i : \theta \in \Theta}\),

\[ [g_\theta(u) \geq a] = [- \log(1 - u_1) \geq (\rho a)^a] \]

so that \(\{[g_\theta(u) \geq a] : \theta \in \Theta\}\) is a VC class (here use the mapping \(T_2 : u \mapsto \log(1 - u_1)\)).

Similarly, \(\{[g_\theta(u) \leq b] : \theta \in \Theta\}\) is a VC class. Finally

\[ D_{bi} = \{[g_\theta(u) \leq h_\theta(u)] \cap [g_\theta(u) \in I_i : \theta \in \Theta\} \]

\[ = \{[g_\theta(u) \leq h_\theta(u)] \cap [g_\theta(u) \geq a] \cap [g_\theta(u) \leq b] : \theta \in \Theta\} \]

is a VC class by Result B.2.

Clearly identical arguments will work for \(D_{i1}\) so long as we can show that

\(\{[h_\theta(u) \geq a] : \theta \in \Theta\}\)

is a VC class. Well

\[ [h_\theta(u) \geq a] = \left[ - \frac{\lambda}{\rho} (- \log(1 - u_1))^{1/\alpha} \log(1 - u_2) \geq a \right] \]

\[ = \left[ \frac{1}{\alpha} \log(- \log(1 - u_1)) + \log(- \log(1 - u_2)) \geq \log \left( \frac{\rho}{\lambda a} \right) \right] \]

so that \(\{[h_\theta(u) \geq a] : \theta \in \Theta\}\) is a VC class; to see this, use the mapping

\(T_3 : u \mapsto (\log(- \log(1 - u_1)), \log(- \log(1 - u_2)))\).

As before then, \(D_{i1}\) is a VC class.
Now let
\[ \mathcal{D}_i = \{ \{u: y(u) \in I_i\} : \theta \in \Theta\} = \mathcal{D}_{0i} \cup \mathcal{D}_{1i} \]
so that
\[ \sup_{\theta} |p_i^*(\theta) - p_i(\theta)| = \sup_{\mathcal{D}_i} |P_B D - PD| \]
\[ \leq \max \left\{ \sup_{\mathcal{D}_{0i}} |P_B D - PD|, \sup_{\mathcal{D}_{1i}} |P_B D - PD| \right\} \]  \hspace{1cm} (5.7)
depending upon the choice of \( I_i \). But both terms in (5.7) go to 0 a.s. since \( \mathcal{D}_{0i} \) and \( \mathcal{D}_{1i} \) are VC classes. This completes our example.

We point out a few issues. First, notice that when we choose a map to apply Result B.3, that map cannot depend upon \( \theta \). Also notice that while VC classes are flexible, they are not arbitrarily flexible. This is somewhat disappointing.

Lastly, observe that the ideas here affect the implementation of SIMEST. For the example we gave here, a simulation size \( B \) requires only \( 2B \) uniform random numbers in order to cover all of \( \Theta \). At every choice of \( \theta \), \( p_i^*(\theta) \) is estimated using the same \( 2B \) random numbers. Generate a single \( u \) and we may calculate \( y_\theta(u) \) for every \( \theta \in \Theta \). This is similar to the construction of the maximum likelihood surface, i.e. \( n \) observations determine the entire likelihood surface; we need not take \( n \) observations for each value of \( \theta \). This is certainly good news from an applied standpoint; we reduce both the simulation noise as well as our computational effort.
5.2 Accommodating Rejection Algorithms in Simulation

Typically, we wish to apply the GC results to show that

$$\lim_{B \to \infty} \sup_{\theta} |p_i^B(\theta) - p_i(\theta)| = 0 \quad \text{a.s.}$$

This is generally not too complicated for simple algorithms. Often, however, simulation algorithms may involve a rejection step. This presents some difficulties. We show how to overcome such difficulties now.

Denote by $Q$ the probability measure governing our simulation mechanism, call this hypothetical simulation algorithm SIM. Algorithm SIM takes a simulated value $u \in U$ and composes it with a function $g_\theta(\cdot)$. For now suppose the simulations $g_\theta(u)$ land in $R$, and suppose under algorithm SIM we will reject $g_\theta(u)$ whenever $g_\theta(u) \in A^c$, where $A \subset R$. Thus $[g_\theta(u) \in A]$ is the acceptance region for algorithm SIM. Generally, this region will depend upon $\theta$. We suggest the following correspondence between the $Q(\cdot)$ world and the $P(\cdot)$ world. Let $\Lambda_i = [g_\theta(u) \in I_i]$ and $\Lambda_A = [g_\theta(u) \in A]$. Then if $Y$ is the random variable we are trying to simulate (on $P(\cdot)$),

$$p_i(\theta) = P[Y \in I_i] = Q[u \in \Lambda_i \mid u \in \Lambda_A] = \frac{Q[u \in \Lambda_i, u \in \Lambda_A]}{Q[u \in \Lambda_A]}$$

That is to say, while $p_i(\theta)$ is an unconditional probability in the $P(\cdot)$ world, it is a conditional probability in the $Q(\cdot)$ world.

Let

$$D_i = \{[u : u \in \Lambda_i(\theta) \cap \Lambda_A(\theta)] : \theta \in \Gamma\} = \{D_i(\theta) : \theta \in \Gamma\}$$
where $D_i(\theta) = \{u : u \in \Lambda_i(\theta) \cap \Lambda_A(\theta)\}$ and $\Gamma$ is compact. Suppose that we are able to show that $D_i$ is a VC class. Let $Q_B$ denote the empirical measure based on $B$ simulations. Then

$$\sup_{D_i} |Q_B D_i - Q D_i| \to 0 \quad \text{a.s.} \quad (5.9)$$

We can show (see below) that (5.9) ensures that

$$\sup_{D} \left| \frac{Q_B D_i}{\sum_i Q_B D_i} - \frac{Q D_i}{\sum_i Q D_i} \right| \to 0 \quad \text{a.s.} \quad (5.10)$$

where $D = \{(D_1(\theta), \ldots, D_k(\theta)) : \theta \in \Gamma\}$. Notice that

$$p_i(\theta) = \frac{Q D_i}{\sum_i Q D_i}, \quad (5.11)$$

so clearly it makes perfect sense to take

$$p_i^*(\theta) = \frac{Q_B D_i}{\sum_i Q_B D_i}. \quad (5.12)$$

If we do so, then (5.10) becomes (5.8). This approach is perfectly logical since $p_i^*(\theta)$ divides the number of simulations which fall in $\Lambda_i(\theta) \cap \Lambda_A(\theta)$ by the number of simulations which fall in $\Lambda_A(\theta)$.

We now demonstrate the claim made that (5.9) implies (5.10).

**Result 5.1** If the acceptance probability, $Q(\Lambda_A(\theta))$, is strictly bounded away from 0 on $\Gamma$, then (5.9) implies (5.10).
Proof. First,

$$\sup_{\mathcal{D}} \left| \sum_i Q_B D_i - \sum_i Q D_i \right| \leq k \max_{i=1, \ldots, k} \sup_{\mathcal{D}_i} |Q_B D_i - Q D_i| \to 0 \text{ a.s.} \quad (5.13)$$

by (5.9). Also,

$$\sup_{\mathcal{D}} \left| \frac{\sum_i Q D_i}{\sum_i Q_B D_i} - 1 \right| \leq \sup_{\mathcal{D}} \frac{1}{\sum_i Q_B D_i} \sup_{\mathcal{D}} \left| \sum_i Q_B D_i - \sum_i Q D_i \right|$$

$$\leq c_\Gamma \sup_{\mathcal{D}} \left| \sum_i Q_B D_i - \sum_i Q D_i \right| \to 0 \quad (5.14)$$

by (5.13), where $c_\Gamma$ is finite due to the assumption concerning the acceptance probability. To see that $c_\Gamma$ exists (and is finite), observe that

$$Q D_i(\theta) = p_i(\theta) Q(\Lambda_A(\theta))$$

which implies that

$$\sum_i Q D_i(\theta) = Q(\Lambda_A(\theta)) \sum_i p_i(\theta) = Q(\Lambda_A(\theta)) \geq c_1 > 0 \quad , \quad \forall \theta \in \Gamma$$

where $c_1 > 0$ because the acceptance probability is strictly bounded away from 0 on $\Gamma$. Thus there is $c_2$ and $\epsilon$ such that

$$\sum_i Q D_i(\theta) - \epsilon \geq c_2 > 0 \quad , \quad \forall \theta \in \Gamma \quad .$$

By (5.13), there exists $B_\epsilon$ such that
\[ \sum_i QD_i - \epsilon \leq \sum_i Q_B D_i \leq \sum_i QD_i + \epsilon, \forall B \geq B_\epsilon, \forall D(\theta) \in \mathcal{D} \]

\[ \Rightarrow \sup_{\mathcal{D}} \frac{1}{\sum_i Q_B D_i} \leq \sup_{\mathcal{D}} \frac{1}{\sum_i QD_i - \epsilon} \]

\[ = \sup_{\theta \in \Gamma} \frac{1}{\sum_i QD_i(\theta) - \epsilon} \leq \frac{1}{c_2} < \infty, \forall B \geq B_\epsilon, \]

hence \( c_\Gamma \) exists and is finite (take \( c_\Gamma := c_2^{-1} \)). Finally, a little algebra shows that

\[ \sup_{\mathcal{D}} \left| \frac{Q_B D_i}{\sum_i Q_B D_i} - \frac{QD_i}{\sum_i QD_i} \right| \]

\[ \leq \sup_{\mathcal{D}} \frac{1}{\sum_i QD_i} \left\{ \left( \sup_{\mathcal{D}} QD_i + \epsilon \right) \sup_{\mathcal{D}} \left| \frac{\sum_i QD_i}{\sum_i Q_B D_i} - 1 \right| + \sup_{\mathcal{D}} |Q_B D_i - QD_i| \right\} \]

\[ \leq \frac{1}{c_1} \left\{ (1 + \epsilon) \sup_{\mathcal{D}} \left| \frac{\sum_i QD_i}{\sum_i Q_B D_i} - 1 \right| + \sup_{\mathcal{D}} |Q_B D_i - QD_i| \right\} \]

which tends to 0 by (5.14) and (5.9).

The bottom line then for adapting the rejection step in any simulation algorithm

is establishing that

\[ \mathcal{D}_i = \{ D_i(\theta) : \theta \in \Gamma \} = \{ \mathbf{u} : \mathbf{u} \in \Lambda_i(\theta) \cap \Lambda_\mathcal{A}(\theta) \} : \theta \in \Gamma \}

is a VC class. Once this is done, we may conclude that

\[ \sup_{\Gamma} |p^*_i(\theta) - p_i(\theta)| \to 0 \]

where \( p^*_i(\theta) \) and \( p_i(\theta) \) are defined as at (5.11) and (5.12), subject to the usual type of condition that the acceptance probability is strictly bounded away from 0 on \( \Gamma \).
5.3 Cancer Model of Atkinson, Bartoszyński, Brown and Thompson

Certainly an excellent application which illustrates the usefulness of SIMEST is the cancer modeling work of Atkinson, Bartoszyński, Brown, and Thompson [1, 1983]. For a more complete exposition than what we will present here, we refer to Thompson, Atkinson, and Brown [63, 1987]. We quote some directly from that source:

Let us now consider a stochastic process model for the description of the occurrence and growth of secondary tumors. The model was motivated by [3, 1982], in which there was an indication that the hazard of discovery of secondary tumors after removal of the primary appeared to be nearly constant in time. This led to the postulating of a model in which secondary tumors were sometimes produced by a systemic mechanism with constant intensity in addition to, rather than solely by, an accepted metastatic process whose intensity is proportional to primary tumor size. We had data sets in which we had records of the time from the removal of the primary tumor to the first discovery of a secondary tumor. The model was based on four axioms:

- **H1**: For each patient each tumor originates from a single cell and grows exponentially at rate $\alpha$.

- **H2**: The probability that a tumor of size $Y_j(t)$, not previously detected and removed to time $t$, is detectable in $[t, t + \Delta t]$ is $bY_j(t)\Delta t + o(\Delta t)$.

- **H3**: Until the removal of the primary the probability of metastasis in $[t, t + \Delta t]$ is $aY_0(t)\Delta t + o(\Delta t)$.

- **H4**: The probability of systemic occurrence of a tumor in $[t, t + \Delta t]$ is $\lambda\Delta t + o(\Delta t)$, independent of the prior history of the patient.

While axioms such as H1-H4 are easily stated, they lead to enormous tangles in the likelihood (cf. Thompson, Atkinson, and Brown [63] for more details). However,
axioms H1-H4 lend themselves quite easily to simulation. We define some random variables

\[ W_D = \text{time of detection of primary} \]
\[ W_M = \text{time of origin of first metastasis} \]
\[ W_S = \text{time of origin of first systemic tumor} \]
\[ W_R = \text{time of origin of first recurrent tumor} \]
\[ W_d^* = \text{time from } W_R \text{ to detection of first recurrent tumor} \]
\[ W_d^{**} = \text{time from } W_D \text{ to detection of first recurrent tumor} \]

Leaving the question of how to simulate these random variables aside for the moment, we write down a simulation algorithm (to be subsequently referred to as “algorithm SM”) for the model postulated in H1-H4:
SM Input \( \alpha, \lambda, a, b \)

Repeat until \( s > 0 \)

Generate \( w_D \)

Generate \( w_M \)

If \( w_M > w_D \), then \( w_M \leftarrow \infty \)

Generate \( w_S \)

\( w_R \leftarrow \min(w_M, w_S) \)

Generate \( w_d^* \)

\( w_D^* \leftarrow w_R + w_d^* - w_D \)

\( s = w_D^* \)

If \( s < 0 \), discard

End repeat

Return \( s \)

Observe that algorithm SM proceeds as follows:

1. Let \( w_D^* \) be \( w_M + w_d^* - w_D \) on \([w_M \leq w_D] \cap [w_M \leq w_S]\) and let \( w_D^* \) be \( w_S + w_d^* - w_D \) on \( ([w_M \leq w_D] \cap [w_M > w_S]) \cup [w_M > w_D] \).

2. Reject \( w_D^* \) if \( w_D^* < 0 \)

For clarity, consider the following notation:

\[
A_1(\theta) = [w_M(u) + w_d^*(u) - w_D(u) \geq 0] \quad , \quad A_2(\theta) = [w_M(u) \leq w_D(u)]
\]
\[ A_3(\theta) = [w_M(u) \leq w_S(u)] \ , \ A_1^a(\theta) = [w_M(u) + w_d^*(u) - w_D(u) \geq a] \]

\[ A_1^b(\theta) = [w_M(u) + w_d^*(u) - w_D(u) < b] \ , \ B(\theta) = [w_S(u) + w_d^*(u) - w_D(u) \geq 0] \]

\[ B^a(\theta) = [w_S(u) + w_d^*(u) - w_D(u) \geq a] \ , \ B^b(\theta) = [w_S(u) + w_d^*(u) - w_D(u) < b] \ . \]

These are sets in \( \mathcal{U}^4 \), where \( \mathcal{U} = (0, 1) \), that is to say \( u = (u_1, u_2, u_3, u_4) \) is a vector of any four numbers between 0 and 1. The reader will anticipate the fact that we will simulate these numbers from a uniform \((0, 1)\) distribution; however, in the construction of VC classes the distribution of \( u \) does not matter at all.

By the time we begin our simulations, we will have decided upon some set of bins for \( \mathbb{R}^+ \), denote them by \( \{[a_i, b_i] : i = 1, \ldots, k\} \) where \( a_1 = 0 \) and \( b_k = \infty \). We recall from our discussion concerning rejection algorithms that we must show that

\[ \{[\Lambda_i(\theta) \cap \Lambda_A(\theta)] : \theta \in \Gamma \} \]

is a VC class, where \( \Lambda_i(\theta) = [w_D(u) \in I_i] \) and \( \Lambda_A = [w_D(u) \geq 0] \). Well

\[ \{[\Lambda_i(\theta) \cap \Lambda_A(\theta)] : \theta \in \Gamma \} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3 \]

\( (5.15) \)

where

\[ \mathcal{H}_1 = \{A_1^{a_i}(\theta) \cap A_1^{b_i}(\theta) \cap A_1(\theta) \cap A_2(\theta) \cap A_3(\theta) : \theta \in \Gamma \} \]
\[ \mathcal{H}_2 = \{ B^{a_i}(\theta) \cap B^{b_i}(\theta) \cap B(\theta) \cap A_2(\theta) \cap A_3(\theta)^c : \theta \in \Gamma \} \]

\[ \mathcal{H}_3 = \{ B^{a_i}(\theta) \cap B^{b_i}(\theta) \cap B(\theta) \cap A_2(\theta)^c : \theta \in \Gamma \} \]

We recall from Result B.2 that (5.15) will be a VC class provided that we demonstrate that

\[ \{ A_j(\theta) : \theta \in \Gamma \} , \quad j = 1, 2, 3 , \quad \{ A_i^{a_i}(\theta) : \theta \in \Gamma \} , \quad \{ A_i^{b_i}(\theta) : \theta \in \Gamma \} \]

\[ \{ B(\theta) : \theta \in \Gamma \} , \quad \{ B_i^{a_i}(\theta) : \theta \in \Gamma \} , \quad \{ B_i^{b_i}(\theta) : \theta \in \Gamma \} \]

are all VC classes. We do this now.

By H1, the tumor volume at time \( t \) is \( v(t) = c \exp(\alpha t) \) where \( c \) is the volume of one cell. From the model axioms H1-H4, we are led to\(^6\)

\[ F_D(t_D) = 1 - \exp \left( -\frac{bc}{\alpha} \left[ \exp(\alpha t_D) - 1 \right] \right) , \quad F_{d^*}(t_{d^*}) = 1 - \exp \left( -\frac{bc}{\alpha} \left[ \exp(\alpha t_{d^*}) - 1 \right] \right) \]

\[ F_M(t_M) = 1 - \exp \left( -\frac{ac}{\alpha} \left[ \exp(\alpha t_M) - 1 \right] \right) , \quad F_S(t) = 1 - \exp(-\lambda t) \]

so that if we generate \( u \) as uniform \((0, 1)\), we generate simulated values of the desired random variables via

---

\(^6\)These cdf's correct typos in Thompson, Atkinson, and Brown [63].
\[ t_D = \frac{1}{\alpha} \log \left( -\frac{\alpha}{bc} \log(1 - u) + 1 \right) = h_D^D(u), \quad t_{d^*} = \frac{1}{\alpha} \log \left( -\frac{\alpha}{bc} \log(1 - u) + 1 \right) = h_{d^*}^D(u) \]

\[ t_M = \frac{1}{\alpha} \log \left( -\frac{\alpha}{ac} \log(1 - u) + 1 \right) = h_M^D(u), \quad t_S = -\frac{1}{\lambda} \log(1 - u) \]

It will be convenient to replace \( \log(1 - u) \) with \( v \) here, so that \( t_D = h_D^D(u) = g_D^D(v) \)

where \( g_D^D(v) := \alpha^{-1} \log(-\frac{\alpha}{bc}v + 1) \) and similarly for \( g_{d^*}^D(v) \) and \( g_M^D(v) \). This is justified

by the mapping argument and Result B.3, but would clutter the presentation if we

waited any longer to do it. We now treat each of the sets of (5.16) in turn. First

\[ A_1(\theta) = [t_M + t_{d^*} - t_D \geq 0] = [g_M^D(v) + g_{d^*}^D(v) - g_D^D(v) \geq 0] \]

\[ = \left\{ \frac{1}{\alpha} \log \left( -\frac{\alpha}{ac}v_2 + 1 \right) + \frac{1}{\alpha} \log \left( -\frac{\alpha}{bc}v_3 + 1 \right) - \frac{1}{\alpha} \log \left( -\frac{\alpha}{bc}v_1 + 1 \right) \geq 0 \right\} \]

\[ = \left\{ \frac{\alpha}{ac}v_2v_3 - \frac{\alpha}{ac}v_2 - \frac{\alpha}{bc}v_3 + 1 \geq -\frac{\alpha}{bc}v_1 + 1 \right\} \]

so that \( \{A_1(\theta) : \theta \in \Theta\} \) is a VC class by Result B.1. Clearly \( \{A_i^D(\theta) : \theta \in \Theta\} \) and

\( \{A_i^b(\theta) : \theta \in \Theta\} \) are also. Next

\[ B^a(\theta) = \left[ -\frac{1}{\lambda} v_4 + \frac{1}{\alpha} \log \left( -\frac{\alpha}{bc}v_3 + 1 \right) - \frac{1}{\alpha} \log \left( -\frac{\alpha}{bc}v_1 + 1 \right) \geq a \right] \]

\[ = \left[ -\frac{\alpha}{\lambda} v_4 + \log \left( -\frac{\alpha}{bc}v_3 + 1 \right) - \log \left( -\frac{\alpha}{bc}v_1 + 1 \right) - \log(\exp(a\alpha)) \geq 0 \right] \]
\[ \frac{1}{v_4} \log \left( \frac{-\frac{\alpha}{bc} v_3 + 1}{e^{\alpha \left( -\frac{\alpha}{bc} v_1 + 1 \right)}} \right) \leq \frac{\alpha}{\lambda} = \left[ -\frac{\alpha}{bc} v_3 v_5 + v_5 \leq -e^{\alpha} e^{\alpha/\lambda} \frac{\alpha}{bc} v_1 + e^{\alpha} e^{\alpha/\lambda} \right] \]

where \( v_5 := e^{1/v_4} \) and the last equality follows by exponentiating both sides in the previous set, then rearranging. Thus \( \{B^\alpha(\theta) : \theta \in \Theta\} \) is a VC class by Result B.1 and Result B.2. A similar argument works for \( B^\beta(\theta) \). Also \( B(\theta) = B^\alpha(\theta) \mid_{n=0} \), so that \( \{B(\theta) : \theta \in \Theta\} \) is also a VC class. Next

\[ A_2(\theta) = [g_\theta^M(v_2) \leq g_\theta^P(v_1)] = \left[ \frac{1}{\alpha} \log \left( -\frac{\alpha}{ac} v_2 + 1 \right) \leq \frac{1}{\alpha} \log \left( -\frac{\alpha}{bc} v_1 + 1 \right) \right] \]

\[ = \left[ -\frac{\alpha}{ac} v_2 + 1 \leq -\frac{\alpha}{bc} v_1 + 1 \right] \]

and

\[ A_3(\theta) = [g_\theta^M(v_2) \leq \bar{g}_\theta^P(v_4)] = \left[ \frac{1}{\alpha} \log \left( -\frac{\alpha}{ac} v_2 + 1 \right) \leq -\frac{1}{\lambda} v_4 \right] \]

\[ = \left[ -e^{-\lambda/\alpha} \frac{\alpha}{ac} v_2 + e^{-\lambda/\alpha} \geq e^{v_4} \right] \]

so that \( \{A_2(\theta) : \theta \in \Theta\} \) and \( \{A_3(\theta) : \theta \in \Theta\} \) are VC classes.

The reader may observe that at times we are nearly stretched to the limit of our techniques for VC classes based on finite dimensionality. For the case of \( \{B^\alpha(\theta) : \theta \in \Theta\} \) (as the toughest example) we point out that we were about one non-linearity away from not being able to show what we needed.
Chapter 6

A Data-based Selection Criterion for SIMDST for Improved Parameter Estimation with Multidimensional Data

In this chapter, we propose the following data-based method for constructing a hybrid parameter estimate when one has multivariate data. Consider the case of 5D data as an example. Suppose that parameter estimation is feasible using:

1. only the first dimension of this data,

2. the first-third dimensions of this data, and

3. all five dimensions of this data.

Denote by $Q_i(\cdot)$, $i = 1, 3, 5$, Pearson’s $X^2$ based on 1D, 3D, and 5D data. For 1D data, assume the equal probability bins approach is used.\(^1\) For 3D and 5D data, assume that our nearest neighbor based concentric ellipsoidal bins approach is used (as we described in Section 3.2). Suppose $k$ is fixed to be the same for all three $Q_i$’s (e.g. if $k = 30$, use 30 bins for 1D, 3D, and 5D). One obtains three estimators

$$
\hat{\theta}_i = \arg\min_{\theta \in \Theta} Q_i(\theta) , \quad i = 1, 3, 5
$$

\(^1\)By “equal probability bins approach” is meant that the bins are chosen so that roughly the proportion $1/k$ of the observations fall in each bin.
and it is desired to determine which $\hat{\theta}_i$ is best.

We propose the following data-driven method for selecting a good $\hat{\theta}$. Suppose $s = \dim(\theta) = 4$ (as it is in this chapter). Construct a collection of $\hat{\theta}$ candidates (call this collection $\Xi$) by considering all $3^4 = 81$ possible combinations of $\hat{\theta}_1, \hat{\theta}_3,$ and $\hat{\theta}_5$ (i.e. there are 3 choices for each parameter, the 1D, 3D, or 5D estimate; and there are 4 parameters). Choose a "best" estimator from $\Xi$ via

$$\hat{\theta}_{hyb} = \arg\min_{\theta \in \Xi} \{Q_1(\theta) + Q_3(\theta) + Q_5(\theta)\} \quad (6.1)$$

As demonstrated by a lengthy example in this chapter, such a hybrid estimator will generally outperform or match all of its competitors (its competitors being $\hat{\theta}_1, \hat{\theta}_3, \hat{\theta}_5,$ and $\hat{\theta}_{13}$) in terms of MSE (e.g. if $\theta = (\alpha, b, a, \lambda)$ then the MSE of $\hat{\alpha}_{hyb}$ will match or beat the MSE of $\hat{\alpha}_1, \hat{\alpha}_3,$ etc.). This data-driven performance is remarkable and is the main success of this chapter.

The fact that a construction such as (6.1) is even discussed illustrates the dramatic differences between univariate and multivariate problems. As should be made clear in this chapter, there is nothing but greater advantage to be gained from a multivariate approach. However, our investigations may fail to discover such advantages if we proceed too naively (in particular, if we insist on clinging to our univariate intuition).

These ideas are illustrated in this chapter with a lengthy investigation of the cancer model of Atkinson, Bartoszyński, Brown and Thompson [1, 1983]. In Section 6.1, we discuss some of the relevant details of this model again (some aspects of
the model were already discussed in Section 5.3). In Section 6.2, we investigate the performance of our proposed multivariate SIMEST methodology applied to the situation of 3-dimensional data observed from such a model (it is assumed that one has (i) the time from the detection and removal of the primary until the discovery of the first secondary tumor, (ii) the volume of the primary, and (iii) the volume of the first secondary tumor). In Section 6.3, we investigate the performance of this same methodology applied to the situation of 5-dimensional data (in addition to the data assumed available for the 3D case, it is assumed that one has (iv) the time from the discovery of the first secondary tumor until the discovery of the second secondary tumor, and (v) the volume of the second secondary tumor). In Section 6.4, we discuss at length various computational aspects involved in the work done in this chapter. In Section 6.5 we briefly mention an exciting idea for an extension of this methodology.

6.1 Cancer Model of Atkinson, Bartoszyński, Brown and Thompson

This chapter begins by reviewing some of the basics of the cancer model of Atkinson, Bartoszyński, Brown and Thompson [1, 1983]. Recall from Section 5.3 of Chapter 5, the 5 axioms of this model:

- H1: For each patient each tumor originates from a single cell and grows exponentially at rate $\alpha$. 
• H2: The probability that a tumor of size $Y_j(t)$, not previously detected and removed to time $t$, is detectable in $[t, t + \Delta t]$ is $bY_j(t)\Delta t + o(\Delta t)$.

• H3: Until the removal of the primary the probability of metastasis in $[t, t + \Delta t]$ is $aY_0(t)\Delta t + o(\Delta t)$.

• H4: The probability of systemic occurrence of a tumor in $[t, t + \Delta t]$ is $\lambda\Delta t + o(\Delta t)$, independent of the prior history of the patient.

At first glance, this set of axioms may appear fairly complicated, but such is not the case. Figure 6.1 describes this axiom set pictorially, and the mechanisms involved in the model can be easily understood from this picture. The data considered by Atkinson, Bartoszyński, Brown, and Thompson in [1, 1983] consisted of the times from the discovery and removal of the primary until the discovery of the first secondary tumor. Those authors found that a model for breast cancer where secondary tumors could only be generated by a metastatic mechanism was just impossible to fit. This led them to postulate that perhaps the secondary tumors could also be generated by a systemic mechanism (i.e. a mechanism which is independent of the primary tumor). Under this model, then, secondary tumors arise as the result of 2 competing stochastic mechanisms: (a) a metastatic mechanism and (b) a systemic mechanism.

The model makes (mild) parametric assumptions about the processes which govern these mechanisms. The systemic process is stationary Poisson with rate $\lambda$. The metastatic process is a nonhomogeneous Poisson process (sometimes abbreviated
n.h.p.p.) with intensity function $aY_0(t)$ where $Y_0(t)$ is the volume of the primary at time $t$ and $a$ is a parameter to be estimated. The n.h.p.p. will be discussed more in Section 6.3. For the reader not familiar with the n.h.p.p., the intuition is simply that with ever-increasing volume of the primary tumor metastatic arrivals become more and more frequent.

With enormous effort, Bartoszyński, Brown, and Thompson ([3, 1982]) were able to obtain the closed form likelihood for this model. But even having the likelihood in hand, there was a substantial amount of numerical difficulty remaining. In particular, the order of numerical effort required to make pointwise function evaluations of this
likelihood was comparable to 4D quadrature. The maximum likelihood fit of this four parameter model yielded the estimates:

$$\alpha = .31 \quad , \quad b = .23 \times 10^{-8} \quad , \quad a = .17 \times 10^{-9} \quad , \quad \lambda = .003 .$$  

With these parameters, roughly 95% of the secondary tumors are due to the systemic mechanism. Intuition suggests that with such parameter choices, the estimation of \( \alpha \) should be difficult, while the estimation of \( \lambda \) should be quite easy.

### 6.2 Estimation for 3D Data

Suppose that some additional information about tumor display for each person was available. Would multivariate data improve the parameter estimation for this model dramatically? Specifically, start by assuming that, in addition to the 1D time data (already described) that one also has (a) the volume of the primary tumor at discovery and (b) the volume of the first secondary tumor at discovery. While the author's advisor (Thompson) has seen such data (as a matter of fact volume data was also available for the work in [1, 1983]; however, the quality of such data was very poor), good volume data does not currently seem to be readily available. In light of this difficulty, it was decided that this investigation should be conducted using simulated data, hoping that good volume data may become available in the future.

As the true model parameters for the simulated data, the MLE estimates at (6.2) are used (these values being based on real data). Also, as the most obvious first
approximation to reality, it is assumed that the growth of tumors is deterministic (specifically $c \exp(\alpha t)$, where $t$ is time and $c$ is the volume of a single cell).

Some Important Simulation Details about the Model

In Section 5.3, a simulation algorithm (algorithm SM) for this model was described in some detail. One important simplifying assumption used in algorithm SM is that the first secondary tumor generated is the first secondary tumor to be observed. Such an assumption ignores the possibility that the (final) ordering of the discoveries may be reversed due to the stochastic nature of the discovery process. For the present section (Section 6.2) this simplifying assumption will also be used. In the next section (Section 6.3, which deals with 5D data), this assumption is relaxed. Note that this simplification was originally used in order to aid in obtaining the likelihood in closed form. Such simplifications are of course not needed in a simulation-based approach.

In terms of the notation used to describe algorithm SM in Section 5.3, the deterministic assumption on the volumes leads to

$$y_0 = c e^{\alpha w_D}, \quad y_t = c e^{\alpha w_d^*}$$

since $w_D$ is the (simulated) time from the origin of the primary until its detection and $w_d^*$ is the time from the origin of the first secondary tumor until its detection.\(^2\)

\(^2\)Here $y_0$ denotes the volume of the primary at discovery and $y_t$ denotes the volume of the first secondary tumor at discovery.
Otherwise, algorithm SM is unchanged. Thus our 3D simulations will return the vector \((s, y_o, y_1)\), one time and two volumes.

For future reference (in Section 6.3), the reader will notice in Section 5.3 that \(w_M\) is the first arrival from a nonhomogeneous Poisson process. Importantly, one does not generate subsequent arrivals from the same c.d.f., i.e.

\[
F_M(t_m) = 1 - \exp \left( -\frac{ac}{\alpha} (e^{at_M} - 1) \right) .
\]

Generating subsequent arrivals will be discussed in Section 6.3. Looking only at the first arrivals from two competing stochastic processes as is done in this section (Section 6.2) is not nearly so dramatic an illustration of the power of SIMEST as is looking at second and third arrivals from competing stochastic processes as is done in the next section (Section 6.3). For second and third arrivals, one expects that the estimation theory would be so difficult that a simulation-based approach may very well be the only viable option.

Another important detail of the simulations for this 3D model (which may not be so obvious from SM at first glance) is that once the primary has been removed only metastases which have already been born will survive. That is to say, no metastases can be born once the primary has been removed. Still further, we do not allow secondary tumors to themselves metastasize.
Sample Sizes, Bin Choices, and Speed Limitations

For most of the details about the multivariate fitting of the ellipsoids, the numerical search algorithm used and other computational aspects related to the work in this chapter see the Computational Section (Section 6.4). Essentially due to constraints on computing speed, it was decided to investigate samples of size 200, 400, and 600 (the sample size will be referred to as $n$) and bin choices of 10, 20, and 30 (the number of bins will be referred to as $k$). These choices of $n$ are of interest from a practical standpoint. For example, the data originally used to obtain the MLE's for the 1D model consisted of roughly 200 observations. After having examined all of the simulation results (details to follow soon), it became evident that additional results for larger $n$ and for more choices of $k$ (particularly for the 5D results in Section 6.3) would have been desirable. Computing speed, however, is still a prohibitive factor for such choices. Small $n$ and small $k$ (like those used here) result in reasonable speed. But the computational effort goes up linearly with $k$ (as should be fairly obvious), and also goes up linearly in $n$, as the simulation sizes used for SIMEST must increase.

---

As a point of detail, there is a rejection step in the simulation algorithm which rejects simulations for which the time of discovery of the first secondary tumor precedes the time of discovery of the primary tumor (this rejection step is pointed out in algorithm SM). On a computational level it is faster, for example, to ask for $n = 200$ and return only those sims not rejected, than it is to ask that one continues to simulate until exactly $n = 200$ are obtained. Typically only about 2% or so of the simulations are rejected under this model so the effective sample sizes are very close to $n = 200, 403,$ and 600, but in all cases will be a little less than that.

Further along this line, the code used in this implementation adjusts $k$ depending on the number of observations accepted, so that the number of bins actually used is not necessarily going to be precisely the number of bins which were requested. For $k = 10$, for example, sometimes there may be as few as 8 or as many as 12 bins used. This adjustment in the code is more important for smaller $n$ and/or for models where the proportion of simulations rejected may be quite large.
proportionally to the sample size. For example, in the results to be discussed below the simulation sizes on the "last pass"\footnote{We explain what is meant by the "last pass" in the Computational Section, Section 6.4. Quickly, several passes of the search routine are used as a technique to improve the performance of the search. Simulation sizes are increased with each additional pass. The largest number of simulations used is on the final pass.} were $6 \cdot n$, where $n$ is the sample size. So for $n = 200, 400, 600$, the number of simulations increases like $1200, 2400, 3600$, and it is clear that computing times (at least for looking at many data sets) become prohibitive fairly quickly. More details concerning computing speed are given in the Computational Section (Section 6.4).

At any rate, the simulated data sets used here were generated along the same sequence as $n$ was increased, that is to say, 200 additional (simulated) observations were added to a given set of 200 (simulated) observations in going from $n = 200$ to $n = 400$ (rather than the alternative which would be to simulate 400 brand new observations independently of the first 200). This was also done across choices of $k$, e.g. for $n = 200$, the same set of 200 simulated observations are used for $k = 10, 20,$ and 30. Thus there are 40 simulated data sets used here, each of size $n = 600$. This seems reasonable, as - for example - one would like for apparent differences in performance due to different choices of $k$ to reflect only the effects of the different choices of $k$ rather than to be confounded with effects that may merely be due to the randomness in the data.
Some Measures of Good Performance

One wonders what a natural measure of "good performance" might be. One measure which may be used in a simulation scenario (i.e. where we know the true values of the parameters) is mean square error (abbreviated MSE), i.e.

\[
\frac{1}{n_s - 1} \sum_{j=1}^{n_s} (\hat{\theta}_j - \theta_0)^2
\]

where \( \theta_0 \) is the true parameter and \( \hat{\theta}_1, \hat{\theta}_2, \ldots \) are the parameter estimates (for us, \( n_s = 40 \), there are 40 estimates of \( \theta_0 \) for each of the different choices of \( n \) and \( k \) considered). But right away it is clear that such an approach has the disadvantage that one must look at one parameter at a time. With a likelihood in hand it might be possible to do a sensitivity analysis to determine what a large change for each parameter might be. With SIMEST of course one has no such luxury. Hence while MSE will be one of the tools used to investigate performance, one must choose some other measure which reflects well the aggregate performance of the whole vector \( \hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\alpha}, \hat{\lambda}) \) jointly.

An obvious aggregate measure is just the SIMEST metric itself, i.e. \( X^2(\theta) \). With multivariate data, there are several different versions of \( X^2 \) to consider. Let \( Q_1(\theta) \) denote Pearson's chi-square based on 1D data (specifically, for 1D data the equal probability approach of putting a proportion of roughly \( 1/k \) observations in each bin is used), and let \( Q_3(\theta) \) denote Pearson's chi-square based on 3D data (using the concentric elliptoidal bins, as previously described). Since there are several different
versions of $X^2$ floating around, there are also several different parameter estimates floating around, namely

$$\hat{\theta}_1 = \arg \min_{\theta \in \Theta} Q_1(\theta)$$

and

$$\hat{\theta}_3 = \arg \min_{\theta \in \Theta} Q_3(\theta).$$

One would obviously expect that $Q_3(\hat{\theta}_3)$ is smaller than $Q_3(\hat{\theta}_1)$. It was hoped that $Q_1(\hat{\theta}_3)$ would also generally be smaller than $Q_1(\hat{\theta}_1)$; however, this turned out not to be the case. Hence it was decided to aggregate over both $Q_1(\cdot)$ and $Q_3(\cdot)$ in some reasonable fashion to obtain a better overall measure. The choice of such a metric is somewhat arbitrary, of course. Suppose that $\Xi$ is some collection of estimates of $\theta_0$ under consideration, then

$$g_{13}(\theta) := \{Q_1(\theta) + Q_3(\theta)\}$$

(6.3)

and

$$h_{13}(\theta) := \max\{Q_1(\theta), Q_3(\theta)\}$$

(6.4)

are 2 reasonable criteria which may be used in order to select a "good" $\theta$ from $\Xi$. Since small $Q_i$'s are good, this rule takes

$$\hat{\theta}_{agg} = \arg \min_{\theta \in \Xi} g_{13}(\theta)$$

(6.5)
and similarly

$$\hat{\theta}_{agg2} = \arg\min_{\theta \in \Xi} h_{13}(\theta) \quad (6.6)$$

as parameter estimates. These estimates should perform well with respect to both $Q_1$ and $Q_3$ relative to the elements in the $\Xi$ class.

But even more can be done than this. Preliminary results for this 3D data (to follow shortly) suggested that $\hat{\theta}_1$ estimates $\lambda$ much better than $\hat{\theta}_3$, while $\hat{\theta}_3$ estimates $\alpha$, $b$, and $a$ generally better than $\hat{\theta}_1$. Hence it seemed clear that a good overall estimator might be

$$\hat{\theta}_{hyb} = (\hat{\alpha}_{3D}, \hat{\beta}_{3D}, \hat{\alpha}_{3D}, \hat{\lambda}_{1D}) \ .$$

But of course such a choice of a hybrid estimator was based on MSE performance, i.e. knowledge of the true $\theta_0$, which is cheating. What is clearly indicated then is to choose a hybrid estimator, call it $\hat{\theta}_{13}$ constructed from the components of $\hat{\theta}_1$ and $\hat{\theta}_3$ in some data-based manner which will outperform both $\hat{\theta}_1$ and $\hat{\theta}_3$ with respect to the aggregate measure. There is no reason to expect that improved MSE performance will automatically follow merely because a hybrid estimator performs well with respect to the aggregate measure, but - as we show later - it turns out that this does happen. This happy circumstance is the main success of this chapter. So as the pool of candidates, $\Xi$, for this situation take all $2^4 = 16$ possible combinations of $\hat{\theta}_1$ and $\hat{\theta}_3$. 
For the aggregate criterion, take the $L_1$ criterion, i.e. the criterion at (6.3). Such a method for choosing the hybrid estimator is entirely data-driven

3D Estimation and Results

Refer now to Figure 6.2. Clearly if $\hat{\theta}_1$, $\hat{\theta}_3$ and $\hat{\theta}_{13}$ are ranked with respect to the criterion (6.3), $\hat{\theta}_{13}$ will always rank first (as this is precisely how it was constructed). But $\hat{\theta}_{13}$ also performs well with respect to the minimax metric (6.4) as indicated in Figure 6.2. Roughly 70-80% of the time $\hat{\theta}_{13}$ ranks first out of the three contenders ($\hat{\theta}_1$, $\hat{\theta}_3$, and $\hat{\theta}_{13}$) with respect to the minimax rule. This figure also clearly indicates the fact that $\hat{\theta}_3$ generally ranks second and that $\hat{\theta}_1$ generally ranks last. Thus, although (i) $Q_1(\hat{\theta}_3)$ tends to be worse (i.e. larger) than $Q_1(\hat{\theta}_1)$ and (ii) $Q_3(\hat{\theta}_1)$ tends to be worse than $Q_1(\hat{\theta}_1)$, more frequently it is the case that $\max\{Q_1(\hat{\theta}_1), Q_3(\hat{\theta}_1)\}$ exceeds $\max\{Q_1(\hat{\theta}_3), Q_3(\hat{\theta}_3)\}$. That is, $\hat{\theta}_1$ is generally worse in terms of $Q_3$ than $\hat{\theta}_3$ is in terms of $Q_1$.\(^5\)

Refer now to Figure 6.3. In this figure, an investigation is made into which parameter estimates the hybrid estimator seems to prefer (that is to say, for each of the four parameters the hybrid has to decide whether it prefers the 1D estimate or the 3D estimate). The plots clearly indicate that the trends reflected here do not depend very much on the choice of $k$ or $n$. The plots also indicate that the hybrid estimator

\(^5\)As an aside, there was nothing at all special about (6.3), it was just the first criterion that was tried, and it worked out fine.

\(^6\)Note that the magnitudes of $Q_1$ and $Q_3$ are being compared fairly, both are $\chi^2_{k-s-1}$, this doesn’t depend on the dimension of the data.
Figure 6.2
prefers $\hat{a}_{3D}$ and $\hat{b}_{3D}$ over 1D, $\hat{\lambda}_{1D}$ over 3D, and that it does not seem to express much preference for the $\alpha$ parameter.

Refer now to Figure 6.4. In this figure, the MSE performance of $\hat{\theta}_1$, $\hat{\theta}_3$ and $\hat{\theta}_{13}$ is investigated. The plots are on a log scale (natural logarithm) to indicate the dramatic performance of the 1D estimate of $\lambda$. To read these plots, notice that the log of $(MSE_x/MSE_y)$ being less than 0 indicates that $x$ is better than $y$ in terms of MSE, similarly log $MSE_x/MSE_y$ being greater than 0 indicates that $y$ is better than $x$. Again, each MSE calculated here is based on 40 data sets. First, observe that $\hat{\theta}_{13}$ outperforms or ties $\hat{\theta}_3$ all the time in terms of MSE. Notably (as mentioned already but shown here for the first time), $\hat{\lambda}_{1D}$ is tremendously better than $\hat{\lambda}_{3D}$. Generally $\hat{a}_{3D}$ and $\hat{b}_{3D}$ are better than 1D and $\hat{a}_{3D}$ seems to be better for larger samples while $\hat{a}_{1D}$ seems to be better for smaller samples (recall that only 5% of the simulated data sets result from the metastatic mechanism, one would be surprised to do well estimating $a$ with samples of $n = 200, 400, \text{or} 600$). As mentioned previously, the practical implications of Figure 6.4 are very important. Constructing the hybrid $\hat{\theta}_{13}$ in an entirely data-based fashion, one is able to uniformly outperform or tie $\hat{\theta}_1$ and $\hat{\theta}_3$ in terms of MSE performance. This uniformity is across different choices of $n$ and $k$. One would believe that such a positive outcome is clearly not a foregone conclusion merely by choosing a hybrid estimator according to (6.5). (6.5) is just one arbitrary (although reasonable) aggregate measure (aggregating across dimensions).
Figure 6.4
It is fortunate that this measure also seems to be closely related to better MSE performance of the resulting hybrid estimator.

Refer now to Figures 6.5-6.8. While MSE is one useful measure of performance, it is not too helpful in indicating skew in the distributions of the parameter estimates. Figures 6.5-6.8 present boxplots of $\hat{\theta}_1$, $\hat{\theta}_3$, and $\hat{\theta}_{13}$ (for each of the 4 parameters separately) for each choice of $n$ and $k$ considered. Again, each boxplot is based on 40 estimates. For the reader not familiar with boxplots, each “box” indicates the central 50% of the data, the line in the middle of the box indicates the median, the lower boundary of the box marks the 25 percentile, and the upper boundary of the box marks the 75th percentile. Values considered to be “outlying” to an extreme degree are indicated with a hash mark (e.g. for $\alpha$, $n = 200$ and $k = 20$, $\hat{\theta}_1$ has many outliers - there appear to be at least 8 or so, while $\hat{\theta}_3$ and $\hat{\theta}_{13}$ have none). All of these boxplots tend to illustrate in detail the general trend seen in the MSE plots. $\hat{\theta}_{13}$ generally tends to choose the most attractive looking estimates from $\hat{\theta}_1$ and $\hat{\theta}_3$. There is dramatic improvement of $\hat{\theta}_{13}$ over $\hat{\theta}_1$ and notable improvement of $\hat{\alpha}_{13}$ over $\hat{\alpha}_1$. Clearly $\hat{\lambda}_{13}$ seems to like $\hat{\lambda}_1$. It is interesting to note that $\hat{\alpha}_1$ is erratic compared to $\hat{\alpha}_3$, suggesting that 1D estimation of $\alpha$ is unstable with respect to the choice of $k$. Interestingly, sometimes the box for $\hat{\alpha}_{1D}$ is tightly clustered around $\alpha_0$ while almost all of the remaining 25% (10 values) are hugely outlying. It is also clear that $\hat{\alpha}_{13}$ tends to match the distribution of $\hat{\alpha}_3$, ignoring that of $\hat{\alpha}_1$. Recall from Figure 6.3
Figure 6.5
that generally 60% of the time or greater, $\hat{\theta}_{13}$ prefers $\hat{\alpha}_3$; this is clearly indicated in the boxplots.

Finally, in Figure 6.9, a more detailed look at the boxplots for $\hat{\theta}_{13}$ is presented, since $\hat{\theta}_{13}$ would be the parameter estimate one would choose as the “best” parameter estimate using the data-driven method. Indicated in these plots are the starting values as well as the true values. For $\alpha$ and $b$ the starting values are a factor of ±2; for $a$ and $\lambda$ the starting values are a factor of ±4. The difficulties associated with the search routine and reasons for starting so close on $\alpha$ and $b$ are discussed elsewhere (see the Computational Section, Section 6.4).

Right away, one notices the outstanding estimation of $\lambda$, this starting a factor of 4 away from the truth and essentially using the 1D estimate. Clearly under the given choice of $\theta$ (that at (6.2)), $\lambda$ seems to dominate the model. The performance for $a$ seems surprising as well. Although the $a$ estimates are clearly biased downwards, the search procedure has no trouble marching down from the upper starting value which is a factor of 4 away.\footnote{Note that in the Computational Section, Section 6.4, it is pointed out that the search starts roughly half the time from the upper starting value and half the time from the lower starting value. Which start to choose is determined by the flip of a fair coin.} Perhaps this can be explained as: near perfect knowledge of the systemic parameter allows for easier estimation of the metastatic parameter.

For $\alpha$ and $b$, the reader might clearly argue that the estimates have surely not moved far from where they started. This is a valid criticism. It is hoped that the reasons for starting so close on $\alpha$ and $b$ have been satisfactorily addressed in the
Figure 6.9
Computational Section (Section 6.4). In defense of the boxplots of $\alpha$ and $b$ in Figure 8, the reader will note that the parameter estimates are completely free to wander *in the wrong direction* if they so desire (and are permitted to wander as far away as a factor of 10). But this does not happen. Further it is clear that the estimates of $\alpha$ and $b$ are not just stalling on the starting value. Rather, in most cases, fully the central 50\% of the estimates are closer to the truth than where they started with the median clearly tending to line up on the truth with increasing $n$ (see $k = 10$ for $b$ as an example). In light of these observations, the position that $\hat{\alpha}_{13}$ and $\hat{b}_{13}$ are merely stalling on the starting values seems untenable, although more work on the problem of global optimization for SIMEST is probably in order.

### 6.3 Estimation for 5D Data

The prospect of using the multivariate approach proposed in this dissertation with 5D data is certainly a very exciting one. As it turns out, however, the 5D results to be presented in this section (Section 6.3) do not add a great deal more to the understanding that has already been obtained with the 3D data of the last section (Section 6.2), although similar estimation performance will be observed for a more sophisticated model. That is to say, nothing dramatic occurs in going from 3D to 5D (as was hoped) at least for the choices of $n$ and $k$ and simulation sizes that were considered here.
A More Sophisticated Model and Some Simulation Details

In Section 6.2, it was assumed that three data values for each individual would be available, namely (i) the time from the discovery of the primary to the discovery of the first secondary tumor, (ii) the volume of the primary and (iii) the volume of the first secondary tumor. In this section, suppose that two additional data values for each individual are available, (iv) the time from the discovery of the first secondary tumor until the discovery of the second secondary tumor and (v) the volume of the second secondary tumor. This is depicted in Figure 6.10.

The reader may wonder if such data would ever be available. The author’s advisor has seen time data (but no volumes) on the first and second secondary tumors in Poland, but has been unable to obtain permission to use it due to proprietary reasons. Hence collecting such data is feasible. With this in mind, we would like to ask: if one did have such data would the estimation improve dramatically? Could one

![5D Data (3 volumes, 2 times)](image)

**Figure 6.10**
support a more sophisticated model with such data that might not be well supported otherwise? If the answer to any of these questions is a resounding yes, then perhaps this simulation work may motivate the collection of such data in the future.

Recall from Section 6.2 the simplifying assumption used that the first secondary tumor generated is the first secondary tumor observed. In this 5D situation where second secondary tumors are considered, such a simplifying assumption should clearly be relaxed if possible. This relaxation is done now.

Generate several "candidate" secondary tumor times, choosing the smallest feasible one as the first secondary tumor time and the next smallest feasible one as the second secondary time. By the term "feasible" it is meant mainly that a candidate secondary tumor time must arrive after the primary discovery time. In symbols, define the random variables

- \( W_{M_i} (i = 1, 2, \ldots) \) = time from origin of primary to the birth of the \( i \)th metastasis (the model assumes that no metastases can be born once the primary has been removed)

- \( W_{S_i} (i = 1, 2, \ldots) \) = time from the origin of the primary to the birth of the \( i \)th systemic tumor

- \( W_{d_i}^* (i = 1, 2, \ldots) \) = time from origin of any secondary tumor until its discovery

Then the "candidate" secondary tumor times will be
\[ W_{M_i} + W_{2^i-1}^*, ~ i = 1, 2, \ldots, ~ W_{S_i} + W_{2^i}^*, ~ i = 1, 2, \ldots. \]

As a practical matter, of course, one cannot wait to simulate infinitely many of anything. To avoid too much of a computational burden (speed is a real issue), just 3 candidate secondary tumor times are generated from each mechanism (6 total). If less than 2 of these candidate times are feasible, then all 6 are rejected and another attempt is made (again, it was found here as before that the rejection percentage under the assumed model is very small \( \approx 2\% \)). For clarity, these ideas are reexpressed as an algorithm:

- Input \( \alpha, b, a, \lambda \)

- Repeat until \( \geq 2 \) feasible candidate secondary times

  - generate \( w_D \) (recall that \( w_D \) is the time from the origin of the primary until its discovery)

  - generate \( w_{M_i}, ~ i = 1, 2, 3 \)

  - generate \( w_{S_i}, ~ i = 1, 2, 3 \)

  - generate \( w_{d_i}^*, ~ i = 1, \ldots, 6 \)

  - then the candidate secondary times are:

\[ W_{M_i} + w_{d_{2^i-1}}^*, ~ i = 1, 2, 3 \]
\[ w_{S_i} + w_{S_d^*}, \quad i = 1, 2, 3 \]

- End repeat

- \( s_1 \leftarrow \) smallest feasible candidate secondary time

- \( s_2 \leftarrow \) second smallest feasible candidate secondary time

- \( w_{D_i} \leftarrow \) time from birth of \( s_i \) tumor until its discovery \((i = 1, 2)\)

- \( t_1 \leftarrow s_1 - w_{D} ; t_2 \leftarrow s_2 - s_1 \)

- \( y_0 = c e^{\alpha w_{D}} ; y_1 = c e^{\alpha w_{D_1}} ; y_2 = c e^{\alpha w_{D_2}} \)

- Return \((t_1, y_0, t_2, y_1, y_2)\)

This is the 5D model used throughout this section (Section 6.3). As promised in Section 6.2, some careful attention must now be given to the simulation of the non-homogeneous Poisson process. Ross [50, 1990, p. 76] explains how to simulate a n.h.p.p. with intensity function "\( \lambda(t) \)" (those details will not be repeated here). For the cancer model, "\( \lambda(t) = aY_0(t) = ace^{at} \) since the starting point for this model is the arrival of the primary. In the same manner as the Ross example, it is easy to show that the sequence \( w_{M_1}, w_{M_2}, \ldots \) may be simulated via

\[ w_{M_1} = \frac{1}{\alpha} \log \left( 1 - \frac{\alpha}{ac} \log(1 - u_1) \right) \]
\[ w_{M_2} = w_{M_1} + \frac{1}{\alpha} \log \left( 1 - \frac{\alpha}{ac} e^{-\frac{\alpha}{ac} w_{M_1}} \log(1 - u_2) \right) \]

\[ w_{M_3} = w_{M_2} + \frac{1}{\alpha} \log \left( 1 - \frac{\alpha}{ac} e^{-\frac{\alpha}{ac} w_{M_2}} \log(1 - u_3) \right) \]

\[ \vdots \]

where the \( u_i \)'s are simulations of independent uniform \((0,1)\) random variables. It is immediately evident how difficult the closed form for such highly nonlinear competing mechanisms must be.

Recall again two other important assumptions (which were also used for the 3D model in Section 6.2). First, all metastases are assumed to come from the primary; metastases are not allowed to throw off their own metastases. Second, once the primary is removed only the systemic mechanism can generate any additional secondary tumors.

5D Estimation and Results

The discussion here will proceed in a fashion similar to that of Section 6.2, as the results are similar. Refer now to Figure 6.11. For 5D, there are now 3 estimators from which to choose: \( \hat{\theta}_1, \hat{\theta}_3, \) and \( \hat{\theta}_5 \). The data used for 1D and 3D is the same as that used before,\(^8\) namely

\(^8\)Except for the fact that it is now generated from the more sophisticated model.
Figure 6.11
• 1D

  1. time from the discovery of the primary to the discovery of the first secondary tumor

• 3D

  1. time from the discovery of the primary to the discovery of the first secondary tumor

  2. volume of primary

  3. volume of first secondary tumor

• 5D

  1. time from the discovery of the primary to the discovery of the first secondary tumor

  2. volume of primary

  3. volume of first secondary tumor

  4. time from discovery of first secondary tumor until the discovery of the second secondary tumor

  5. volume of the second secondary tumor

In Figure 6.11, ̂θ_1, ̂θ_3, and ̂θ_5 are ranked relative to the minimax criterion
\begin{equation}
  h_{135}(\theta) := \max\{Q_1(\theta), Q_3(\theta), Q_5(\theta)\} .
\end{equation}

\(\hat{\theta}_5\) is the clear winner, and perhaps importantly (again) \(\hat{\theta}_3\) is generally the second runner up.

For 5D, the hybrid estimator \(\hat{\theta}_{135}\) was chosen in a similar manner to the manner which was used before. Since there are now three candidates \(\hat{\theta}_1, \hat{\theta}_3,\) and \(\hat{\theta}_5\), there are \(3^4 = 81\) possible combinations to consider in constructing the \(\hat{\theta}_{135}\) hybrid estimator. Letting \(\Xi\) denote the collection of all 81 of these candidates, the hybrid estimator is

\begin{equation}
  \hat{\theta}_{135} = \arg\min_{\theta \in \Xi} \{Q_1(\theta) + Q_3(\theta) + Q_5(\theta)\}
\end{equation}

where \(Q_5(\cdot)\) denotes Pearson's \(X^2\) based on the 5D ellipsoidal bins, \(Q_1(\cdot)\) and \(Q_3(\cdot)\) are the same as before.

In Figure 6.12, \(\hat{\theta}_1, \hat{\theta}_3, \hat{\theta}_5,\) and \(\hat{\theta}_{135}\) are compared with respect to the minimax criterion (6.7) (again, of course \(\hat{\theta}_{135}\) would win all the time if comparisons were being made with respect to the criterion at (6.3)). For \(k = 20\) and 30, the hybrid wins roughly 80% of the time, for \(k = 10\) roughly 70% of the time. Again the aggregate minimax criterion (6.4) seems to like \(\hat{\theta}_5\) second best, \(\hat{\theta}_3\) third best, and so on.

Figure 6.13 reflects which parameter estimates the hybrid seems to prefer. For \(\alpha\) and \(\lambda\) there seems a clear preference for the 1D estimates, while the hybrid seems to like 3D and 5D roughly the same (in second place). For \(\alpha\) and \(b\), there is a clear preference for the 3D estimate. For \(\alpha\), the hybrid seems to like the 5D estimate equally
well or better than the 1D estimate. For \( b \), the 5D estimate is never chosen. Based on this plot alone, it appears that the 5D data does not improve the estimation for \( b \), whereas much improvement was observed in the estimation of \( b \) when going from 1D data to 3D data.

Unlike in the 3D case, in the 5D case there is an additional hybrid to consider namely \( \hat{\theta}_{13} \). For clarity the \( \hat{\theta}_{13} \) to be discussed here is almost the same as the \( \hat{\theta}_{13} \) discussed in Section 6.2, the only difference being that the \( \hat{\theta}_{13} \) hybrid in the present section is constructed using the more sophisticated model. Another way to say this is that for the 5D case, there are now 5 competing estimators, \( \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_{13} \), and \( \hat{\theta}_{135} \). \( \hat{\theta}_{13} \) is constructed using only \( \hat{\theta}_1 \) and \( \hat{\theta}_3 \). There are \( 2^4 = 16 \) possible combinations for \( \hat{\theta}_{13} \) and (as in Section 6.2)

\[
\hat{\theta}_{13} := \arg\min_{\theta \in \Xi_2} \{Q_1(\theta) + Q_3(\theta)\}
\]

where \( \Xi_2 \) denotes the collection of the 16 candidate \( \hat{\theta} \)'s constructed from the permutations of \( \hat{\theta}_1 \) and \( \hat{\theta}_3 \).

In Figure 6.14 and 6.15, \( \hat{\theta}_1, \hat{\theta}_3, \) and \( \hat{\theta}_{13} \) are compared finding similar performance to that for \( \hat{\theta}_{13} \) in the previous section (see Figures 6.2 and 6.3). Figure 6.16 plots the logarithm (natural log) of the ratios of the MSE's comparing \( \hat{\theta}_1, \hat{\theta}_3, \) and \( \hat{\theta}_5 \). Notable in this plot is the fact that \( \hat{\theta}_3 \) and \( \hat{\theta}_5 \) perform roughly the same (except on \( \lambda \)); except perhaps that \( \hat{\theta}_5 \) seems to lose when there are too few bins (\( k = 10 \)). For this more sophisticated model, 5D and 3D beat 1D on \( a \) and \( b \) but seem to lose to 1D on \( a \) (and
Figure 6.14
Figure 6.15
Figure 6.16
of course on \( \lambda \)), which seems to duplicate similar performance of 3D vs 1D on the simpler model. One wonders if the asymptotics are kicking in on \( a \) for \( k = 30 \) and \( n = 600 \), but the boxplots (which are more detailed) in Figure 6.22 clearly indicate that it is the outlying estimates which hurt the 1D MSE (although one notes that there appear to be at least 7 or 8 of these outliers \( \approx 25\% \)).

Figure 6.17 compares \( \hat{\theta}_{13} \) and \( \hat{\theta}_{135} \) in terms of MSE. Notice the scale on the plot. With only 40 values to calculate the MSE, one has to conclude that these estimators are performing roughly the same with respect to this measure.

Figure 6.18 compares \( \hat{\theta}_{135} \) to \( \hat{\theta}_1, \hat{\theta}_3, \) and \( \hat{\theta}_5 \) in terms of MSE. Figure 6.19 compares \( \hat{\theta}_{13} \) to \( \hat{\theta}_1, \hat{\theta}_3, \) and \( \hat{\theta}_5 \) in terms of MSE. As noted about Figure 6.17, there is very little difference between \( \hat{\theta}_{13} \) and \( \hat{\theta}_{135} \) in terms of MSE, so the 2 plots in Figure 6.18 and 6.19 are somewhat redundant, but are included for completeness. In Figure 6.18, notice that \( \hat{\theta}_{135} \) either beats or ties \( \hat{\theta}_1, \hat{\theta}_3, \) and \( \hat{\theta}_5 \) in terms of MSE on \( a, b, \) and \( a \). \( \hat{\theta}_{135} \) beats \( \hat{\theta}_3 \) and \( \hat{\theta}_5 \) on \( \lambda \), but loses somewhat to 1D. While not perfect (due to \( \lambda \)), we again point out that this plot is somewhat remarkable in that the choice of \( \hat{\theta}_{135} \) is entirely data-driven, yet it’s performance in terms of MSE is quite good overall.

Figures 6.20-6.23 present detailed comparisons of \( \hat{\theta}_1, \hat{\theta}_3, \hat{\theta}_5, \hat{\theta}_{13} \) and \( \hat{\theta}_{135} \) via boxplots. It is clear from these plots that \( \hat{\theta}_{135} \) tends to match or even improve some upon \( \hat{\theta}_{13} \) in nearly all cases. Even where Figures 6.18 and 6.19 indicate some loss of \( \hat{\theta}_{135} \) to \( \hat{\theta}_{13} \) in the estimation of \( \lambda \), Figure 6.23 does not seem to bear that out dramatically. \( \hat{\theta}_{135} \) is still clearly competitive overall, although \( \hat{\theta}_{13} \) seems better for \( \lambda \) (but of course
Figure 6.17
Figure 6.19
Figure 6.21
Figure 6.22
\( \hat{\theta}_1 \) is better than everybody for \( \lambda \), but the data-based selection rule does not seem to be able to discern this).

Figure 6.24 compares \( \hat{\theta}_{13} \) and \( \hat{\theta}_{135} \) in more detail (note on this plot that 2.13 is an abbreviation for \( n = 200 \) and \( \hat{\theta}_{13} \), similarly 4.135 is an abbreviation for \( n = 400 \) and \( \hat{\theta}_{135} \)). Also overlaid in this figure are the starting values and the true parameter. Finally, detailed boxplots for \( \hat{\theta}_{135} \) are given in Figure 6.15, as under our data-based rule one would be required to take \( \hat{\theta}_{135} \) as the best estimator. The trends in Figure 6.25 match well with those in Figure 6.9.

6.4 Computational Section

**Ellipsoid Fitting Algorithm**

Recall the basic (and very simple) strategy which has been proposed (and theoretically justified) in this dissertation for multivariate SIMEST. Assume that one is given \( x_1, x_2, \ldots, x_n \) observations where each \( x_i \) is a vector in \( \mathbb{R}^d \). One also assumes that the \( x_i \)'s are realizations from some \( d \) dimensional continuous density in \( \mathbb{R}^d \), thus the main concern here is binning the continuous multivariate problem. With such \( x_i \)'s in hand, the proposed method proceeds as follows:

1. Standardize each dimension to mean = 0, standard deviation = 1.
Figure 6.24
Figure 6.25
2. Calculate some estimate of the center call it $\hat{\eta}$. For the models discussed in this dissertation, the simplest such $\hat{\eta}$ was used, namely the mean.\(^9\)

3. Set $i = 0$

- Repeat until $i = k - 1$
  
  - $i = i + 1$
  
  - choose the $m_i = i \cdot n/k$ nearest neighbors of $\hat{\eta}$
  
  - fit the ellipsoid of minimum volume containing these $m_i$ neighbors

- End repeat

4. This determines $k - 1$ ellipsoids, where each concentric “ring” contains roughly the proportion $1/k$ of the total number of observations.

And to a statistician, the computational steps required in such an algorithm are quickly discerned as being trivial, except possibly for fitting an ellipsoid of minimum volume to a given set of $m_i$ data points. It may perhaps be surprising, but even this

\(^9\)Note here that in general one would like $\hat{\eta}$ to be something more robust than just $\hat{x}$; a more robust estimator is given by the mean update algorithm of Fwu, Tapia, and Thompson [18, 1981] (also see Boswell [5, 1983] for related work on dimensions beyond 3). Such generality was not required for the models used in this dissertation.

Another note: for the consistency theory explicated in this dissertation, all one needs (to be rigorous, at least) is some consistent estimate for the mode (or even just some fixed point of the density). There have been some objections to the approach proposed in this dissertation on the grounds that perhaps there are more efficient binning strategies than the equal probability approach. In response the author points out that the consistency theory in this dissertation requires only a consistent mode estimator in order for the entire method to be completely rigorous, whereas any discussion about AMISE or similar measures of asymptotic efficiency (measures used in NDE) must generally make strong assumptions involving several derivatives of $f(x, \theta)$, assumptions not made in the theory of this dissertation.
task is a computationally trivial one. In Cook, Hawkins, and Weisberg [8, 1993], the authors investigate 2 algorithms for fitting a smallest ellipsoid to a given set of points. The task in their 1993 paper is to find the ellipsoid of minimum volume containing a certain percentage of the (multivariate) data. This is referred to as the Minimum Volume Ellipsoid (or MVE) and is a robust alternative to a covariance matrix as an estimator for the dispersion of the data. In the MVE problem, finding the ellipsoid of minimum volume containing a specific set of \( m \) data points is a subcomponent of the computationally intensive task of finding the (exact) MVE. Of the 2 algorithms investigated, the authors find that the Titterington [65, 1976] algorithm is far superior in terms of speed. This is the algorithm we chose to implement for multivariate SIMEST. It proceeds as follows:

1. Let \( \pi = (\pi_1, \pi_2, \ldots, \pi_m) \) be a vector of weights such that \( \sum_i \pi_i = 1 \) (on the first iteration one begins with the weights all equal)

2. Calculate the weighted mean and covariance based on the given choice of \( \pi \)

\[
\bar{x}_\pi = \sum_{i=1}^{m} \pi_i x_i \quad , \quad S_\pi = \sum_{i=1}^{m} \pi_i (x_i - \bar{x}_\pi)(x_i - \bar{x}_\pi)^T
\]

3. Calculate \( D_i(\pi) = (x_i - \bar{x}_\pi)^T S_\pi^{-1} (x_i - \bar{x}_\pi), i = 1, \ldots, m \)

4. Update the weights via

\[
\pi_i^{\text{iter}+1} = \frac{\pi_i^{\text{iter}} D_i(\pi_i^{\text{iter}})}{d}
\]
5. Continue until \( \max_i D_i(\pi^{iter}_i) < d + \epsilon \) where \( \epsilon \) is some (arbitrary) tolerance.

Then the ellipsoid found by this procedure is defined by

\[
(x - \bar{x}_\pi)^T S^{-1}_\pi (x - \bar{x}_\pi) = d + \epsilon
\]  

(6.8)

Technically, this is not exact, but could be made numerically exact by choosing \( \epsilon \) to be very small. In theory, the ellipsoid of (6.8) would have \( \epsilon = 0 \) (i.e. the right hand side of (6.8) would just be \( d \)). In practice a reasonable approximation is to choose \( \epsilon \in (0, 1) \). Cook, et. al. [8, 1993] used \( \epsilon = 0.01 \). In the work done for this dissertation values anywhere from \( \epsilon = 0.01 \) to \( \epsilon = 0.3 \) were used. The practical reason for using a larger \( \epsilon \) is that the algorithm then requires fewer iterations, making it much faster.

It is important to note that any time involved in fitting these ellipsoids is negligible in terms of the entire SIMEST procedure. This follows since one essentially trades in the data set for a set of ellipsoids. Then with these ellipsoids in hand, one embarks upon a rather long (comparatively) sojourn searching around \( \Theta \), using the ellipsoids only to count where simulations fall. By far the greatest computational effort and time is devoted to \((a)\) simulating and \((b)\) counting where the simulations fall.

For reference, on a 400 MHz Pentium II PC using FORTRAN code, fitting 30 or 40 ellipsoids (i.e. the whole set of ellipsoids needed for the method if \( k = 30 \) or \( k = 40 \)) to anywhere from 200 to 600 or so observations with a tolerance of \( \epsilon \approx 0.2 \) or 0.1 probably takes about 10-15 seconds. Notably this time probably doubles or triples if
\( \varepsilon \) is cut in half. Also the fitting tends to bog down in a hurry for larger sample sizes.

The author was able to determine that \( \varepsilon \leq 0.2 \) or so for \( n \leq 600 \) seems to work just fine; however, the importance of the choice of \( \varepsilon \) for larger \( n \) was never determined.

A conjecture based on the work done for this dissertation is that it does not matter a great deal; the critical choices of parameters seem to be \( n, k \), and the simulation size(s), most other parameters (even those involved in the search routine\(^{10}\)) are of much less significance.

The slowness of the fitting algorithm for large \( n \) and \( k \) contributed to - but was only partly responsible for - the inability of the author to do much work in the way of many bins (e.g. in Thompson, Atkinson, and Brown [63, 1987], a quick theoretical treatment - on pages 393-402 - suggests that the more bins the better).

Interestingly, the difference in computing speed going from 3D to 5D was not enormous. It was approximately double the time or a little less. This contrasts with the author's experience in simulations for the multivariate studentized range (Schwalb and Thompson [55, 1998]) where increasing the dimension by only one led to enormous increases in computing time.

**Polygonal Bins - an Idea for Increased Speed**

As noted above, the far majority of time doing SIMEST is spent in \((a)\) simulating the observations and \((b)\) counting where they fall. The time required (the time needed to

\(^{10}\)More discussion on the search routine follows below.
solve the appropriate system of equations, which is of dimension $d \times B$, where $B$ is the number of simulations used) in computing where the simulations fall is probably about half of the computing time. It seemed likely that there might be some faster way to do this. The author thought of the following idea which seemed promising, but took the same amount of time as the ellipsoidal bins. It is recorded here in the event it may prove useful elsewhere in the future.

**The Polygonal Bins Idea**

With the ellipsoids determined by our nearest neighbor approach in hand, it seems natural to consider either (i) some inscribed polygon or (ii) some polygon circumscribing each ellipsoid. See Figure 6.26. The clear advantage of the polygon is that instead of needing to solve a system of equations, one merely needs to check a few inequalities to determine where a given observation falls. Still further there are savings to be had in that generally one will determine if an observation is not in a particular polygon before one has to check it against every side of the polygon. One knows which side of the polygon is “out” (of the bin), and one may stop checking as soon as the observation is determined to be out. The neat thing about this approach is that - like the ellipsoids - it is also algebraic, extending straightforward to dimensions greater than three. Some of the interesting details involved in these calculations are now described.
2D ellipse, 2 axes, 4 vertices
4-sided inscribed polygon

3D ellipsoid, 3 axes, 6 vertices
8 sided inscribed polygon
(front panel of polygon is cut away)

Figure 6.26

So how does one extend this polygon idea to high dimensions? Notice that any set of $d$ vertices each lying on a different axis determines a side of the inscribed polygon. So immediately one sees that there are $2^d$ sides for each inscribed polygon. To determine whether an observation falls in a given polygon then, one has (at most) $2^d$ inequalities to check. Each set of $d$ vertices (each lying on different axes) determines a hyperplane in $\mathbb{R}^d$. One knows the “correct” side of each hyperplane where the observation must fall to be in the current bin, since one knows that $\mu$ (where $\mu$ denotes the ellipsoid center) falls on the correct side of all the $2^d$ defining hyperplanes.

Skipping the (straightforward) justification, the algorithm for determining these $2^d$ hyperplanes for a given ellipsoid is as follows:
1. Assume one is given $\mu$ and $\Sigma$ for the ellipsoid defined by $(x-\mu)^T \Sigma^{-1} (x-\mu) = c^2$.

2. Calculate the eigenstructure of $\Sigma$, call it $(\alpha_j, e_j)$, $j = 1, \ldots, d$, making sure that the $e_j$'s are orthonormal. Here the $\alpha_j$'s denote the eigenvalues and the $e_j$'s denote the eigenvectors of $\Sigma$.

3. The $2d$ vertices of this ellipsoid are

$$v_{ij} = (-1)^i \frac{c}{\sqrt{\lambda_j}} e_j + \mu, \; i = 1, 2, \; j = 1, \ldots, d$$

where $\lambda_j := 1/\alpha_j$.

4. Suppressing the index $i$ (only choosing a set of $d$ vertices which do not lie on the same axis), let $v_1, \ldots, v_d$ denote the set of vertices for a given hyperplane (i.e. $d$ vertices lying on non-adjacent axes). Calculate

$$\alpha_1 = v_1 - v_2, \; \alpha_2 = v_2 - v_3, \; \ldots, \; \alpha_{d-1} = v_{d-1} - v_d$$

Find $\gamma$ orthogonal to the hyperplane spanned by $\alpha_1, \ldots, \alpha_{d-1}$ by the calculation

$$\gamma = A_*^{-1} b$$

where

$$A_* = \begin{pmatrix}
\alpha_1^T \\
\alpha_2^T \\
\vdots \\
\alpha_{d-1}^T \\
1^T
\end{pmatrix}$$

and

$$b_* = \begin{pmatrix}
0 \\
\vdots \\
0 \\
1
\end{pmatrix}$$
where $1^T$ denotes a row vector of 1's and there are $d - 1$ 0's in the $b_*$ vector. This determines $\gamma$ (and $\gamma$ satisfies $\sum \gamma_i = 1$ but length of $\gamma$ doesn’t matter; a length is specified here only because otherwise the system is underdetermined).

5. The equation $\gamma^T(x - v_1) = 0$ defines this plane. Use the calculation $\gamma^T(\mu - v_1)$ to determine the "correct" side of this plane.

Summary of Failure of Polygonal Bins

As mentioned, the only point in attempting the concentric polygonal bins was in order to speed things up. As it turned out the polygonal bins did not even match the ellipsoidal bins in terms of speed although the speeds were very close (the ellipsoidal bins were about 5% faster). It was expected that the speed of the polygonal bins might at least double the speed of the ellipsoidal bins, but it did not which was disappointing.

Nelder Mead and Miscellaneous Computational Details

While the big strength of SIMEST is its ability to handle models of extreme complexity, one of its biggest downfalls is the numerical problem of trying to find the global maximum in the absence of any derivative information, and even (as pointed out already) more or less in the absence of any sensitivity information. This aspect of the SIMEST problem is fundamentally difficult and deserves more attention in the
future. In this dissertation, this aspect of the SIMEST problem is not dealt with at all but is treated as a necessary evil, making the best out of the situation as it stands.

Following the suggestion of Thompson, et al, ([63, 1987]), the author chose Nelder Mead for the numerical search procedure. The code for this algorithm was found at:

http://www.stat.cmu.edu/apstat/47

There are 2 versions of code there, we used the second one (coded by D.E. Shaw).

One of the main difficulties found in preliminary work with the algorithm is that relatively speaking SIMEST is still rather slow. In real life, one will have only one data set to deal with. In such a situation, one would have ample time and opportunity to "fiddle around" (for lack of a better phrase) with the search routine in an ad-hoc manner. Based on the experience of the author in this dissertation, such fiddling around is likely to be both timely and successful. For example, with the models considered in this chapter the only problem the search algorithm seemed to exhibit was premature stalling. In situations where the algorithm has stalled, one can always easily tell it has stalled by the magnitude of $X^2$. Never in the models considered in this chapter (and all the data sets which were considered) was a good $X^2$ value obtained for parameter values far from the truth. Unlike an arbitrary optimization problem, in this statistical situation one has a pretty good rough idea of what value the function should be near its minimum (i.e. roughly $k - s - 1$ where $k$ is the number of bins and $s$ is the number of parameters estimated).
From a one data-set-only standpoint, stalling is not such a difficult problem to overcome. The user can restart the search as many times as he or she cares to. In this dissertation, however, the situation is such that it is desirable to consider the performance of SIMEST on many data sets. Hence it was not considered timely in the work done here to supervise every run of Nelder Mead. It was necessary that the search routine be set up in such a way that it could run unsupervised and still be relatively successful.

This situation of needing the search routine to run unsupervised happens to make it difficult to start the search very far from the true parameters. It was originally hoped that we might start as far away as a factor of 5 or a factor of 10 from the truth, but this hope was later reduced to a factor of 3. Finally, the SIMEST results shown in this chapter used a factor of 2 for \( a \) and \( b \) and a factor of 4 for \( a \) and \( \lambda \). As justification for starting this close, hopefully the following 2 arguments are satisfactory:

1. It is based on the need to let the search run many times over and over unsupervised.

2. Two real issues to be addressed in SIMEST can legitimately be separated, they are

   (a) Finding the global minimum (search problems), and

   (b) Investigating the behavior of the parameter estimates once they are found.
If one agrees to not deal with the practical issue of 2 (a) here, then starting close does not detract very much at all from the investigation of 2 (b). Also, for example, even if one starts close, there is nothing to keep the search routine from wandering off in the wrong direction.\textsuperscript{11} More could be said here, but the problem of global optimization is to be left for another time. It is admitted that this problem is not dealt with here.

It was discovered in running the Nelder Mead search routine that a single run with many iterations and a large number of simulations per $X^2(\theta)$ evaluation was just not very effective. Typically, the algorithm would stall after about 75 iterations unless it happened to be well along the way to the minimum. A more effective approach seemed to be to repeat the search say 4 to 5 times, using the ending values of the previous search as the starting values for the next search. In more detail, this algorithm is described as follows

- $\theta_{in} \leftarrow \theta_{\text{start}}$

- $\text{step}_{in} \leftarrow \text{step}_{\text{start}}$

- $\text{pass} \leftarrow 1$ (i.e. this is the first pass)

- $B \leftarrow 2 \cdot n$ ; $\text{maxit} \leftarrow 130$

- $(\theta_{out}, \text{step}_{out}) \leftarrow \text{search}(\theta_{in}, \text{step}_{in}, B, \text{maxit})$

- Repeat until $\text{pass} = 5$

\textsuperscript{11}In the modeling results presented in this chapter, the search routine was free to wander away as far as a factor of 10 if it wished.
- \( pass \leftarrow pass + 1 \) (i.e. 2nd pass through 5th pass)

- \( B \leftarrow 2 \cdot pass \cdot n \) (or some increase in simulation size with each pass)

- \((\theta_{in}, step_{in}) \leftarrow (\theta_{out}, step_{out})\)

- \((\theta_{out}, step_{out}) \leftarrow search(\theta_{in}, step_{in}, B, maxit)\)

- End repeat

- \( \hat{\theta} \leftarrow \theta_{out} \)

- Return \( \hat{\theta} \)

Here the notation \( search(\theta_{in}, step_{in}, B, maxit) \) is used to indicate a call to the search routine. The search routine is passed \( \theta_{in} \) (a starting value for \( \theta \)), \( step_{in} \) (step sizes), \( B \) (the number of simulations to use per function evaluation), and \( maxit \) (search terminates and returns best \( \theta \) so far if the number of iterations exceeds \( maxit \)). A call to the search routine returns \( \theta_{out} \) (the best value of \( \theta \) found during the search) and \( step_{out} \) (the step sizes at the end of the search). The assignment \((\theta_{out}, step_{out}) \leftarrow search(\theta_{in}, step_{in}, B, maxit)\) indicates a call to \( search \) where \( \theta_{in}, step_{in}, B, \) and \( maxit \) are given to \( search \), and \( \theta_{out} \) and \( step_{out} \) are returned from \( search \) on termination. The rest of the steps in this algorithm should be clear.

Notice that with each repetition the number of simulations is increased. When far away from the minimum very few simulations per evaluation are required for the search algorithm to find its way to a more promising region of \( \Theta \), and it is only when
the algorithm is relatively close to the minimum that lots of simulations seem to
improve the parameter estimates.

Using this approach it was found that the Nelder Mead code used has the following
drawback. As written, the routine always returns to the user the value of \( \hat{\theta} \) where it
stops, or where it believes to have converged. But never in all of the work done for
this dissertation was it found that the algorithm was near the correct value of \( \theta \) when
it believed it had converged. A modification which should be made to this code to
have it run more effectively would be to remove the part which tests for convergence
as it just does not seem to work (at least not for SIMEST). The author (of this
dissertation) never made this particular modification.

A modification the author (of this dissertation) did make\(^{12}\) was to have the al-
gorithm keep track of the \( \hat{\theta} \) where it had the smallest \( X^2 \) and to be sure to return
that \( \hat{\theta} \). It may seem surprising, but, for example, one could watch the algorithm hit
\( X^2 = 20 \) at iteration number 50, then march around to \( X^2 = 75 \) or so, exceed the
iteration limit at 130 iterations and instead of returning the \( \hat{\theta} \) which had \( X^2 = 20 \),
it would return the \( \hat{\theta} \) where it was when it stopped (say at \( X^2 = 75 \)). It is hard
to believe that a minimization routine would be coded in such a Markovian fashion,
but that is how this routine was coded. The aforementioned modification resulted in
a substantial improvement in performance, probably one of the most effective single
adjustments in the SIMEST code that the author (of this dissertation) made.

\(^{12}\)This modification is already reflected in the algorithm described above.
Some problems which had to be dealt with were (a) zero denominator for $X^2$ and (b) non-negativity requirements on the parameters. Again, the Nelder Mead code used does not provide the capability of placing even simple bounds on the parameters, and the author was unable to find a Nelder Mead code which does. This problem was overcome by assigning $X^2$ some huge value whenever the search routine stepped any of the parameters into $(-\infty, 0)$. For (a) whenever $np^2(\theta)$ - the denominator of $X^2(\theta)$ - was 0, it would be replaced with .001; assigning a stiff penalty for choosing a $\theta$ for which one of the probability $\approx 1/k$ bins collects no simulations. The implementation checks for a zero denominator for each component of $X^2$ rather than just adding .001 to each $np^2(\theta)$ indiscriminantly. Thus this adjustment is typically only made while searching farther away from the minimum.

The following strategy was implemented in the preliminary work for this chapter as one idea of making the search more global (but note that this implementation is not used in the simulation results presented in this chapter - as for the results presented in this chapter the starts were rather close), recording it here due to the fact that it was an idea which seemed quite successful and promising. Let the user specify some large interval within which $\theta_0$ (the true parameter) is believed to lie. This large interval could be something like a factor of 10 to 20 above and below $\theta_0$. Generate a starting value uniformly in each of the given intervals, call the resulting $\theta$, $\theta_1^{(a)}$. If
fewer than say $p\%$ or so\textsuperscript{13} of the bins collect no simulations under this starting value, reject it and generate another one, $\theta_2^{(s)}$. Continue in this fashion until one obtains a start for which at least $p\%$ of the bins collect some simulations. Give this starting value to Nelder Mead. Typically such an approach is very successful. Generally the global search performs excellently if started with a healthy percentage of the bins collecting simulations.

An important practical implication of the theoretical work in this dissertation (thanks to the empirical process theory) is that one may keep the same seed when going from one $\theta$ to another in the search algorithm. This helps since, for example, it makes $X^2(\theta)$ a bona-fide function, i.e. each $\theta$ value returns only one $X^2(\theta)$ value, whereas if the seed could not be reset then one would keep getting different $X^2$ values for the same choice of $\theta$. Undoubtedly, resetting the seed is improving the performance of the search algorithm.

Starting values were chosen either above or below by the appropriate factor according to a fair coin flip, so roughly half the time the starting value for a particular parameter will be high, the other half of the time it will be low (a factor of 2 was used for $\alpha$ and $b$; a factor of 4 was used for $a$ and $\lambda$). Starting step sizes were chosen to be equal to the starting value (as perhaps opposed to (starting value)/2 or (starting value)/3). Only two passes of the search routine were used (for the results

\textsuperscript{13}This percentage is a parameter one can set, i.e. it is arbitrary. In the experimentation done for this dissertation, any percentage $\geq 50\%$ seemed to work well.
in this chapter), and each pass was limited to 130 iterations. At the end of the second pass, the value of \( \hat{\theta} \) returned by the search routine (which is in essence the \( \hat{\theta} \) which led to the smallest value of \( X^2 \) before the routine is stopped at 130) was taken as the simulation-based parameter estimate of \( \theta_0 \). The simulation size for the first pass was \( 3 \cdot n \), where \( n \) is the sample size, and for the second pass was \( 6 \cdot n \). It is relatively easy to determine the point of diminishing returns when increasing the simulation size, the number of iterations, and/or the number of passes. Also, (perhaps importantly) all of the 1D, 3D, and 5D results used this same setup (i.e. same simulation sizes, starting values, maximum number of iterations, etc.). Certainly by any reasoning this "level playing field" across the dimensions must confer an unfair advantage to the parameter estimation based on 1D data. It is believed in fact that more dramatic results might have been (or still may be) obtained for the 3D and 5D estimation with either more bins or with an interesting overlapping bins idea proposed by the author's advisor (Thompson), which will be discussed elsewhere. At any rate, the main point here is that the 1D-5D results discussed are all on a level playing field.

**Computing Speeds**

The computing in this dissertation was done using FORTRAN on a Dell PC, 400MHz Pentium II machine (at the moment this is the fastest machine in the department).

For the simpler model of Section 6.2, the running times were

- 1D - roughly 1/2 to 1 hour
• 3D - about 2.5 hours

For the more complicated model in Section 6.3, the running times were

• 1D - 1.5 hours
• 3D - 3 hours
• 5D - 4 hours

For example, 1D SIMEST was performed for \( n = 200, 400, 600 \) and \( k = 10, 20, 30 \) (9 different SIMEST runs) and 40 (simulated) data sets which implies \( 9 \times 40 = 360 \) total SIMEST runs (including in this the time required to generate the simulated data sets and fit the ellipsoids) in about 1.5 hours for the more complicated model.

Interestingly, the calculations needed to obtain the hybrid estimators are fairly intensive. Doing these calculations for \( \hat{\theta}_{135} \), the hybrid estimator constructed from \( \hat{\theta}_1, \hat{\theta}_3, \) and \( \hat{\theta}_5 \), took

• 1D - 1 hour (on a 150MHz machine)
• 3D - 1 hour (on 400MHz)
• 5D - 1.5 hours (on 400MHz)

which is a good proportion of the time required to do the entire estimation.

The code used for the random number generators is in the “randlib.f” library and is available at:
http://odin.mdacc.tmc.edu/anonftp/source.html.

In this dissertation, version 1.3 (the most recent version) was used.

6.5 A Future Direction of Research

As the final note of this chapter, we present an idea proposed by the author's advisor which attempts to improve the performance of our ideas by using a more local binning strategy. The idea proceeds as follows.

Choose some value of \( m \) corresponding to a certain percentage of the data (e.g. \( m = 10 \) or \( 20 \) for \( n = 200 \)). At each data point determine the sphere (or ellipsoid if computing speed is fast enough) of minimum volume centered at that data point which contains \( m \) observations. Construct such a sphere for each data point. Observe that this collection of \( n \) spheres will overlap and that the amount of overlapping should be substantial.\(^{14}\) Do SIMEST using these \( n \) overlapping bins. This approach is both exciting and promising. In particular, one expects to improve small sample performance with such a localized approach. At the present time, the author has been unable to adapt his consistency proofs to work for such an approach, but as the reader should undoubtedly see, the necessary theoretical machinery to do so certainly seems to be the same machinery which has been used in this dissertation. Hopefully we will be able to make such a generalization of the theory very soon.

\(^{14}\) Anyone familiar with the goodness of fit literature will know that overlapping multivariate bins has not been done at all.
Bibliography


Appendix A

Background on Banach-valued Strong Laws

A.1 Measurability and Banach-valued Random Variables

We start with a map $Y : \Omega \to \Xi$, where $\Omega$ is some set and $\Xi$ is a Banach space. As usual we associate a sigma algebra of subsets, say $\mathcal{F}$, with $\Omega$. We also associate a sigma algebra of subsets, say $\mathcal{M}$ with $\Xi$. With real-valued random variables, the concept of measurability is unambiguous; but with $\Xi$-valued random variables there are several meaningful ways to define measurability. Some of these follow (most of these definitions are taken from Hille and Phillips [23, 1957]; we will abbreviate Hille and Phillips to HP).

**Definition A.1** A mapping $Y : \Omega \to \Xi$ is Borel measurable if $[Y \in M] \in \mathcal{F}$ for each Borel set $M \in \mathcal{M}$.

**Definition A.2** (HP [23], p. 72) A mapping $Y : \Omega \to \Xi$ is strongly measurable if there is a sequence of countably-valued functions $\{Y_n\}$ with $\lim_n \|Y_n - Y\| = 0$ a.e.

---

1Such a $Y$ is called a generalized random variable by Hanš [21, 1956], although there seems to be no standard usage.
Definition A.3 (HP [23], p. 72) A mapping \( Y : \Omega \to \Xi \) is weakly measurable if for all \( f \in \Xi^* \) (where \( \Xi^* \) denotes the dual of \( \Xi \)), \( f(Y) \) is measurable (in the scalar sense), i.e. \([\omega : f(Y(\omega)) \in B] \in \mathcal{F} \) for all \( B \in \mathcal{B} \) (\( \mathcal{B} \) denotes the Borel sets on \( \mathbb{R} \)).

For the case of finite measures, we may replace "countably-valued" in Definition A.2 by "simple" or "finite-valued" (cf. HP [23, p. 72]). We demonstrate this later for the special case where \( \Xi \) is separable (see Theorems A.1 and A.2), as it is somewhat instructive.

We are interested only in the case where \( \Xi \) is a separable Banach space (indeed, much of the usual behavior of random variables we are accustomed to in the real case carries over to the \( \Xi \)-valued case so long as \( \Xi \) is a separable metric space; see e.g. Dudley [14, 1989]). In that case all reasonable definitions of measurability coincide. Restricted to the case where \( \Xi \) is separable, the following definition is appropriate.

Definition A.4 Suppose \( \Xi \) is separable. A Borel measurable mapping \( Y : \Omega \to \Xi \) is an \( \Xi \)-valued random variable (or a random variable).

A.2 Continuity of Mappings and Measurability

In constructing \( \mathbb{B} \)-valued random variables, it is generally quite easy and natural to compose two mappings like

\[
X : \Omega \to \mathbb{R}^d, \quad g : \mathbb{R}^d \to \Xi
\]  

(A.1)
to arrive at a mapping \( g(X) : \Omega \rightarrow \Xi \) which is some mapping of interest. A case of great interest (to us) is when \( \Xi = C(\Gamma) \), the space of continuous functions on a compact set \( \Gamma \). We know that \( C(\Gamma) \) is a separable Banach space.

Although one can quickly discern the necessary component mappings for the composition (as expressed in (A.1)), determining measurability of the composition is not so obvious. Usually we start with the assumption that \( X \) is a measurable map. Next, we ask: what does it mean to say that the mapping \( g \) is measurable? Concentrating on the case where \( \Xi = C(\Gamma) \) we may make some headway as follows.

Since \( C(\Gamma) \) is a metric space, the closed sets in \( C(\Gamma) \) generate the Borel sigma field \( \mathcal{M} \). Equivalently, we may generate \( \mathcal{M} \) with the open sets in \( C(\Gamma) \). Denote by \( \mathcal{O} \) the open sets of \( C(\Gamma) \). Since

\[
g^{-1}(\mathcal{M}) = \sigma(g^{-1}(\mathcal{O}))
\]

it is enough to check that

\[
g^{-1}(\mathcal{O}) \subseteq \mathcal{B}^d \tag{A.2}
\]

where \( \mathcal{B}^d \) denotes the Borel sets on \( \mathbb{R}^d \) as usual. Rewrite (A.2) as

\[
[x : g(x) \in O] \in \mathcal{B}^d , \quad O \in \mathcal{O} .
\]

Since \( \Xi \) is separable, we also know (see for example the details of the proof of Proposition 6, Royden [52, 1988, p. 142]) that each open set is expressible as a
countable union of open balls, say \( \{O_i\}_{i=1}^{\infty} \), (moreover, the centers of these balls are members of the separating set), i.e. for each \( i = 1, 2, \ldots \)

\[
O_i = \{ y : \|y - y_i\| < \delta \}
\]

for some rational \( \delta \) and some \( y_i \) in the separating set. Thus

\[
[x : g(x) \in O] = [x : g(x) \in \bigcup_i O_i] = \bigcup_i [x : g(x) \in O_i]
\]

\[
= \bigcup_i [x : g(x) \in \{ y : \|y - y_i\| < \delta \}]
\]

\[
= \bigcup_i [x : \|g(x) - y_i\| < \delta]
\]

And it is intuitively obvious that so long as \( g(\cdot) \) is reasonably well behaved (in some sense), there is no reason why \( [x : \|g(x) - y_i\| < \delta] \) should fail to be a \( \mathcal{B}^d \) set. In fact, all we must require of \( g(x, \theta) \) is that (a) for each \( x \), \( g(x, \theta) \) is continuous on \( \Gamma \) and (b) for each \( \theta \), \( g(\cdot, \theta) \) is measurable \( \mathcal{B}^d/\mathcal{B} \). To see this, suppose that for each choice of \( \theta \), \( g(\cdot, \theta) \) is measurable \( \mathcal{B}^d/\mathcal{B} \). Let \( \theta_1, \theta_2, \ldots \) be an enumeration of the rationals in \( \Gamma \). Let \( f_n(x) = |g(x, \theta_n) - y_i| \) and observe that each \( f_n \) is measurable \( \mathcal{B}^d/\mathcal{B} \). Let \( f(x) = \sup_{\Gamma'} |g(x, \theta) - y_i| \) where \( \Gamma' = \Gamma \cap \mathcal{Q} \) (the symbol \( \mathcal{Q} \) is used to denote the rationals). Then \( \sup_n f_n(x) = f(x) \), so that \( f \) is measurable \( \mathcal{B}^d/\mathcal{B} \). If we suppose that for each \( x \), \( g(x, \theta) \) is a continuous function of \( \theta \), then

\[
\sup_{\Gamma'} |g(x, \theta) - y_i| = \sup_{\Gamma} |g(x, \theta) - y_i|
\]
because $|g(x, \theta) - y_i|$ is then continuous (recall that the $y_i$'s are members of the separating set of $C(\Gamma)$, hence are polynomials). Hence $[x : f(x) < \delta]$ is a measurable set. But

$$[x : f(x) < \delta] = [x : \sup_{\Gamma} |g(x, \theta) - y_i| < \delta] = [x : ||g(x) - y_i|| < \delta].$$

Thus $g : \mathbb{R}^d \to \Xi$ is Borel measurable. We summarize this discussion in the following result.

**Result A.1** Suppose that $g(x, \theta) : \mathbb{R}^d \times \Theta \to \mathbb{R}$, where $\dim(\Theta) = s$. If (a) for each $x$, $g(x, \cdot)$ is a continuous function of $\theta$ and (b) for each $\theta$, $g(\cdot, \theta)$ is a (Borel) measurable function of $x$, then the mapping $g : \mathbb{R}^d \to \Xi = C(\Gamma)$ is measurable.

For the sake of avoiding the topic of measurability, it may perhaps be of interest to have a calculus-type condition on $g(x, \theta)$ which is sufficient to make $g : \mathbb{R}^d \to \Xi$ Borel measurable. Of course, continuity of the map $g$ is sufficient for Borel measurability. But here continuity means that for $x_1$ and $x_2$ close in $\mathbb{R}^d$, $g(x_1)$ and $g(x_2)$ will be close in sup norm over $\Gamma$. Such a statement is a little unnatural from a calculus-type perspective. This is easily remedied, however. Consider the following result.

**Result A.2** Suppose $g^*(x, \theta) : \mathbb{R}^d \times \Theta \to \mathbb{R}$ where $\dim(\Theta) = s$. Suppose further that for any compact region $I$ in $\mathbb{R}^d$ that $g^*(x, \theta)$ satisfies a Lipschitz condition

$$|g^*(x_1, \theta_1) - g^*(x_2, \theta_2)| \leq K_1 \|(x_1, \theta_1) - (x_2, \theta_2)\| \quad \text{on } I \times \Gamma \quad (A.3)$$
where $\Gamma$ is a compact region of $\Theta$. On $\Gamma$, define $g(\cdot) := g^*(\cdot, \theta)$. Then $g$ is a continuous map from $\mathbb{R}^d$ to $\Xi = C(\Gamma)$.

**Proof.** By problem 8.6 in Ruckle [53, 1991] or 7.16 in Royden [52, 1988], if $M_1$ and $M_2$ are metric spaces and $h : M_1 \to M_2$, then $h$ is continuous at $m_0$ if and only if for each sequence $\{m_n\}$ converging to $m_0$, $\{h(m_n)\}$ converges to $h(m_0)$ in $M_2$.

So pick $x_0 \in \mathbb{R}^d$. Let $x_n \to x_0$. Pick any $I$ such that $x_0$ is properly contained in $I$. Then $\exists n_0$ such that $x_n \in I$ for $n \geq n_0$. Then on $I \times \Gamma$ and for $n \geq n_0$,

$$|g^*(x_n, \theta) - g^*(x_0, \theta)| \leq K_I|| (x_n, \theta) - (x_0, \theta)|| = K_I||x_n - x_0||$$

which implies that

$$\sup_{\theta \in \Gamma} |g^*(x_n, \theta) - g^*(x_0, \theta)| \leq K_I||x_n - x_0||$$

$$\iff \sup_{\theta \in \Gamma} |g(x_n) - g(x_0)| \leq K_I||x_n - x_0|| \to 0$$

$$\iff ||g(x_n) - g(x_0)|| \to 0 \quad \text{as } n \to \infty.$$

So $g(x_n) \to g(x_0)$ in $C(\Gamma)$. Thus $g(x)$ is continuous at $x_0$. But $x_0$ was arbitrary.

\[\square\]

**Result A.3** Suppose $g^*(x, \theta)$ and $g(x)$ are as in the previous theorem except that in lieu of the Lipschitz condition we assume that $\nabla g^*(y, \theta)$ exists and is bounded on $\Gamma \times I$
for any choice of a compact $I$ (and $\Gamma$ some fixed compact set). Then $g$ is a continuous map from $\mathbb{R}^d$ to $\Xi = C(\Gamma)$.

Proof. The existence of the bound on the gradient ensures that $g^*(x, \theta)$ satisfies (A.3). Apply Result A.2.

We remark that the assumptions about $g(x, \theta)$ in Result A.1 are extremely flexible. Still, even the assumptions about $g(x, \theta)$ in Results A.2 and A.3 are rather flexible, particularly because they need only hold over compact regions.

A.3 Strong Measurability When $\Xi$ is Separable

Theorem A.1 (Hanš [21], 1956) Suppose that $\Xi$ is a separable Banach space.\(^2\) The mapping $Y : \Omega \to \Xi$ is Borel measurable if and only if it is strongly measurable.\(^3\)

Proof. This proof can be found in Hanš [21], but it will be instructive for us to work through the sufficiency part of the proof in detail. We will not work through the necessity. Pick $\epsilon_n = n^{-1}$. Let $D = \{y_1, y_2, \ldots\}$ be a countable dense subset of $\Xi$. Let

$$M_i = \{y : \|y - y_i\| \leq \epsilon_n\} - \bigcup_{k=1}^{i-1} M_k.$$

We claim that the $M_i$'s partition $\Xi$. First observe that each $M_i \subseteq \Xi$ and $M_i \in \mathcal{M}$, so $\cup_i M_i \subseteq \Xi$. We need to show that $\Xi \subseteq \cup_i M_i$.

\(^2\)Hanš uses "metric space" instead of Banach space.

\(^3\)Note for reference that the $Y_n$'s are expressible as $\mathcal{M}$ measurable functions of $Y$.\n
Clearly $D \subseteq \bigcup_{k} M_{k}$. Let $D'$ denote the limit points of $D$. Let $y \in D'$. Then the ball of radius $\epsilon_{n}$ around $y$, $B(y, \epsilon_{n})$, must contain one of the $y_{i}$'s, call it $y_{io}$ (definition of a limit point). But then

$$y \in B(y_{io}, \epsilon_{n}) = \{ y : \| y - y_{io} \| \leq \epsilon_{n} \} \subseteq \bigcup_{i \leq i_{0}} M_{i} \subseteq \bigcup M_{i} .$$

Since $\Xi = D \cup D'$ (by the density of $D$), it follows that $\Xi \subseteq \bigcup_{i} M_{i}$. Thus $\Xi = \bigcup_{i} M_{i}$.

By construction, the $M_{i}$'s are disjoint. Thus the $M_{i}$'s partition $\Xi$ as claimed.

On $M_{i}$, $\| Y - y_{i} \| \leq \epsilon_{n}$ a.s. Define $Y_{n} := \sum_{i=1}^{\infty} y_{i} I_{Y \in M_{i}}$. $Y_{n}$ is countably-valued, Borel measurable, and $\| Y_{n} - Y \| \leq \epsilon_{n} = n^{-1}$ a.s. Thus $\lim_{n} \| Y_{n} - Y \| = 0$ a.s. \(\Box\)

**Theorem A.2** (Hans [21], 1956) Suppose that $\Xi$ is a separable Banach space. The mapping $Y : \Omega \to \Xi$ is Borel measurable if and only if there is a sequence of simple functions $Y_{n}$ such that $\lim_{n} \| Y_{n} - Y \| = 0$ a.s.\(^{4}\)

**Proof.** Again we prove only sufficiency. Even though this direct proof is instructive, note that strong measurability is equivalent to the existence of a sequence of simple functions $Y_{n}$ such that $\lim_{n} \| Y_{n} - Y \| = 0$ a.s. (cf. Hille and Phillips [23, p. 72]) for finite measures, hence Theorem A.2 is equivalent to Theorem A.1.

Let $\{ y_{1}, y_{2}, \ldots \}$ be a countable dense subset of $X$. For each $n = 1, 2, \ldots$ and $i = 1, 2, \ldots, n$ let

$$M_{11} = \Xi , \quad M_{nn} = \{ y : \| y - y_{n} \| < \min_{1 \leq k \leq n-1} \| y - y_{k} \| \}$$

\(^{4}\)Note for reference that the $Y_{n}$'s are expressible as $M$ measurable functions of $Y$.\[\]
and

\[ M_{in} = M_{in-1} \cap (\Xi - M_{nn}) \]

Although this construction is clear, we save the reader a little trouble by writing out the first few terms:

\( n = 2: \)

\[ M_{22} = \{ y : \| y - y_2 \| < \| y - y_1 \| \} : \text{points closer to } y_2 \text{ than to } y_1, \]

\[ M_{12} = M_{11} \cap (\Xi \cap M_{22}) = \Xi - M_{22} = M_{22}^c : \text{points closer to } y_1 \text{ than to } y_2. \]

\( n = 3: \)

\[ M_{33} = \{ y : \| y - y_3 \| < \min_{1 \leq k \leq 2} \| y - y_k \| \} : \text{closer to } y_3 \text{ than to } y_1 \text{ or } y_2, \]

\[ M_{13} = M_{12} \cap M_{33}^c : \text{closer to } y_1 \text{ than to } y_2 \text{ and closer to } y_1 \text{ or } y_2 \text{ than to } y_3, \]

\[ M_{23} = M_{22} \cap M_{33}^c = : \text{closer to } y_2 \text{ than to } y_1 \text{ and closer to } y_1 \text{ or } y_2 \text{ than to } y_3. \]

So that \( M_{in}, i = 1, \ldots, n \) is a partition of \( \Xi \) into \( n \) regions where points in \( M_{in} \) are closer to \( y_i \) than to any of the remaining points \( \{ y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n \} \). Now for each \( n \), let \( Y_n = \sum_{i=1}^{n} y_i I_{\{ y \in M_{in} \}} \). Again, as in the previous result, \( Y_n \) is Borel measurable and it is clear from the construction that for each \( \omega \) in a set of probability 1, \( \lim_n \| Y_n(\omega) - Y(\omega) \| \to 0. \) Equivalently, \( \| Y_n - Y \| \to^a 0. \)

\( \square \)

For Hanš [21] and Hoffman-Jørgensen, et al. [24, 1977], attention is restricted to the case where \( \Xi \) is separable. An important point to note is that if we wish to relax
the separability assumption, we must then face the question of which definition of measurability to use.\textsuperscript{5}

Also, the $\sigma$-field on $\Xi$ (the Banach space) is determined by the norm on $\Xi$. Then $\| \cdot \|$ determines a topology $\mathcal{S}$, so $(\Xi, \mathcal{S})$ is a topological space. The closed sets of this $\mathcal{S}$ generate the Borel sets $\mathcal{M}$ on $\Xi$. How is this affected by non-separability? One example is that the ball $\sigma$-field and the Borel $\sigma$-field are the same if the space is separable. For non-separable spaces, maps can be ball measurable while failing to be Borel measurable (see van der Vaart and Wellner [69, 1996, p. 45]). An example of this is the space of cadlag functions on $[0,1]$ under the sup norm.

### A.4 The Bochner Integral and Expected Value

Of course, one of the first things we have to do if we want to talk about generalizing random variables is to define expectation in some reasonable fashion. The natural way to do this is via the Bochner integral. The reader may note at this point that the Bochner integral is a clever way to define the integral of a vector-valued function with respect to a scalar measure. There are meaningful theories of integration with respect to vector (valued) measures.

\textsuperscript{5}There is at least one other definition of measurability we have not mentioned, i.e. Price measurability.
Definition A.5 (HP [23], p. 78-79) Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. A countably-valued function $g : \Omega \to \Xi$ is integrable (Bochner) if and only if $\|g\|$ is integrable Lebesgue. In that case

$$(B) \int_A g \, d\mu = \sum_{i=1}^{\infty} c_i \mu(A_i \cap A)$$

where $A \in \mathcal{F}$.

We note that to say $g$ is countably-valued, we mean that $g = \sum_{i=1}^{\infty} c_i I_{A_i}$, where the $A_i$'s are a partition $\Omega$ into $\mathcal{F}$ sets.

Definition A.6 (HP [23], p. 79) Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. A function $g : \Omega \to \Xi$ is integrable (Bochner) if and only if there is a sequence of countably-valued integrable functions $\{g_n\}$ such that $g_n \to g$ a.s. and

$$\lim_n \int \|g_n - g\| d\mu = 0.$$

Then the Bochner integral of $g$ is defined as

$$(B) \int_A g \, d\mu = \lim_n (B) \int_A g_n \, d\mu$$

where $A \in \mathcal{F}$.

Bochner integrable functions are characterized by the following theorem.

Theorem A.3 (HP [23], p. 80) $g$ is integrable Bochner with respect to $\mu$ if and only if $g$ is strongly measurable and $\int \|g\| \, d\mu < \infty$. 
These definitions are rather formal. From an applied point of view, one finds that there is still more work to do in order to make any useful probability calculations. In the real case, we know that to calculate expectations our main options are sums and Riemann integrals. Further, these calculations cannot be justified without invoking a transformation theorem. This same problem persists for the $\Xi$-valued case. We need a transformation theorem. We start with (the usual) change of variable discussion.

Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be such that $\Omega$ and $\Omega'$ are sets and $\mathcal{F}$ and $\mathcal{F}'$ are sigma algebras of subsets of $\Omega$ and $\Omega'$, respectively. Let $X : \Omega \to \Omega'$ and suppose $X$ is measurable $\mathcal{F}/\mathcal{F}'$. If $P$ is a measure on $\mathcal{F}$, define another measure $P_X$ on $\mathcal{F}'$ by

$$P_X(A') = PX^{-1}A' \ , \ A' \in \mathcal{F}'$$

so $P_X$ is the induced measure on $\mathcal{F}'$. Let $g : \Omega' \to \Xi$, where $\Xi$ is a separable Banach space,

$$g : (\Omega', \mathcal{F}') \longrightarrow (\Xi, \mathcal{M})$$

and assume $g$ is measurable $\mathcal{F}'/\mathcal{M}$. Thus $gX$ is a Banach-valued function with domain on $\Omega$ and measurable $\mathcal{F}/\mathcal{M}$.

**Theorem A.4 (Transformation)** $gX$ is integrable Bochner (with respect to $P$) if and only if $g$ is integrable Bochner (with respect to $P_X$), in which case

$$\int_{X^{-1}A'} g(X\omega)P(d\omega) = \int_{A'} g(\omega')P_X(d\omega')$$

for each $A' \in \mathcal{F}$. 
Proof. As usual, we start with indicators. Let \( g = I_{A'} \) so that \( gX = I_{A'}(X) \). Then

\[
\int_{\Omega} g(X(\omega)) P(d\omega) = \int_{\Omega} I_{A'}(X(\omega)) P(d\omega) = \int_{X^{-1}A'} P(d\omega) = P(X^{-1}A')
\]

and

\[
\int_{\Omega'} g(\omega') P_X(d\omega') = \int_{\Omega'} I_{A'}(\omega') P_X(d\omega') = \int_{A'} P_X(d\omega') = P(X^{-1}A')
\]

Next consider simple functions \( g = \sum_i c_i I_{A'_i} \) where \( c_i \in \Xi, A'_i \in \mathcal{F}' \). Then

\[
\int_{\Omega} g(X(\omega)) P(d\omega) = \int_{\Omega} \sum_i c_i I_{A'_i}(X(\omega)) P(d\omega) = \sum_i c_i \int_{\Omega} I_{A'_i}(X(\omega)) P(d\omega)
\]  

where the linearity (second equality) follows from HP [23], Theorem 3.7.5, page 81.

Then (A.4) is

\[
= \sum_i c_i \int_{\Omega'} I_{A'_i}(\omega') P_X(d\omega') = \int_{\Omega'} \sum_i c_i I_{A'_i}(\omega') P_X(d\omega') = \int_{\Omega'} g(\omega') P_X(d\omega')
\]

where the first equality follows by the result above for simple functions.

Next consider integrable functions (i.e. Bochner integrable). There is an if and only if to prove, so we have two implications to show.

\[\iff\] Suppose \( g \) is Bochner integrable with respect to \( P_X \). By Definition A.6, \( g \) is strongly measurable, hence Borel measurable because \( \Xi \) is separable. By Theorem A.2, there is a sequence of simple functions \( g_n \) such that \( \lim_n \|g_n - g\| = 0 \) a.s. \( P_X \).

Now \( \{g_n(X)\} \) is a sequence of simple functions converging to \( g(X) \). Why? First,
since \( g_n \) is simple, \( g_n(\omega') = \sum_i c_i I_{A_i}(\omega') \Rightarrow g_n(X(\omega)) = \sum_i c_i I_{A_i}(X(\omega)) \) so that \( g_n(X) \) is simple. Second, pick \( \epsilon > 0 \). There is \( N_\epsilon \) such that

\[
\|g_n - g\| \leq \epsilon \quad \text{a.s. } P_X , \quad \forall n \geq N_\epsilon
\]

\[
\Leftrightarrow \quad P_X\{\|g_n - g\| \leq \epsilon\} = 1 , \quad \forall n \geq N_\epsilon
\]

\[
\Leftrightarrow \quad P_X\{\omega' : \|g_n(\omega') - g(\omega')\| \leq \epsilon\} = 1 , \quad \forall n \geq N_\epsilon . \quad (A.5)
\]

Let \( A'_\epsilon = [\|g_n - g\| \leq \epsilon] \). Then

\[
P_X A'_\epsilon = P_X^{-1} A'_\epsilon = P[X \in A'_\epsilon] .
\]

But \( P_X A'_\epsilon \) is the left hand side of (A.5) and

\[
P[X \in A'_\epsilon] = P\{\omega : \|g_n(X(\omega)) - g(X(\omega))\| \leq \epsilon\} = P\{\|g_n(X) - g(X)\| \leq \epsilon\}
\]

so that \( P[X \in A'_\epsilon] = 1, \forall n \geq N_\epsilon \). That is, \( \lim_n \|g_n(X) - g(X)\| = 0 \) a.s. \( P \).

By the previous result for simple functions,

\[
\int_{\Omega'} g_n(\omega')P_X(d\omega') = \int_{\Omega'} g_n(X(\omega))P(d\omega)
\]

and let \( n \to \infty \) on both sides (apply Dominated Convergence, Theorem A.5 to each side; see discussion following this proof for the details of applying DC here) to get

\[
\int_{\Omega'} g(\omega')P_X(d\omega') = \int_{\Omega} g(X(\omega))P(d\omega) . \quad (A.6)
\]
On the other hand, suppose $gX$ is Bochner integrable with respect to $P$. By Definition A.6, $gX$ is strongly measurable, hence Borel measurable because $\Xi$ is separable. By Theorem A.2, there is a sequence of simple functions $\{h_n\}$ such that $\lim_n \|h_n - gX\| = 0$ a.s. $P$. From Theorem A.2, we also know that the $h_n$'s are expressible as functions of $gX$, i.e. $h_n = f_n(gX)$. So let $g_n(\cdot) := f_n(g(\cdot))$. By properties of the induced measure $P_X$ (same argument as we did for $\implies$) it follows that $\{g_n\}$ is a sequence of simple functions converging to $g$ a.s. $P_X$. Continue as before, obtaining (A.6) again.

Theorem A.5 (Dominated Convergence) \footnote{This theorem can be found in Cohn, \textit{Measure Theory}, 1980, pg. 353.} Let $(\Omega, F, \mu)$ be a measure space, let $\Xi$ be a Banach space, and let $h$ be a $[0, +\infty]$-valued integrable function on $\Omega$. Suppose that $g$ and $g_1, g_2, \ldots$ are strongly measurable $\Xi$-valued functions on $\Omega$ such that $g_n \to g$ a.s. $\mu$ and $\|g_n\| \leq h$, $n = 1, 2, \ldots$ a.s. Then $g$ and $g_1, g_2, \ldots$ are integrable and $\int gd\mu = \lim_n \int g_n d\mu$.

In the previous proof, we needed a dominating $h$ both for $\implies$ and for $\impliedby$. For $\implies$, we had $g$ Bochner integrable and $\{g_n\}$ a sequence of simple functions such that $\lim_n \|g_n - g\| = 0$ a.s. For $\epsilon = 1$, there is $N_1$ such that

$$\|g_n\| \leq 1 + \|g\|, \quad \forall n \geq N_1.$$ 

Define $h = \max\{1 + \|g\|, \|g_1\|, \ldots, \|g_{N_1 - 1}\|\}$. By the integrability of $g$ and $g_1, \ldots, g_{N_1 - 1}$, $h$ is integrable. Also $h$ dominates $\{g_n\}$. Apply Theorem A.5 to get
\[
\lim_n \int_{\Omega'} g_n(\omega') P_X(d\omega') = \int_{\Omega'} g(\omega') P_X(d\omega')
\]

where we take \(\mu = P_X\).

For \(\Rightarrow\), we had \(gX\) Bochner integrable and showed that the sequence of simple functions \(\{h_n\}\) was expressible as \(\{g_n(X)\}\). The same dominating argument just made still works with \(g_n^* = g_n(X)\), \(g^* = g(X)\) and \(\mu^* = P\) yielding

\[
\lim_n \int_{\Omega} g_n^* d\mu^* = \int_{\Omega} g^* d\mu^* \iff \lim_n \int_{\Omega} g_n(X(\omega)) P(d\omega) = \int_{\Omega} g(X(\omega)) P(d\omega).
\]

\(\square\) **Example A.1**

In this example, we illustrate the calculation of \(Eg(X)\). Consider the case where \(X: \Omega \to \mathbb{R}\) and \(g: \mathbb{R} \to \Xi\), where \(\Xi\) is \(C(\Gamma)\). If both maps are measurable, then \(g(X)\) is a Borel measurable map from \(\Omega\) to \(\Xi\) and

\[
Eg(X) := \int_{\Omega} g(X) dP
\]

is the Bochner integral of \(g(X)\), provided that \(g(X)\) is Bochner integrable with respect to \(P\). By our transformation theorem \(g(X)\) is Bochner integrable with respect to \(P\) if and only if \(g\) is Bochner integrable with respect to \(P_X\). By Theorem A.3, \(g\) is Bochner integrable with respect to \(P_X\) if and only if \(g\) is strongly measurable and \(\int \|g\| dP_X < \infty\). Since \(g\) is Borel measurable and \(\Xi\) is separable, \(g\) is strongly measurable, so it is enough to verify that (assume the density \(f(x)\) of \(X\) exists)
\[ \int |g| dP_X = \int \sup_{\Gamma} |g| dP_X = \int \sup_{\Gamma} |g(x)| P_X(dx) = \int \sup_{\Gamma} |g(x)| f(x) dx < \infty \quad (A.7) \]

So if (A.7) is finite, then \( E_g(X) \) exists and (again by Theorem A.4) is given by

\[ E_g(X) = \int g(x) f(x) dx \]

\[ \circ \]

A.5  The Strong Law for \( \Xi \)-valued Random Variables

**Theorem A.6 (SLLN)**  Let \( Y_1, Y_2, \ldots \) be iid random variables which take values in a separable Banach space \( \Xi \). If \( E\|Y_1\| < \infty \) then

\[ \|Y_n - \mu\| \to 0 \quad a.s. \]

where \( \mu = EY_1 \).

**Proof.** Pick \( \epsilon > 0 \). Since \( \Xi \) is separable, by Theorem A.2 there is a simple random variable \( Z \) such that

\[ \|Y_1 - Z\| \leq \epsilon/3 \quad a.s. \quad (A.8) \]

and \( Z = \sum_i c_i I_{[Y_1 \in M_i]} \) where the \( M_i \)'s partition \( \Xi \), so that \( Z = f(Y_1) \) where \( f(y) = \sum_i c_i I_{M_i}(y) \). Since the \( M_i \) are \( \mathcal{M} \) sets, \( f : \Xi \to \Xi \) and is a measurable map. Let \( \mu_\epsilon = EZ \), then
\[ \| \mu - \mu_\epsilon \| = \| E Y_1 - EZ \| = \| E (Y_1 - Z) \| \leq E \| Y_1 - Z \| \leq \epsilon/3 \]

by (A.8). Now let \( W_n = f(Y_n) \), then \( W_1, W_2, \ldots \) are iid random variables with mean \( \mu_\epsilon \). Since they are finite-valued the ordinary SLLN still works (see notes after this proof) so

\[ W_n - \mu_\epsilon \rightarrow 0 \text{ a.s. } \Leftrightarrow \| W_n - \mu_\epsilon \| \rightarrow 0 \text{ a.s.} \quad (A.9) \]

Now \( \zeta_j = \| Y_j - W_j \| \) are iid real random variables (measurability is fine because of the separability assumption, e.g. for the sup norm, take the supremum over the rationals) so

\[ \frac{1}{n} \sum_{j=1}^{n} \| Y_j - W_j \| \overset{a.s.}{\rightarrow} E \| Y_1 - W_1 \| \leq \epsilon/3 \]

by (A.8). Also

\[ \| \overline{Y}_n - \mu \| = \| \overline{Y}_n - W_n + W_n - \mu_\epsilon + \mu_\epsilon - \mu \| \leq \| \overline{Y}_n - W_n \| + \| W_n - \mu_\epsilon \| + \| \mu_\epsilon - \mu \| \]

which implies that

\[ \lim \| \overline{Y}_n - \mu \| \leq \lim \| \overline{Y}_n - W_n \| + \lim \| W_n - \mu_\epsilon \| + \| \mu_\epsilon - \mu \| \]

\[ \leq \lim_{n} \frac{1}{n} \sum_{j=1}^{n} \| Y_j - W_j \| + \lim \| W_n - \mu_\epsilon \| + \| \mu_\epsilon - \mu \| \]

\[ = E \| Y_1 - W_1 \| + 0 + \| \mu_\epsilon - \mu \| \leq \epsilon/3 + \epsilon/3 \quad \square \]
Note about Theorem A.6: We need to discuss the claim made in the proof of Theorem A.6 just before (A.9). The finite dimensional SLLN referred to there is simple. Write

\[ W_j = \sum_i c_i I_{M_i}(Y_j), \quad W_j : \Omega \to \Xi \]

\[ \overline{W}_n = \frac{1}{n} \sum_{j=1}^n W_j = \frac{1}{n} \sum_{j=1}^n \sum_i c_i I_{M_i}(Y_j) \]

\[ = \frac{1}{n} \sum_i \sum_j c_i I_{M_i}(Y_j) = \sum_i c_i \frac{1}{n} \sum_{j=1}^n I_{M_i}(Y_j) \]

and notice that the random variable \( C_i := n^{-1} \sum_{j=1}^n I_{M_i}(Y_j) \) maps from \( \Omega \) to \( \mathbb{R} \), so is an ordinary random variable. Thus by the ordinary SLLN

\[ C_i = \frac{1}{n} \sum_{j=1}^n I_{M_i}(Y_j) \to P(Y \in M_i) \quad \text{a.s.} \]

\[ \Leftrightarrow \quad |C_i - P(Y \in M_i)| \to 0 \quad \text{a.s.} \]

So

\[ \| c_i C_i - c_i P(Y \in M_i) \| = \| c_i \| |C_i - P(Y \in M_i)| \to 0 \]

since \( \| c_i \| < \infty \) and fixed. Thus

\[ \left\| \sum_i c_i C_i - \sum_i c_i P(Y \in M_i) \right\| \leq \sum_i \| c_i \| |C_i - P(Y \in M_i)| \to 0 \]

\[ \Leftrightarrow \quad \overline{W}_n - \mu_c \to 0 \quad \text{a.s. where} \quad \mu_c := \sum_i c_i P(Y \in M_i) = E W_1. \]

\[ \blacksquare \]
Appendix B

Background on Generalized Glivenko-Cantelli Results

B.1 Generalized Glivenko-Cantelli Results

Consider independent sampling from a fixed probability measure $P$ on a set $S$ (e.g. $P_{\theta_0}$ is a probability measure on $\mathbb{R}^d$). The empirical measure is the probability measure that puts equal mass at each of the $n$ observations $\zeta_1, \zeta_2, \ldots, \zeta_n$. We will be concerned with classes of sets $\mathcal{D}$ and the behavior of

$$\sup_{D \in \mathcal{D}} |P_n D - PD|$$

which is equivalently written

$$\sup_{D \in \mathcal{D}} \left| \frac{1}{n} \sum_{j=1}^{n} I_{[\zeta_j \in D]} - P[\zeta \in D] \right| .$$

We would like to know conditions under which (B.1) converges almost surely to 0. Such a result represents a generalization of the univariate Glivenko-Cantelli theorem.

As an aside, we note that there are Glivenko-Cantelli results which are even more general (e.g. where the sup is taken over a class of functions); however, the generality of statements like (B.1) will be sufficient for our purposes. We recall the univariate G-C theorem.

\textsuperscript{1}This section is based on Pollard [44, 1984].
Theorem B.1 (Glivenko-Cantelli) If \( \zeta_1, \zeta_2, \ldots \) are iid, then

\[
\sup_{t \in (-\infty, \infty)} |F_n(t) - F(t)| \xrightarrow{a.s.} 0
\]

(B.2)

It is informative to reexpress (B.2) in empirical process notation. Let \( D := \{(-\infty, t] : t \in \mathbb{R}\} \).

We may rewrite (B.2) as

\[
\sup_{D \in \mathcal{D}} |P_nD - PD| \xrightarrow{a.s.} 0
\]

where

\[
F_n(t) = \frac{1}{n} \sum_{j=1}^{n} I_{(-\infty,t]}(\zeta_j) = \frac{1}{n} \sum_{j=1}^{n} I_{[\zeta_j \leq t]} = P_nD
\]

and

\[
F(t) = P [ \zeta \in (-\infty, t]] = PD
\]

Theorem B.1 is a powerful result which holds under quite general conditions. A natural question to ask is: What are the important theoretical properties utilized in the proof of Theorem B.1? Are these properties which we may generalize?

There are two important properties used in the proof of Theorem B.1. The first essential property is independence. The second essential property is: For a given set of \( n \) observations, one may choose at most \( n + 1 \) subsets (out of the \( 2^n \) possible subsets) of the observations using sets of the form \((-\infty, t], t \in \mathbb{R}\). Stated another way,
$\mathcal{D} = \{(-\infty, t] : t \in \mathbb{R}\}$ can pick out at most $(n + 1)$ subsets from a set of $n$ points in $\mathbb{R}$. See Figure B.1.

![Diagram](image)

**Figure B.1**

Similarly, $\mathcal{D} = \{(-\infty, t_1] \times (-\infty, t_2] : t_1, t_2 \in \mathbb{R}\}$ can pick out at most $(n + 1)^2$ points from a set of $n$ points in $\mathbb{R}^2$ (there are $n + 1$ places to put $t_1$ and $n + 1$ places to put $t_2$). It turns out that we will continue to observe the Glivenko-Cantelli behavior so long as $\mathcal{D}$ picks out at most a polynomial number of subsets from $S$.

**Definition B.1** A class $\mathcal{D}$ of subsets of some set $S$ is said to have polynomial discrimination\(^2\) if there is a polynomial $\rho(\cdot)$ such that for any set $S_0$ consisting of $n$ points in $S$, there are at most $\rho(n)$ distinct sets of the form $S_0 \cap \mathcal{D}$ with $\mathcal{D}$ in $\mathcal{D}$.

\(^2\) $\mathcal{D}$ is also called a Vapnik-Červonenkis class or a VC class. Often we abbreviate by saying that $\mathcal{D}$ is a polynomial class.
Why does this polynomial class idea work? The independence is used in applying Hoeffding’s Inequality to obtain an exponential term in the bound\(^3\)

\[
P \left\{ \sup_{D \in \mathcal{D}} |P_n D - PD| > \epsilon \right\} \leq \rho(n) e^{-\alpha^2 n}.
\]

Any polynomial \(\rho(n)\) is overwhelmed by the exponential term, which is still fast enough to apply Borel-Cantelli. We may now state the generalized Glivenko-Cantelli theorem.

**Theorem B.2 (generalized Glivenko-Cantelli)**  
*Let \(P\) be a probability measure on a space \(S\). For every permissible\(^4\) class \(\mathcal{D}\) of subsets of \(S\) with polynomial discrimination,*

\[
\sup_{\mathcal{D}} |P_n D - PD| \xrightarrow{a.s.} 0.
\]

**An application: concentric ellipsoidal bins**

As an illustration of an important application of this theorem, we will show that if the bins \(I_i\) are chosen from the class of concentric ellipsoidal bins (previously described), call this class \(\mathcal{C}\), then

---

\(^3\)See Pollard [64, 1984, p. 16], for example.

\(^4\)The permissibility requirement deals with measurability issues. The discussion of permissibility is postponed until Section B.2.
\[
\sup_{i \in C} \left| \frac{N_i}{n} - p_i(\theta_0) \right| \xrightarrow{a.s.} 0 \quad .
\]  

(B.3) is a critical step in the proof of our strong consistency theorem.

There are several useful results concerning the construction of complicated polynomial classes from more basic (polynomial) classes. We list several of these results without proof.

**Result B.1** If \( \mathcal{G} \) is a finite dimensional vector space of real-valued functions on \( S \), then the class

\[
\mathcal{D}_1 = \{g \geq t : t \in \mathbb{R}, g \in \mathcal{G}\} \quad \text{or} \quad \mathcal{D}_2 = \{g > t : t \in \mathbb{R}, g \in \mathcal{G}\}
\]

is a polynomial class.

**Result B.2** If \( \mathcal{D} \) and \( \mathcal{E} \) are polynomial classes, then so are \( \mathcal{D} \cup \mathcal{E}, \{D^c : D \in \mathcal{D}\}, \{D \cup E : D \in \mathcal{D}, E \in \mathcal{E}\}, \) and \( \{D \cap E : D \in \mathcal{D}, E \in \mathcal{E}\} \).

**Result B.3 (mappings)** If \( T \) is a map from a set \( \mathcal{X} \) into a set \( \mathcal{Y} \) and if \( \mathcal{D} \) is a polynomial class of subsets of \( \mathcal{Y} \), then \( \{T^{-1}D : D \in \mathcal{D}\} \) is a polynomial class of subsets of \( \mathcal{X} \).

Let \( g(c, \mu, A) := (x - \mu)^T A(x - \mu) - c \), where \( \mu \in \mathbb{R}^d, c \in \mathbb{R}^+ \), and \( A \) is symmetric and positive definite. Then sets of the form

\[
\{g(c, \mu, A) \leq 0\}
\]  

(B.4)
are ellipsoids in $\mathbb{R}^d$. Of course, $g(c, \mu, A)$ is a second degree polynomial in $\mathbb{R}^d$, and as $c$, $\mu$ and $A$ vary, the class $G$ of all such $g$ is a finite dimensional vector space of real-valued functions. Apply Result B.1 to get that

$$C' = \{g(c, \mu, A) > 0\} \quad (B.5)$$

is a polynomial class (of subsets of $\mathbb{R}^d$). Apply Result B.2 (complements) to get that (B.4) is also a polynomial class. Any ellipsoidal bin can either be expressed as (B.4), (B.5), or as

$$\{g(c, \mu, A) > 0\} \cap \{g(c, \mu, A) \leq 0\} \quad (B.6)$$

and again (B.6) is a polynomial class by Result B.2. Hence $C$, the class of all ellipsoidal concentric bins in $\mathbb{R}^d$, is a polynomial class. Assuming this class is permissible\(^5\) we may apply Theorem B.2 to establish (B.3).

### B.2 Permissible Classes and Measurability

In this section, we present Pollard’s definition of permissible classes and define some related terminology. The reader interested in more details than those given here should refer to Appendix C of Pollard [44, 1984].

\(^5\)See Section B.2 about permissibility.
Definition B.2 (Cohn [7], p. 251) A Polish space is a separable topological space that can be metrized by means of a complete metric.

Definition B.3 (Cohn [7], p. 261) Let $X$ be a Polish space. A subset $A$ of $X$ is analytic if there is a Polish space $Z$ and a continuous function $f : Z \to X$ such that $f(Z) = A$.

Note that every Borel subset of a Polish space is analytic, but there are analytic sets that are not Borel. Given a set $M$ and a $\sigma$-field $\mathcal{M}$ of subsets of $M$, denote by $\mathcal{A}(M)$ the analytic subsets of $M$. If $\mathcal{M}$ is complete for some probability measure $\mu$, then $\mathcal{A}(M) = \mathcal{M}$. To avoid ambiguity in the notation, if $M$ is a metric space choose $\mathcal{M}$ to be the Borel $\sigma$-field. For product spaces, choose the $\sigma$-algebra to be the product $\sigma$-field.

Definition B.4 (Pollard, 1984) Let $\zeta_1, \zeta_2, \ldots$ be measurable maps from $(\Omega, \mathcal{E}, \mathbf{P})$ into a set $S$ equipped with a sigma-field $\mathcal{S}$. A class $\mathcal{F}$ of $\mathcal{S}/\mathcal{B}(\mathbb{R})$-measurable, real-valued functions is given. We suppose without loss of generality that $\mathcal{F} = \{f(\cdot, t) : t \in T\}$ where $T$ is a separable metric space. We say that the class $\mathcal{F}$ is a permissible class if it can be indexed by a $T$ in such a way that

(i) the function $f(\cdot, \cdot)$ is $\mathcal{S} \otimes \mathcal{B}(T)$-measurable as a function from $S \otimes T$ to $\mathbb{R}$, and

(ii) $T$ is an analytic subset of a compact metric space $\overline{T}$. 
Notice here that $\otimes$ denotes the Cartesian product when applied to two sets, but denotes the product $\sigma$-field when applied to two $\sigma$-fields.\(^6\) The manner in which permissibility is used to establish the measurability of certain random objects is somewhat detailed, but is discussed fully in Appendix C of Pollard [44, 1984]. In those details certain properties of analytic sets are used. As discussed in the next section, the generality of analytic sets is not needed for most of the applications we have in mind.

B.3 Easy Ways to Establish Permissibility

For the cases of interest to us, the class $\mathcal{F}$ will be a class of indicator functions of sets. As mentioned in Pollard (1984), there are several easy ways to demonstrate permissibility. We mention two of these now.

For the case of concentric ellipsoidal bins, consider the class

$$\mathcal{C} = \{ [g(c, \mu, A) \leq 0]: c \in \mathbb{R}^+, \mu \in \mathbb{R}^d, A \text{ p.s.d} \}$$

$$\cup \{ [g(c, \mu, A) > 0]: c \in \mathbb{R}^+, \mu \in \mathbb{R}^d, A \text{ p.s.d} \}$$

$$\cup \{ [g(c, \mu, A) > 0] \cap [g(c, \mu, A) \leq 0]: c \in \mathbb{R}^+, \mu \in \mathbb{R}^d, A \text{ p.s.d} \}.$$

The class $\mathcal{C}$ contains all of the concentric ellipsoidal bins. If $\mathcal{C}_Q$ is this same class restricted to rational $c$, $\mu$, and $A$ then of course

\(^6\)By product $\sigma$-field is meant the $\sigma$-field generated by the measurable rectangles.
\[ \sup_{C \in \mathcal{Q}} |P_n C - PC| = \sup_{C} |P_n C - PC| \]  \quad (B.7)

Since each \( C \) is Borel, \( |P_n C - PC| \) is measurable for each choice of \( C \). The sup on the left in (B.7) is taken over a countable collection, hence it is measurable. While a certain measurability of cross-sections is required in Pollard’s proof of a Symmetrization Lemma ([44], p. 14), observe that this can be dispensed with in cases where the index set can be taken to be countable, as is the case for the class \( C \) here.

It is important to note that in general measurability of a quantity like (B.7) may not be the only measurability issue to take care of, although it is the only measurability issue for the situation here.

For a second example, consider the case where we wish to show that

\[ \sup_{\Gamma} |p_{i}^{*}(\theta) - p_{i}(\theta)| \to 0 \quad \text{a.s.} \]

where \( \Gamma \) is some compact subset of \( \Theta \). Often we will simulate\(^7\) values \( x \) which land in some set \( \mathcal{X} \) and then take \( g_{\theta}(x) \) (for some function \( g_{\theta}(.\cdot) \)) as our simulated realization of the observed process. If we let

\[ \mathcal{D}_i = \{ [x : g_{\theta}(x) \in I_i] : \theta \in \Gamma \} \]

then \( \sup_{\mathcal{D}_i} |P_{\sigma} D - PD| = \sup_{\Gamma} |p_{i}^{*}(\theta) - p_{i}(\theta)| \), which will be measurable provided that \( \mathcal{D}_i \) is permissible. Now if \( g_{\theta}(x) \) is measurable \( \mathcal{B}(\Gamma) \) for each choice of \( x \) and is measurable \( \mathcal{X} \) for each choice of \( \theta \), then

\(^7\)We will discuss this particular scenario in greater detail later on.
\[ [x : g_\theta(x) \in I_i] \]

is an \( S \otimes B(\Gamma) \)-measurable set, so that its indicator function is an \( S \otimes B(\Gamma) \)-measurable function. Further, \( \Gamma \) itself is compact. Hence \( D \) is permissible by definition.
Appendix C

Literature search for Perlman papers

In this appendix, we summarize some of the results of our efforts to search for a multiparameter version of Perlman [42, 1983]. Another paper which is very closely related to Perlman [42, 1983] is Perlman [41, 1972]. Perlman himself does not revisit the issues treated in [42, 1983] (in anything published).

Besides much other searching in the literature, the Citation Index was also used in an attempt to follow up on any subsequent work. The searching based on the references found in the Citation Index is discussed in this appendix. The Citation Index does not cover all of the statistical journals, but it covers many of the important ones. The journals it does cover are: Advances in Applied Probability, American Statistician, Annals of Probability, Annals of Statistics, Applied Statistics - JRSS Series C, Biometrics, Biometrika, Fuzzy Sets and Systems, International Statistical Review, Journal of Applied Probability, Journal of Multivariate Analysis, JASA, JRSS - Series A - Statistics in Society, JRSS - Series B - Methodological, Probability Theory and Related Fields, Scandinavian Journal of Statistics, Stochastic Processes and their Applications, Technometrics, and Theory of Probability and its Applications. The following discussion refers to journals which are covered by this

---

1This appendix was completed on November 10, 1997.
Index, i.e. if we say "this paper is never referenced" we mean with respect to the journals covered by the Index.

Perlman (1983) is only referenced once. That reference is Reeds [49]. The subject of that paper is the Cauchy location problem (i.e. 1-d), hence it is not relevant to the issue of the multiparameter version of Perlman [42, 1983]. It is possible that Perlman [42, 1983] may have been generalized with a reference being made only to Perlman [41, 1972] and not to Perlman [42, 1983]. Thus it is necessary that we also follow up on possible references to Perlman [41, 1972].

The Perlman [41, 1972] paper is in many ways an important heir to the work of Wald [71, 1949] (as we mention in Appendix D). It provides more general sufficient conditions for strong consistency than all previous papers (i.e. papers subsequent to and including Wald [71, 1949]). It also provides some necessary and sufficient conditions for strong consistency. Some of these conditions (both the sufficient conditions and the necessary and sufficient conditions) have been relaxed by some more recent papers (this is also discussed in Appendix D). It suffices to keep in mind that more general formulations of the conditions in Perlman [41, 1972] are now established, particularly ones which handle non-parametric situations (e.g. where the parameter \( \theta \) in question is our unknown density \( f \)).

According to the Citation Index, the Perlman [41, 1972] paper is referenced 15 times. The complete list of all 15 of these references is given at the end of this appendix. Not all of them are relevant to our interests. The references which are directly
relevant are Vajda [68, 1995], Wang [72, 1985], and Strasser [60, 1981]. In addition to the important theory given in those papers, some important counterexamples (of where consistency fails, etc.) are given in Ferguson [16, 1982] and Lehmann [29, 1980]. None of these papers generalizes the Perlman [42, 1983] paper to the multiparameter case (they all deal largely with generalizations and/or topics related to strong consistency of the MLE).

Appendix D

Consistency Summary for Maximum Likelihood

Although it is not directly relevant to our main focus, we outline in this appendix the major developments in the sufficient conditions for consistency (both strong and weak, but our emphasis is on strong) over the past 50 years. This summary came about naturally as a result of our literature search on the Perlman papers, and we believe it may be of some use to anyone who is interested in the topic. It is not appropriate (nor would it be very helpful) to write down all the conditions here. The purpose of this appendix is to point the interested reader in the correct direction. The interested reader should probably note that there is no fast way to get to the conditions. All of these papers are fairly dense. In the absence of time, I would advise looking at Strasser [60, 1981]. Certainly, the last place to start if time is an issue would be Wald [71, 1949] or Perlman [41, 1972].

Generally, the Wald [71, 1949] paper is cited as the most important paper along these lines. Authors since Wald have relaxed the sufficient conditions somewhat. The general theme, however, in all of these results is that the conditions given are usually fairly mathematical and are difficult to work with even for very simple distributions. And the irony is that in the case of difficult problems - which are the cases where one needs to know whether or not maximum likelihood estimators are consistent -
the conditions as given in all these papers are unlikely to be very helpful. There is an apparent need here then (as in many areas of statistics) for someone to connect the theory with practice. It seems unlikely that this will occur for some time to come.

- **Cramér [10, 1946]** - Cramér proves that the MLE is consistent and asymptotically normal under some assumptions. His biggest assumption is that \( \frac{\partial^3}{\partial \theta^3} \log f(x, \theta) \) exists a.s. For the rest of these assumptions, see Sen and Singer [55, 1993], page 205. Note that the existence of third derivative can be relaxed to their requirement in (ii) (b).

- **Wald [71, 1949]** - Wald proves (strong) consistency of the MLE under very weak assumptions. In particular, he does not require existence of any derivatives (hence, we do not even need to have the existence of the likelihood equation). Most of his assumptions have to do with technical things like (i) certain things have to be measurable, (ii) certain things have to be integrable, etc. There are 8 assumptions altogether, but several of them can be relaxed. Wald does not show proofs for greatest generality because he wants the exposition to be as simple as possible. Notably, the proofs involving singular distributions are not included. Densities are assumed to be absolutely continuous or discrete throughout.

- **Perlman [41, 1972]** - This paper is very important, and it subsumes all papers on the same subject after Wald [71, 1949]. The main purpose of the paper is to
generalize the sufficient conditions for strong consistency as much as possible.

Perlman finds conditions which are more general than LeCam [28, 1953], Kiefer and Wolfowitz [27, 1956], Huber [26, 1967], Bahadur [2, 1967], and Crawford [11, 1958] (these references are given in Perlman [41, 1972]). These conditions are given (in Perlman [41, 1972]) on page 269. His \textit{global dominance} assumption is referred to as $D$.

\textbf{Result 1:} \(D\) and Condition 1 \(\implies\) strong consistency.

\textbf{Result 2:} \(D\) and Condition 2 \(\implies\) strong consistency.

The hypothesis in either \textbf{Result 1} or \textbf{Result 2} subsumes all previous (work by other authors). However, it was known (at that time) that consistency holds in some examples where the hypotheses of these two results were not satisfied. This led Perlman to introduce another hypothesis which required a \textit{global uniformity} assumption, he calls $U$, as well as a different condition (Condition 3):

\textbf{Result 3:} \(U\) and Condition 3 \(\implies\) strong consistency (see page 275).

Condition 3 is weaker than conditions 1 and 2, however I don’t believe that the hypotheses for \textbf{Result 3} have been shown to be either stronger or weaker than the hypotheses for the other two results. As is evident from our main focus in this dissertation, there are many important ideas in this paper.
• **Strasser [60, 1981]** - He establishes nothing new in the way of sufficient conditions, but has a very concise restatement of all the assumptions and definitions required for Perlman [41, 1972]. If one has very little time to spend, this would be the first place to look.

• **Wang [72, 1985]** - She generalizes Perlman's sufficient conditions. Again, some of her theorem statements may be more concise and easier to follow than Perlman's. However, she refers the reader back to Perlman for many of the definitions and notation. Strasser [60, 1981] has it all completely restated.

• **Vajda [68, 1995]** - Some results on weak consistency. Context is stochastic processes, but iid is a special case.
Appendix E

Goodness-of-Fit Literature

It was desired to find (or develop) rigorous theoretical justification of the asymptotic properties SIMEST. That is, it was desired to determine under what conditions we have consistency and/or asymptotic normality of $\hat{\theta}_n$, where $\hat{\theta}_n$ is the simulation based estimate of $\theta_0$. Since SIMEST is closely related to MCS (minimum chi-square) and other methods, it seemed logical to start with results for MCS and subsequently determine what alterations needed to be made in order to account for the simulation error. Unfortunately, it was not possible to find what we were looking for in the MCS literature, as will be discussed in this chapter.

In section E.1, we discuss the obvious approaches one would like to use in a MCS scenario. In section E.2, we discuss how parameter estimation is related to the goodness-of-fit (GOF) literature. Section E.2 also discusses how the GOF literature provides inadequate justification for the approaches suggested in Section E.1.¹

E.1 Obvious Approaches to Minimum Chi-square

For clarity, we restrict our attention to MCS, although there are many other reasonable methods to use. MCS chooses the parameter estimate $\hat{\theta}_n$ for $\theta_0$ as

¹Due to some misplaced notes, I have been forced to rely on my memory in some parts of this literature review; there will undoubtedly be errors. Apologies to the reader.
$$\hat{\theta}_n = \arg \min_{\theta \in \Theta} X_n^2(\theta) = \arg \min_{\theta \in \Theta} \sum_{i=1}^{k} \frac{(N_i - np_i(\theta))^2}{np_i(\theta)}.$$ 

Now we have in mind the situation where we observe $X_1, X_2, \ldots$ from a continuous density in $\mathbb{R}^d$, and we desire to bin these observations in some reasonable fashion and calculate $\hat{\theta}_n$. We would like to have consistency for $\hat{\theta}_n$ and also some asymptotic distribution for $\hat{\theta}_n$. Now straightforward intuition suggests that we should be sure to satisfy

1. $k \to \infty$ as $n \to \infty$ and $k/n \to 0$, and

2. the bins should be chosen based upon the observed data.

Now 1. and 2. are seemingly innocuous enough. It is perhaps an **astounding fact** that no theory exists today which (under any assumptions whatever) allows both 1. and 2. for a general setup.

### E.2 Parameter Estimation in the Goodness-of-Fit Literature

The nearest statisticians have come to treating both 1. and 2. simultaneously is in the excellent paper of Drost [13, 1989]. He treats 1. and 2. simultaneously for the case of one dimensional observations from a location-scale family. No improvements have been made since Drost [13]. In light of the fact that this area has been so heavily researched, this should suggest to the reader that the theoretical issues involved here are non-trivial.
MCS receives essentially no treatment at all as a serious method of parameter estimation, except for the early work by Rao.\textsuperscript{2} This fact stands in great contrast to the great theoretical attention which has been given to maximum likelihood estimation. The treatment of MCS and related methods (since Rao's work) appears to have fallen within the context of the goodness-of-fit (GOF) literature.

Within the GOF literature, there are two scenarios, simple \( H_0 \) and composite \( H_0 \). In the case of a simple \( H_0 \), one considers \( H_0 : F = F_0 \), where \( F_0 \) is completely specified and \( H_\alpha : F \neq F_0 \). One compares

\[
X^2_n(\theta_0) = \sum_{i=1}^{k} \frac{(N_i - np_i(\theta_0))^2}{np_i(\theta_0)}
\]

with the appropriate quantile of \( \chi^2_{k-1} \), rejecting for large values of \( X^2_n(\theta_0) \). For fixed \( k \), this is appropriate even when the bin boundaries are determined from the data \textbf{under some conditions on both the underlying model and the method of binning} (cf. A.R. Roy [51, 1956], Watson [74, 75, 1958, 1959]).\textsuperscript{3}

In the case of composite \( H_0 \), one considers \( H_0 : F \in \{F_\theta : \theta \in \Theta\} \) versus \( H_\alpha : F \notin \{F_\theta : \theta \in \Theta\} \). Here \( \theta_0 \) is estimated via MCS, call this estimate \( \hat{\theta}_n \). Then for fixed \( k \), \( X^2_n(\hat{\theta}) \) is asymptotically \( \chi^2_{k-s-1} \), where \( s = \text{dim}(\Theta) \). Again, this is appropriate even when the bin boundaries are determined from the data \textbf{under the appropriate}

\textsuperscript{2}The early work by Rao always deals with the situation where \( k \) is fixed and the bins are non-random.

\textsuperscript{3}It is important to note here that these older papers dealt with the situation where the bin boundaries are functions of the parameter estimates \( \hat{\theta}_n \). Bins based upon order statistics do not fit into these early formulations. Without referring back to some of these papers, I cannot recall any more of the details.
conditions on the underlying model and the method of binning (cf. Chibisov [6, 1971], Moore [34, 1971], and Moore and Spruill [36, 1975]). It turns out that these older papers are not sufficiently flexible to deal with the types of binning strategies that we care about.\textsuperscript{4}

In the literature, much of the time the properties of the parameter estimates based on GOF criteria are treated only as secondary issues. Typically these properties are dealt with only as a nuisance on the way to establishing distributional results such as $X_n^2(\hat{\theta}) \sim \chi^2_{k-J-1}$ asymptotically. More dramatically, consistency of $\hat{\theta}_n$ is often one of the theoretical assumptions in a paper (often times even distributional results about $\hat{\theta}_n$ are taken as assumptions). At times, these assumptions are unjustified.\textsuperscript{5} Probably the only careful treatment of consistency in the any of the random bins papers is Pollard [43, 1979]. We discuss elsewhere why the treatment he gives there does not suit our needs.

In the two subsections to follow, we make a few brief comments on the most important papers. The GOF literature may be classified into papers which (a) treat the case of random bins with $k$ fixed or (b) treat the case of non-random bins where the number of bins goes to infinity. The exception to this classification is the Drost

\textsuperscript{4}Without going into all the details, the older “random-bins” papers do not allow for the possibility that those bins are quantile-based (univariate observations). Put another way, the older papers assume a structure about the bins which depends upon the assumed model. We are interested in binning methods which, though data-based, do not use the model assumptions.

\textsuperscript{5}For example, Moore and Spruill [36] refer back to one of the Watson papers for justification of the random-bins consistency of $\hat{\theta}$. In fact, however, the Watson papers do not work for the situation where the bins are based on the order statistics - which is one of the possibilities under the Moore and Spruill [36] formulation.
[13] paper which treats both (a) and (b) at the same time. Most of what follows are technical notes; hence, the reader may be well advised to skip the details and go directly to Table E.1.

**Random bins, number of bins fixed**

**Chibisov [6, 1971]:** This paper treats the case where the observations are univariate and the number of bins $k$ is fixed. Under certain conditions (which are too restrictive for our needs), he shows that Pearson’s chi-square statistic is asymptotically $\chi^2$. He also treats the Chernoff-Lehmann statistic.

**Moore [34, 1971]:** He extends the Chernoff-Lehmann results$^6$ to the case where the bins are hypercubes in $\mathbb{R}^d$. We are not interested in Chernoff-Lehmann type results. Also, we need more general shapes than rectangles.

**Moore and Spruill [36, 1975]:** Extends Chibisov [6] and Moore [34] to a more general class of statistics, specifically statistics constructed from certain quadratic forms. See line 1.2 in Drost [13] for example. Still in the fixed $k$ case.

In this paper there is a framework which is general enough to deal with bins chosen independently of the model. However, there are still problems with this paper, notably:

---

$^6$“Chernoff-Lehmann results” refers to the situation where the parameters $\hat{\theta}_n$ are estimated via some fully efficient method (like fully parametric maximum likelihood), and one wishes to know what the resulting asymptotic distribution of $X_n^2(\hat{\theta}_n)$ is. Generally, the asymptotic distribution will depend on $\theta_0$; however, some $\theta_0$-free asymptotic distributions are available in special cases.
1. It refers back to the Watson papers (in the 50's, i.e. [74, 75] and others) for justification of a consistency assumption. However, the Watson papers do not treat the case where the bin boundaries are independent of the model. Therefore, there is no consistency proof we can use here.

2. It still (as in Moore [34]) uses rectangular bins in $\mathbb{R}^d$.

**Number of bins goes to infinity, bins non-random**

There do not appear to be rigorous treatments of consistency for methods such as MCS in the case where $k \to \infty$. All that exists in this regard seem to be off-hand comments and very light heuristics. What evidence is there to support this claim?

1. Read and Cressie's [47, 1988] book contains an extensive literature review section and there is nothing relevant there, except perhaps a remark on p. 140, where they say that the work of Tumanyan [67, 1956], Steck [59, 1957], and Morris [37, 1966] "culminated in the landmark paper of Morris [38, 1975]...". But the Morris [38] paper only treats the case of simple $H_0$ (see Table E.1). Thus although one might suggest that Tumanyan [67] appears to be relevant, this cannot be the case.

2. Greenwood and Nikulin's [19, 1996] book has more recent coverage of the literature with a similar story. Greenwood and Nikulin seem to address the issue on page 107 by referring to the Watson [74, 1958] paper. However, the Watson [74]
paper has no proof or development, only a heuristic remark (e.g. "... it seems that the following is probably true..."). As mentioned by Drost [13] - in the context of his setup - several previous authors had claimed that their results were also true as \( n, k \to \infty \); however they offered no proof. Watson offers no proof.

<table>
<thead>
<tr>
<th>author [ref] (year)</th>
<th>binning</th>
<th>null</th>
<th>cells?</th>
<th>real/vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chibisov [6] (1971)</td>
<td>fixed ( k )</td>
<td>composite</td>
<td>random</td>
<td>real-valued</td>
</tr>
<tr>
<td>Moore [34] (1971)</td>
<td>fixed ( k )</td>
<td>composite</td>
<td>random</td>
<td>vector-valued</td>
</tr>
<tr>
<td>Moore &amp; Spruill [36] (1975)</td>
<td>fixed ( k )</td>
<td>composite</td>
<td>random</td>
<td>vector-valued</td>
</tr>
<tr>
<td>Holst [25] (1972)</td>
<td>( k \to \infty )</td>
<td>simple</td>
<td>fixed</td>
<td>multinomial</td>
</tr>
<tr>
<td>Morris [38] (1975)</td>
<td>( k \to \infty )</td>
<td>simple</td>
<td>fixed</td>
<td>multinomial</td>
</tr>
<tr>
<td>Drost [13] (1989)</td>
<td>( k \to \infty )</td>
<td>composite</td>
<td>random</td>
<td>real-valued</td>
</tr>
</tbody>
</table>

**Table E.1** Goodness-of-fit literature summary.
Appendix F

An Asymptotic Result for the Case Where the Number of Bins Tends to Infinity

As described in Appendix E, there are no rigorous treatments of the asymptotics for MCS (minimum chi-square) parameter estimates in the case where the number of bins is allowed to tend to infinity (except for Drost [13, 1989]). This statement is also true of methods similar to MCS, such as MML (multinomial maximum likelihood). In this chapter, we work with the MML criterion. The chapter culminates in Theorem F.1, which gives conditions under which \( \hat{\theta}_n \) is asymptotically normal, where

\[
\hat{\theta}_n = \arg \max_{\theta \in \Theta} \sum_{i=1}^{k} \frac{N_i}{n} \log p_i(\theta).
\]

The reader familiar with the goodness-of-fit literature will recognize how unusual the identifiability assumption of Theorem F.1 seems. This unusualness is a quick indicator that Theorem F.1 is new. The body of the chapter consists of the proofs of many of the preliminary results needed to establish Theorem F.1.

When we began the work in this chapter, we had hoped to develop a proof for the \( k \to \infty \) case when the bins are non-random (which we have succeeded in doing, to a certain extent). With such a proof in hand, we would have hoped to have then shown that the effect of the random bins would disappear asymptotically, thus obtaining our ultimate goal of a (general) theorem treating (a) random bins and (b) \( k \to \infty \),
simultaneously for MML. As it has turned out, the work of the present chapter has fallen short of a completely satisfactory theorem for the non-random bins case. Hence, we were unable to go any further.

This chapter restricts attention to the case of univariate observations from a family with only one parameter (i.e. \(d = 1\) and \(s = 1\)).\(^1\) One of the main problems with the results in this chapter is that we are forced to require that the infinite bins remain fixed. Let \(I_1 \equiv (-\infty, a]\) and \(I_k \equiv [b, \infty)\),\(^2\) so that \(I_1\) and \(I_k\) are the infinite bins. By keeping the infinite bins fixed, we mean that \(a\) and \(b\) remain fixed as \(k \to \infty\), i.e. the refinements of the partition take place only on \([a, b]\). There are major difficulties which develop if one allows the infinite bins to march away to infinity (that is, if one allows \(a \to -\infty\) and \(b \to \infty\)). As a heuristic justification for keeping the infinite bins fixed, we observe that as \(k \to \infty\), if we put roughly equal probabilities in each bin, then the leftmost and rightmost bins (the infinite bins) will more or less sit still, relative to the other bins. Putting roughly equal probabilities in each bin forces \(a \to -\infty\) and \(b \to \infty\) technically, but this will be ever so slow relative to the behavior of the refining process in the \([a, b]\) region, i.e. region of high probability. While these heuristics are intuitively obvious, a proof of the conjecture that we can neglect the rate at which \(a\) and \(b\) tend to \(\pm \infty\) seems out of reach.

\(^1\)The generalization to the multiparameter case is fairly quick, however.

\(^2\)Assume that as \(k \to \infty\), we keep relabeling the bins so that \(I_k\) always refers to the infinite bin on the right.
We have included this chapter in this dissertation mainly in the hope that it might be a building block for something more useful. It has also been included because it represents roughly 4-5 months of work. It has been placed in an appendix due to the fact that we do not feel that it is satisfactory at its present level of development. If we could get around the restriction that the infinite bins remain fixed, then this chapter might perhaps then be considered appropriate for the main body of the dissertation.

F.1 Technical Assumptions

In this brief section, we deal with several issues which tend to underlie most of the arguments throughout this entire chapter, but which are somewhat too bothersome to mention repeatedly. It will be understood by the reader that, for example, if we are to do any operations on a particular quantity, then that quantity must exist (e.g. it cannot be undefined).

Assumption F.1 \( \forall \theta \in \Theta, \) there is a continuous distribution \( P_\theta \) with univariate density \( f(x, \theta). \) The support of each \( f(x, \theta) \) is a connected subset of \( \mathbb{R}, \) possibly all of \( \mathbb{R}. \)\(^3\) The support of this family of densities does not depend upon \( \theta. \)

Assumption F.2 \( p_i(\theta_0) > 0, \) \( i = 1, \ldots, k, \) for all \( k. \)

\(^3\)Notice that in our proofs, we assume that the support is all of \( \mathbb{R}, \) since this is the most difficult case. Any smaller region of \( \mathbb{R} \) only makes the proofs easier.
We observe that this assumption is relatively standard. In the presence of Assumption F.1, Assumption F.2 ensures that \( p_i(\theta) > 0, i = 1, \ldots, k \) for all \( k \).

**Assumption F.3** \( \forall \theta \in \Theta, \frac{\partial}{\partial \theta^r} p_i(\theta), i = 1, \ldots, k, r = 1, 2, 3, \) exist for all \( k \).

As an aside, we observe that we may relax any assumption such as this one which requires certain behavior \( \forall \theta \in \Theta, \) to one which only requires \( \forall \theta \in N(\theta_0), \) where \( N(\theta_0) \) is some neighborhood of \( \theta_0 \). From a practical standpoint such a relaxation is purely formal, hence we do not dwell upon it here.\(^4\)

**Consequences of Assumptions F.1, F.2, and F.3**

Let \( g(\theta) = \sum_{i=1}^{k} N_i \log p_i(\theta) \), the multinomial maximum likelihood ratio criterion function. Let \( g^{(r)}(\theta) = \frac{\partial^r}{\partial \theta^r} g(\theta), r = 1, 2, 3. \) Observe that \( g(\theta) \) and \( g^{(r)}(\theta), r = 1, 2, 3, \) are all defined if Assumptions F.1 - F.3 hold, since in that case

\[
g^{(1)}(\theta) = \sum_{i=1}^{k} N_i \frac{\frac{\partial}{\partial \theta} p_i(\theta)}{p_i(\theta)}
\]

\[
g^{(2)}(\theta) = \sum_{i=1}^{k} N_i \left\{ -\left( \frac{\frac{\partial}{\partial \theta} p_i(\theta)}{p_i(\theta)} \right)^2 + \frac{\frac{\partial^2}{\partial \theta^2} p_i(\theta)}{p_i(\theta)} \right\}
\]

\[
g^{(3)}(\theta) = \sum_{i=1}^{k} N_i \left\{ 2 \left( \frac{\frac{\partial}{\partial \theta} p_i(\theta)}{p_i(\theta)} \right)^3 - 3 \frac{\frac{\partial}{\partial \theta} p_i(\theta)}{p_i(\theta)} \frac{\frac{\partial^2}{\partial \theta^2} p_i(\theta)}{p_i(\theta)^2} + \frac{\frac{\partial^3}{\partial \theta^3} p_i(\theta)}{p_i(\theta)} \right\}.
\]

\(^4\)Note further (as we mention near the end of this technical section), that we could change "for all \( k \)" to "for all \( k \) beyond some fixed \( k_0 \)." That is to say, as long as everything eventually exists and is well-defined and once this happens it stays that way.
Clearly, if Assumption F.2 fails, then \( g(\theta) \) and \( g^{(r)}(\theta) \), \( r = 1, 2, 3 \), all fail to exist even if Assumption F.3 holds. Currently, I am not sure if there are any important examples where Assumption F.3 fails, Assumption F.2 holds, yet \( g^{(r)}(\theta) \) still exists. I don’t think so.

**Comment on Assumption F.3**

The existence in Assumption F.3 is implicitly assumed in Lemma F.2, assumption 4. (although we explicitly mention it there in footnote 34). As we demonstrate in the discussion to follow, however, it is not necessary to require Assumption F.3 in either Lemma F.3 or Theorem F.1.

The subject of the existence of \( \frac{\partial^r}{\partial \theta^r} p_i(\theta) \) is very much related to the question of whether or not the interchange

\[
\frac{\partial^r}{\partial \theta^r} \int_{I_i} f(x, \theta)dx = \int_{I_i} \frac{\partial^r}{\partial \theta^r} f(x, \theta)dx \quad (F.1)
\]

is appropriate.\(^5\) For clarity, we consider the case of \( r = 1 \), i.e.\(^6\)

\[
\frac{\partial}{\partial \theta} \int_{I_i} f(x, \theta)dx = \int_{I_i} \frac{\partial}{\partial \theta} f(x, \theta)dx \quad (F.2)
\]

\(^5\)The reader will observe that much of this discussion is related to advanced calculus.

\(^6\)For a comment on what new issues arise for \( r = 2 \) and \( r = 3 \), see footnote 38.
When the \( I_i \)'s are finite, continuity of the two integrands in (F.1) on \( I_i \times \overline{N(\theta_0)} \) is sufficient\(^7\) for (F.2) to be valid (cf. Trench [66, 1978], *Theorem 5.1*, p. 576), and the existence of the left hand side of (F.2) follows for free.

In the case where \( I_i \) is infinite, we suppose without loss of generality that \( I_i = [\alpha, \infty) \), some finite \( \alpha \) (we treat this in great detail in Result F.11). If the integral on the right hand side of (F.2) is uniformly convergent on \( \overline{N(\theta_0)} \) and the integrands are continuous on \( [\alpha, \infty) \times \overline{N(\theta_0)} \), then it is only necessary to have \( \int_{\alpha}^{\infty} f(x, \theta) \, dx \) converge for some \( \theta \in \overline{N(\theta_0)} \) in order to conclude (F.2).\(^8\) But \( \int_{\alpha}^{\infty} f(x, \theta) \, dx \) of course converges no matter what \( \theta \) we choose since each \( f(x, \theta) \) is a density function by Assumption F.1. Hence not only does (F.2) follow (for \( I_i = [\alpha, \infty) \)), but the existence of the left hand side of (F.2) follows for free.

The point here is that we may assume \( \frac{\partial}{\partial \theta} f(x, \theta) \) exists along with a few other mild assumptions. And from this, the existence in Assumption F.3 is guaranteed to follow.

In the situation we have in mind where one is binning observations from a continuous density (as opposed to the situation where the observations are discrete), it seems more reasonable to ask for \( \frac{\partial}{\partial \theta} f(x, \theta) \) to exist than to ask for Assumption F.3.

---

\(^7\)See footnote 24 about the notation "\( N(\theta_0) \)."

\(^8\)See *Theorem 5.6*, Trench [66], p. 586.
Taylor expansion of $g(\theta)$

Almost immediately in the main development to follow (in Section F.2) we will expand $g'(\hat{\theta}_n)$ in a Taylor series around $\theta_0$. We need to be aware of when this is appropriate. Let us take some neighborhood of $\theta_0$, say $N(\theta_0)$. Firstly, when we choose such a neighborhood, we must be sure that it is contained in $\Theta$. Secondly, to expand $g'(\theta)$ to third order about $\theta_0$ for any $\theta \in N(\theta_0)$, we must assume that $g'''(\theta)$ exists on $N(\theta_0)$. We have just shown how Assumptions F.1 - F.3 guarantee that $g'''(\theta)$ exists on $N(\theta_0)$. Nothing more is needed.$^9$

Things are a little trickier if Assumption F.3 holds only $\forall \theta \in N(\theta_0)$, since in that case $\hat{\theta}_n$ must (of course) fall in $N(\theta_0)$ before the Taylor series argument will be valid. Theoretically, though, we may start watching the behavior for as large an $n$ as we like. Hence we can even wait (in probability) for the appropriate quantities to exist (i.e. until the probability that $\hat{\theta}_n$ falls in $N(\theta_0)$ is large). This is purely formal, of course, but it somewhat unties our hands theoretically. There are no practical implications of this remark.

$^9$As a practical remark, however, even if the various expressions in the Taylor series argument exist (as guaranteed by Assumptions F.1 - F.3), this does not guarantee that the (Taylor) approximation (and hence the normal approximation) will be very good. Obviously, if $\hat{\theta}_n$ falls far from $\theta_0$, $\theta_n^*$ may also, and in such a case the $o_p(1)$ remainder term may still be very large.
F.2 Outline

This chapter establishes asymptotic normality of the parameter estimate (notice only one parameter) based on maximizing the multinomial likelihood criterion when (i) the infinite bins remain fixed,\(^\text{10}\) at say \(I_1 = (-\infty, a]\) and \(I_k = [b, \infty)\) and (ii) the number of bins on the interior of \([a, b]\) is allowed to tend to infinity with \(n\). It is appropriate now to briefly outline the structure of our main argument, which follows Lehmann [30, 1983], pp. 415-417, and is at least as old as Cramér [10, 1946]. Let

\[
g(\theta) := \sum_{i=1}^{k} N_i \log p_i(\theta) \ .\]

Expand \(g'(\hat{\theta}_n)\) in a Taylor series about \(\theta_0\) where

\[
\hat{\theta}_n := \arg \max_{\theta \in \Theta} \sum_{i=1}^{k} \frac{N_i}{n} \log p_i(\theta) = \arg \max_{\theta \in \Theta} g(\theta) \ .
\]

Necessarily we have \(g'(\hat{\theta}_n) = 0\) so

\[
0 = g'(\hat{\theta}_n) = g'(\theta_0) + (\hat{\theta}_n - \theta_0)g''(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^2 g'''(\theta_n^*) , \quad \theta_n^* \in (\theta_0, \hat{\theta}_n) \ .
\]

And upon rearranging, we obtain

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\frac{1}{\sqrt{n}} g'(\theta_0)}{-\frac{1}{n} g''(\theta_0) - \frac{1}{2n}(\hat{\theta}_n - \theta_0)g'''(\theta_n^*)} \ . \tag{F.3}
\]

\(^{10}\)It will be convenient to use a shorthand notation to distinguish the infinite bins from the finite bins. The indices \(i = 1, k\) will be reserved for the two finite bins. That is, \(I_1 = (-\infty, a]\) and \(I_k = [b, \infty)\). Even though \(k\) changes, we will assume that the bins are “labeled” so that \(i = k\) represents the rightmost (infinite bin). Thus the indices \(i = 2, \ldots, k - 1\) will refer to the finite (or interior, that is, interior to \([a, b]\)) bins, and \(i \neq 1, k\) will be a shorthand way of referring to the collection of all these finite bins (for a particular choice of \(k\)). Similarly \(i = 1, k\) will be a shorthand way of referring to the two infinite bins.
Now
\[-\frac{1}{n} g''(\theta_0) - \frac{1}{2n} (\hat{\theta}_n - \theta_0) g''(\theta_n^*) = \]
\[-\sum_{i=1}^{k} h_i p_i + \left( \sum_{i=1}^{k} h_i p_i - \frac{1}{n} g''(\theta_0) \right) - \frac{1}{2n} (\hat{\theta}_n - \theta_0) g''(\theta_n^*) , \]
where

\[h_i = h_i(\theta_0) := \frac{\partial^2}{\partial \theta^2} \log p_i(\theta_0) .\]

The main results to be demonstrated are (under the appropriate assumptions) the following

\[\hat{\theta}_n \overset{P}{\rightarrow} \theta_0 \quad (F.4)\]

\[\frac{\frac{1}{\sqrt{n}} g'(\theta_0)}{\sqrt{\sum_{i=1}^{k} g_i^2 p_i}} \overset{D}{\rightarrow} N(0, 1) \quad \text{where} \quad g_i = g_i(\theta_0) := \frac{\partial}{\partial \theta} \log p_i(\theta_0) \quad (F.5)\]

\[\sum_{i=1}^{k} h_i p_i - \frac{1}{n} g''(\theta_0) = o_p(1) \quad (F.6)\]

\[-\frac{1}{2n} (\hat{\theta}_n - \theta_0) g''(\theta_n^*) = o_p(1) \quad (F.7)\]

\[-\sum_{i=1}^{k} h_i p_i = \sum_{i=1}^{k} g_i^2 p_i \quad (F.8)\]
\[ \sum_{i=1}^{k} g_i^2 p_i \text{ is bounded below.} \quad (F.9) \]

And the reader may observe that with such results established, we may multiply both sides of (F.3) by \( \sqrt{\sum_{i=1}^{k} g_i^2 p_i} \) to get

\[ \sqrt{\sum_{i=1}^{k} g_i^2 p_i} \sqrt{n} (\hat{\theta}_n - \theta_0) = \frac{\sqrt{\sum_{i=1}^{k} g_i^2 p_i} \frac{1}{\sqrt{n}} g'(\theta_0)}{\sum_{i=1}^{k} g_i^2 p_i + o_p(1)} \]

\[ = \{1 + o_p(1)\}^{-1} \frac{\frac{1}{\sqrt{n}} g'(\theta_0)}{\sqrt{\sum_{i=1}^{k} g_i^2 p_i}} \sim N(0,1) \quad (F.10) \]

by Slutsky's theorems.\(^{11}\) The remainder of our work then lies in establishing (F.4)-(F.9). In the following sections, we treat each of these issues in order. The reader uninterested in the details may sip to Theorem F.1 for the conclusions.

### F.3 Weak consistency of \( \hat{\theta}_n \)

In this subsection, we extend the proof of Rao [46, 1965] to the \( k \to \infty \) case.

**Result F.1** Suppose that the following conditions are satisfied.

1. \( \forall \delta > 0, \exists k_0 \text{ and } \eta_1, \eta_2 > 0 \text{ small enough that} \)

\[ \sum_{i=1}^{k} (p_i(\theta_0) - \eta_1) \log \frac{p_i(\theta_0)}{p_i(\theta)} \geq \eta_2, \forall k \geq k_0, \forall \theta \in \{||\theta - \theta_0|| \geq \delta\} \]

\(^{11}\)We point out that if \( \sum_{i=1}^{k} g_i^2 p_i \) is bounded below, then \( \left( \sum_{i=1}^{k} g_i^2 p_i \right)^{-1} \) is bounded above. In turn, this ensures that \( \left( \sum_{i=1}^{k} g_i^2 p_i \right)^{-1} \times o_p(1) = o_p(1) \). This is exactly what is needed to guarantee the equality in (F.10) holds.
2. $p_i(\theta)$, $i = 1, \ldots, k$, are continuous functions of $\theta$.

3. $f(x, \theta_0)$ is bounded above, $k/n \max_{i \neq 1, k} \delta_i \to 0$, $k/n \to 0$ and the infinite bins remain fixed.\footnote{Note that we may relax the conditions on the infinite bins by requiring only that they get no larger as we increase $k$ and $n$.}

Then $\hat{\theta}_n \overset{p}{\to} \theta_0$, where

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} \sum_i \frac{N_i}{n} \log p_i(\theta)$$

**Proof.** By assumption 1., we claim that $\max_i \left| \frac{N_i}{n} - p_i \right| \overset{p}{\to} 0$. To see this, let

$$A_i(\epsilon) = \left\{ \left| \frac{1}{n} N_i - p_i \right| \geq \epsilon \right\}.$$ 

Then

$$P(A_i(\epsilon)) \leq \frac{1}{\epsilon^2} E \left( \frac{N_i}{n} - p_i \right)^2 = \frac{1}{\epsilon^2 n} p_i (1 - p_i)$$

and

$$P \left\{ \max_i \left| \frac{1}{n} N_i - p_i \right| \geq \epsilon \right\} = P \left\{ \bigcup_i A_i(\epsilon) \right\} \leq \sum_i P(A_i(\epsilon)) \leq k \max_i P(A_i(\epsilon))$$

$$\leq k \max_i \frac{1}{\epsilon^2 n} p_i (1 - p_i) \leq \frac{1}{\epsilon^2 n} \max_{i \neq 1, k} f(\zeta_i, \theta_0) \delta_i + \frac{1}{\epsilon^2 n} \max_i p_i$$
\[ \leq K_0 k \max_{i \neq 1, k} \delta_i + \frac{1}{\epsilon^2} \max_{i=1,k} p_i \to 0 \]  

(F.11)

by assumption 3.

Pick \( \delta > 0 \). Let \( S = \{ \theta : ||\theta - \theta_0|| \leq \delta \} \), \( S^+ = \{ \theta : ||\theta - \theta_0|| \geq \delta \} \). By assumption 1., \( \exists k_0 \) and \( \eta_1, \eta_2 \) small enough that

\[ \sum (p_i(\theta_0) - \eta_1) \log \frac{p_i(\theta_0)}{p_i(\theta)} \geq \eta_2 \ , \ \forall \ k \geq k_0 \ , \ \forall \ \theta \in S^+ . \]  

(F.12)

Since \( n \) and \( k \) are related by assumption 3., notice that \( \exists n_\delta = n(k_0) \) such that (F.12) holds for \( n \geq n_\delta \). Now fix a \( \gamma \in (0, 1) \); we may take \( \gamma \) to be as small as we wish. The statements to follow take place on \( A_\gamma \) where \( P(A_\gamma) \geq 1 - \gamma \), \( \forall n \geq n_0 \). From (F.11),

\[ p_i(\theta_0) - \eta_1 \leq \frac{N_i}{n} \leq p_i(\theta_0) + \eta_1 \ , \ i = 1, \ldots, k \ , \ \forall n \geq n_0 \]

which implies that \( \forall n \geq n_0 \) and \( \forall \theta \in \Theta \),

\[ \sum (p_i(\theta_0) - \eta_1) \log \frac{p_i(\theta_0)}{p_i(\theta)} \leq \sum \frac{N_i}{n} \log \frac{p_i(\theta_0)}{p_i(\theta)} \leq \sum (p_i(\theta_0) + \eta_1) \log \frac{p_i(\theta_0)}{p_i(\theta)} \]

\[ \implies \sum \frac{N_i}{n} \log \frac{p_i(\theta_0)}{p_i(\theta)} \geq \eta_2 > 0 \ , \ \forall n \geq \max\{n_0, n_\delta\} \ , \ \forall \theta \in S^+ \]

\[ \iff \sum \frac{N_i}{n} \log p_i(\theta_0) > \sum \frac{N_i}{n} \log p_i(\theta) \ , \ \forall n \geq \max\{n_0, n_\delta\} \ , \ \forall \theta \in S^+ \]  

\[ \text{Of course, we do not need the norm } || \cdot || \text{ here, since we are only dealing with a one-dimensional parameter. However, this argument is completely general.} \]
Since $n^{-1} \sum N_i \log p_i(\theta)$ is continuous by assumption 2, it must attain a maximum in $S$. If there is more than one maximum in $S$, let $\hat{\theta}_n$ be the largest one. Thus $\hat{\theta}_n \in S$, for $n$ large enough on $A_\gamma$. But the choice of $\delta$ was arbitrary.

F.4 Asymptotic normality of $n^{-1/2}g'(\theta_0)$

Result F.2 14 If $k, n \to \infty$ and $n \min_i p_i(\theta_0) \to \infty$ then

$$\frac{\frac{1}{\sqrt{n}} g'(\theta_0)}{\sqrt{\sum_{i=1}^{k} \frac{g_i^2}{p_i}}} = \frac{\sum_{i=1}^{k} N_i \frac{\partial}{\partial \theta} \log p_i(\theta_0)}{\sqrt{\sum_{i=1}^{k} \left( \frac{\partial}{\partial \theta} \log p_i(\theta_0) \right)^2 p_i(\theta_0)}} \xrightarrow{D} N(0, 1).$$

Proof. Recall the notation $g_i = g_i(\theta_0) = \partial / \partial \theta \log p_i(\theta_0)$ and $p_i = p_i(\theta_0).$ 15 We first observe that, although for each fixed $k$ we may rewrite $n^{-1/2} \sum_{i=1}^{k} N_i g_i$ as an average of the $n$ $X_j$'s, the standard central limit conditions don't work because the variance of the terms in the sum changes with each change in $n$. Consider the triangular array

$$X_{11}, \ X_{12}, \ldots, \ X_{1n},$$

$$X_{21}, \ X_{22}, \ldots, \ X_{2n},$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

---

14 We note that this result is completely general and does not require the special situation where the infinite bins remain fixed. We need the infinite bins to remain fixed starting in Section F.5.

15 We suppress the argument $\theta$ in the case where $\theta = \theta_0$. This is a standard notation.
where $X_{nj} := n^{-1/2} \sum_{i=1}^{k} I_{ji} g_i$ and $I_{ji} := I_{[X_j \in I_i]}$. Observe that in each row, $X_{nj}$, $j = 1, 2, \ldots$ are independent.\textsuperscript{16} Since $\sum_{i=1}^{k} p_i(\theta) = 1$, regardless of the choice of $\theta \in \Theta$, we may differentiate both sides to obtain

$$\sum_{i=1}^{k} \frac{\partial}{\partial \theta} p_i(\theta) = 0 \ , \ \forall \theta \in \Theta \ ,$$

so that in particular, \eqref{eq:12} holds for $\theta = \theta_0$.\textsuperscript{17} Thus

$$E[X_{nj}] = \frac{1}{\sqrt{n}} \sum_{i=1}^{k} g_i p_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{k} \frac{\partial}{\partial \theta} p_i = 0 \ .$$

Also,

$$E[X_{nj}^2] = E\left[\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{k} I_{ji} g_i\right)^2\right] = \frac{1}{n} \sum_{i=1}^{k} g_i^2 p_i$$

and

$$s_n^2 = \sum_{j=1}^{n} E\left[X_{nj}^2\right] = \sum_{i=1}^{k} g_i^2 p_i .$$

\textsuperscript{16}For clarity, $X_1, X_2, \ldots$ is the infinite sequence of observations. The $X_{nj}$'s are derived from these observations. Each $X_{nj}$ depends probabilistically only upon $X_j$ (i.e. $X_{nj}$ is measurable $\sigma(X_j)$), but the $X_{nj}$'s also depend upon the choice of $k$ and the location of the bin boundaries.

\textsuperscript{17}The reader may wonder why it is not necessary to assume one interchange here, thinking along the lines of the classical argument which shows that $E\left[\frac{\partial}{\partial \theta} \log f(X, \theta_0)\right] = \int \frac{\partial}{\partial \theta} \log f(x, \theta_0) f(x, \theta_0) dx = \int \frac{\partial}{\partial \theta} f(x, \theta_0) dx = 0$. It is a subtle point that this same classical argument also shows that $\int \frac{\partial}{\partial \theta} f(x, \theta) dx = 0$ no matter what the choice of $\theta \in \Theta$ so long as one interchange of $\frac{\partial}{\partial \theta}$ with $\int$ is valid, i.e.

$$\int f(x, \theta) dx = 1 \Rightarrow \frac{\partial}{\partial \theta} \int f(x, \theta) dx = 0 \ \text{interch.} \Rightarrow \int \frac{\partial}{\partial \theta} f(x, \theta) dx = 0 \ .$$

(Only the interpretation of \eqref{eq:14} as an expectation requires that the “true” distribution be $P_0$.) Notice, however, that in our situation we do not need to require that 1 interchange be valid. Thus there is no exact equivalence between our situation and \eqref{eq:14}. We will see more of this in Result F.9.
Here the Lindeberg condition (cf. Billingsley [4, 1995], equation (27.8)) is that \( \forall \epsilon > 0, \)

\[
\lim_{n \to \infty} \frac{1}{s_n^2} \sum_{j=1}^{n} \int_{|X_{n_j}| \geq \varepsilon s_n} X_{n_j}^2 dP = 0.
\]  

(F.15)

So (let \( B = \{|X_{n_j}| \geq \varepsilon s_n\} \) and \( AI_r = \{X \in I_r\} \))

\[
\int_B X_{n_j}^2 dP = E \left[ X_{n_j}^2 I_B \right] = \sum_{r=1}^{k} E \left[ X_{n_j}^2 I_{B I_{A_r}} \right] = \sum_{r=1}^{k} \left( \frac{1}{\sqrt{n}} g_r \right)^2 \left( I_{B I_{A_r}} \right)
\]

\[
= \frac{1}{n} \sum_{r=1}^{k} g_r^2 \mathbb{P} \left( |X_{n_j}| \geq \varepsilon s_n, X_j \in I_r \right)
\]

\[
= \frac{1}{n} \sum_{r=1}^{k} g_r^2 \mathbb{P} \left( |X_{n_j}| \geq \varepsilon s_n \mid X_j \in I_r \right) \mathbb{P} \left( X_j \in I_r \right)
\]

\[
= \frac{1}{n} \sum_{r=1}^{k} g_r^2 \mathbb{P} \left( \left( \frac{|g_r|}{s_n} \right) \geq \sqrt{n} \varepsilon \right) p_r = \frac{1}{n} \sum_{r=1}^{k} g_r^2 \mathbb{P} \left( \left( \frac{|g_r|}{s_n} \right) \geq \sqrt{n} \varepsilon \right) p_r
\]

where once we apply the conditioning, the event \( \{|g_{kr}|/s_n \geq \sqrt{n} \varepsilon\} \) is non-stochastic, hence has probability 0 or 1. We will show that it will go to 0 due to the assumption on the rate of the \( p_r(\theta_0) \)'s. Clearly,

\[
\frac{|g_r|}{s_n} = \frac{\sqrt{g_r^2}}{\sqrt{\sum_{r=1}^{k} g_r^2 p_r}} = \sqrt{\frac{g_r^2 p_r}{\sum_{r=1}^{k} g_r^2 p_r}} \frac{1}{\sqrt{p_r}} \leq \frac{1}{\sqrt{p_r}} \leq \max_{1 \leq r \leq k} \frac{1}{\sqrt{p_r(\theta_0)}}
\]

But

\[
n \min_i p_i(\theta_0) \to \infty.
\]  

(F.16)
\[
P\left(\frac{|g_r|}{s_n} \geq \sqrt{n}\epsilon\right) \leq P\left(\max_{1 \leq r \leq k} \frac{1}{\sqrt{p_r(\theta_0)}} \geq \sqrt{n}\epsilon\right) = P\left(\frac{1}{\min_{1 \leq r \leq k} \sqrt{p_r(\theta_0)}} \geq \sqrt{n}\epsilon\right)
\]

\[
= P\left(\epsilon^{-1} \geq \sqrt{n} \min_{1 \leq r \leq k} \sqrt{p_r(\theta_0)}\right) = P\left(\epsilon^{-1} \geq \sqrt{n} \min_{1 \leq r \leq k} p_r(\theta_0)\right) \rightarrow 0 \quad (F.17)
\]

by (F.16). Notice that the event in (F.17) is non-stochastic. Thus it will have probability 0 or 1. Thus it will be 0 (identically) forever as soon as \(\sqrt{n} \min_{1 \leq i \leq k} p_i(\theta_0) > \epsilon^{-1}\).

Continuing,

\[
\sum_{j=1}^{n} \int_{\{X_{nj} \geq s_n\}} X_{nj}^2 \, dP = \sum_{j=1}^{n} \frac{1}{n} \sum_{r=1}^{k} g_r^2 p_r \left(\frac{|g_r|}{s_n} \geq \sqrt{n}\epsilon\right)
\]

\[
= \sum_{r=1}^{k} g_r^2 p_r \left(\frac{|g_r|}{s_n} \geq \sqrt{n}\epsilon\right) \leq \max_{1 \leq r \leq k} g_r^2 p_r \sum_{r=1}^{k} P\left(\frac{|g_r|}{s_n} \geq \sqrt{n}\epsilon\right)
\]

which implies that

\[
\frac{1}{s_n^2} \sum_{j=1}^{n} \int_{\{X_{nj} \geq s_n\}} X_{nj}^2 \, dP \leq \frac{\max_{1 \leq r \leq k} g_r^2 p_r}{s_n^2} \sum_{r=1}^{k} P\left(\frac{|g_r|}{s_n} \geq \sqrt{n}\epsilon\right)
\]

\[
\leq \sum_{r=1}^{k} P\left(\frac{|g_r|}{s_n} \geq \sqrt{n}\epsilon\right) \rightarrow 0 .
\]

Notice that the sum here presents no problems even as \(k \rightarrow \infty\). Once \(\sqrt{n} \min_{1 \leq r \leq k} p_k(\theta_0) > \epsilon^{-1}\), all \(k\) probabilities are 0. This is rather remarkable. Thus we have verified the Lindeberg condition (F.15); hence we may apply Theorem 27.2 of Billingsley [4] to get the result.
F.5 Proof of \( \sum_{i=1}^{k} h_i p_i - \frac{1}{n} g''(\theta_0) = o_p(1) \)

Result F.3 Suppose that the infinite bins remain fixed\(^{18}\) and that for any choice of \( \alpha \in \mathbb{R} (\alpha \text{ finite}), \)

\[
\frac{\partial^r}{\partial \theta^r} \int_{-\infty}^{\alpha} f(x, \theta_0) dx = \int_{-\infty}^{\alpha} \frac{\partial^r}{\partial \theta^r} f(x, \theta_0) dx, \quad r = 1, 2, \tag{F.18}
\]

and similarly for \([\alpha, \infty),^{19}\) and suppose that \( f(x, \theta_0) \) and \( f^{(r)}(x, \theta_0), \quad r = 1, 2, \) are continuous,\(^{20}\) where \( f^{(r)}(x, \theta) := \frac{\partial^r}{\partial \theta^r} f(x, \theta). \) Suppose that there exists a \( K_{30} \) such that for \( x \in [a, b], \)

\[
f(x, \theta_0) \geq K_{30} > 0. \tag{F.19}
\]

Further, suppose that

\[
\int_{-\infty}^{\alpha} f^{(r)}(x, \theta_0) dx \quad \text{and} \quad \int_{b}^{\infty} f^{(r)}(x, \theta_0) dx, \quad r = 1, 2, \tag{F.20}
\]

are all finite and that

\[
\frac{k}{\sqrt{n}} \max_i \frac{1}{\sqrt{p_i(\theta_0)}} \to 0. \tag{F.21}
\]

\(^{18}\)From here on, we will fix the left-most infinite bin as \( I_1 \equiv (-\infty, a] \) and the right-most infinite bin as \( I_k \equiv [b, \infty). \) To the symbol "\( \equiv \)," we attribute the standard meaning of "is identically equal to for all \( k \)". From here on, the phrase "the infinite bins remain fixed" refers to the meaning we have just discussed. See also, footnote 10.

As a technical point, one may take \( I_k \equiv (b, \infty) \), we never use the fact that the bin is closed at \( b \).

\(^{19}\)If the interchanges are not valid, we will still be fine; however, the assumptions get a little more complicated. We avoid the complications by asking for a little more than we need.

\(^{20}\)One only needs continuity on \([a, b]\).
Then there exist finite constants $K_{3r}$, $r = 1, 2$, such that for $x \in [a, b]$

$$
|f^{(1)}(x, \theta_0)| \leq K_{31}, \quad |f^{(2)}(x, \theta_0)| \leq K_{32}, \quad (F.22)
$$

and

$$
\sum_{i=1}^{k} h_i p_i - \frac{1}{n} g''(\theta_0) \xrightarrow{L_1} 0.
$$

Proof. Recall from Section F.2 that $\frac{1}{n} g''(\theta_0) := \frac{1}{n} \sum_{i=1}^{k} N_i h_i$. Thus

$$
E \left[ \sum_{i=1}^{k} h_i p_i - \frac{1}{n} g''(\theta_0) \right] = E \left[ \sum_{i=1}^{k} \left( h_i p_i - \frac{h_i N_i}{n} \right) \right] \leq \frac{1}{n} \sum_{i=1}^{k} E \left[ |N_i - n p_i| |h_i| \right]
$$

$$
\leq \frac{1}{n} \sum_{i=1}^{k} \left( n p_i (1 - p_i) \right)^{1/2} |h_i| \leq \frac{1}{\sqrt{n}} \sum_{i=1}^{k} p_i^{1/2} |h_i| = \frac{1}{\sqrt{n}} \sum_{i=1}^{k} p_i |h_i| \frac{1}{p_i^{1/2}}
$$

$$
\leq \left( \frac{k}{\sqrt{n}} \max_{i} \frac{1}{\sqrt{p_i}} \right) \max_{i} |h_i| p_i.
$$

which goes to 0 if $\max_{i} |h_i| p_i$ is eventually bounded, since the other term goes to 0 by (F.21). We show now that $\max_{i} |h_i| p_i$ is eventually bounded. For $i = 1, k$, \footnote{Recall that the notation $i = 1, k$ and $i \neq 1, k$ is shorthand, see footnote 10.}

$$
|h_i| p_i := \left| \frac{\partial^2}{\partial \theta^2} \log p_i \right| p_i = \left| -\frac{\left( \frac{\partial}{\partial \theta} p_i \right)^2}{p_i} + \frac{\partial^2}{\partial \theta^2} p_i \right|
$$

$$
= \left| -\frac{\left( \int \frac{\partial}{\partial \theta} f(x, \theta_0) dx \right)^2}{p_i(\theta_0)} + \int \frac{\partial^2}{\partial \theta^2} f(x, \theta_0) dx \right|
$$
which is finite (and remains constant as \( n, k \to \infty \) since the infinite bins are assumed to be fixed) by the assumption (F.20). Notice also that the last equality follows since the necessary interchanges are valid by (F.18).

For \( i \neq 1, k \),

\[
|h_i| p_i = \left| -\frac{\partial}{\partial \theta} p_i \right|^2 + \frac{\partial^2}{\partial \theta^2} p_i \leq \left( \frac{\partial}{\partial \theta} p_i \right)^2 + \frac{\partial^2}{\partial \theta^2} p_i
\]

\[
= \left( \frac{\partial}{\partial \theta} \int_{I_i} f(x, \theta_0) dx \right)^2 \frac{\partial^2}{\partial \theta^2} \int_{I_i} f(x, \theta_0) dx + \int_{I_i} \frac{\partial^2}{\partial \theta^2} f(x, \theta_0) dx
\]

\[
= \left( \frac{\partial}{\partial \theta} \int_{I_i} f(x, \theta_0) dx \right)^2 p_i + \int_{I_i} \frac{\partial^2}{\partial \theta^2} f(x, \theta_0) dx
\]  \hspace{1cm} (F.24)

We observe that the interchanges are valid due to the continuity assumptions (see, e.g., Trench [66], Theorem 5.1, p. 576 - an advanced calculus result). Also by the continuity assumptions, we may apply the Mean Value Theorem for integrals to get that (F.24) is

\[
= \left( \frac{\partial}{\partial \theta} f(\zeta_1^{(2)}, \theta_0) \right)^2 \frac{\partial^2}{\partial \theta^2} f(\zeta_1^{(1)}, \theta_0) \left| \delta_i \right| \]  \hspace{1cm} (F.25)

where \( \zeta_i^{(r)}, r = 1, 2, 3, \) are points (which are guaranteed to exist by the Mean Value Theorem for Integrals) in the interval \( I_i \).\(^22\) Also, \( \delta_i \) is the bin width for interval \( I_i \).

Notice that the \( \delta_i \)'s in the numerator and denominator cancel in the left hand term of

\(^{22}\)We use the superscript \( r \) to indicate that each \( \zeta_i \) is different and depends upon the integrand, e.g. \( \zeta_i^{(1)} \) depends upon \( I_i \) as well as upon \( f(x, \theta_0) \).
Finally, by continuity of $f^{(r)}(x, \theta_0)$, $r = 1, 2$, there exists constants $K_{31}$ and $K_{32}$ satisfying (F.22). By assumption, there exists a $K_{30}$ satisfying (F.19), so (F.25) is

$$\leq \left( \frac{K_{31}}{K_{30}} \right)^2 p_i + K_{32}\delta_i \to 0$$ (F.26)

since $\max_{i \neq 1, k} p_i \to 0$ and $\max_{i \neq 1, k} \delta_i \to 0$. Combining these results shows that $\max_i |h_i|p_i$ is eventually bounded. Hence (F.23) goes to 0 as desired. □□□

### F.6 Proof of $-\frac{1}{2n}(\hat{\theta}_n - \theta_0)g'''(\theta^*_n) = o_p(1)$

#### Lemma F.1 For any real-valued functions $g(\theta)$ and $h(\theta)$ it holds that\(^{23}\)

$$\sup_{\theta} \left| g(\theta)h(\theta) \right| \leq \sup_{\theta} \left| g(\theta) \right| \sup_{\theta} \left| h(\theta) \right|$$

**Proof.**

$$\left| g(\theta)h(\theta) \right| = \left| g(\theta) \right| \left| h(\theta) \right| \leq \left| g(\theta) \right| \sup_{\theta} \left| h(\theta) \right| \leq \sup_{\theta} \left| g(\theta) \right| \sup_{\theta} \left| h(\theta) \right|$$

which implies that

$$\sup_{\theta} \left| g(\theta)h(\theta) \right| \leq \sup_{\theta} \left| g(\theta) \right| \sup_{\theta} \left| h(\theta) \right|$$

\(\Box\)

\(^{23}\)The reason this is such a general result is that its proof depends entirely upon the definition of supremum and the absolute value property of $|xy| = |x||y|$.
Result F.4 Suppose that

1. The infinite bins remain fixed (see footnote 18),

2. There is a neighborhood of $\theta_0$, say $N(\theta_0)$, and a finite constant $K_{10}$ such that

$$f(x, \theta) \geq K_{10} > 0 \quad \forall \theta \in N(\theta_0), \, \forall x \in [a, b], \text{ and}$$

3. On $[a, b] \times \overline{N(\theta_0)}$, $f(x, \theta)$ and $f^{(r)}(x, \theta)$ are continuous,\(^{24}\) where $f^{(r)}(x, \theta) := \frac{\partial^r}{\partial \theta^r} f(x, \theta)$

Then\(^ {25}\) there is a neighborhood of $\theta_0$, say $N(\theta_0)$, and finite constants $K_{1r}, \, r = 1, 2, 3$, such that

$$|f^{(r)}(x, \theta)| \leq K_{1r} \quad \forall \theta \in N(\theta_0), \, \forall x \in [a, b], \, r = 1, 2, 3,$$

and

$$\sup_{\theta \in N(\theta_0)} \left| \sum_{i \neq 1, k} \frac{N_i}{n} \frac{\partial^3}{\partial \theta^3} \log p_i(\theta) \right|$$

is eventually bounded a.s.

\textbf{Proof.} Observe that for $i \neq 1, k$,\(^ {26}\)

\(^{24}\)Notice that here we use $\overline{N(\theta_0)}$ to denote the closure of $N(\theta_0)$, i.e. $N(\theta_0)$ together with its limit points. There is no practical difference between $N(\theta_0)$ and $\overline{N(\theta_0)}$. We merely must have a closed set in order to apply an advanced calculus theorem. It suffices to ask for the smallest closed set around $N(\theta_0)$, which is $\overline{N(\theta_0)}$.

\(^{25}\)Recall that the notation $i = 1, k$ and $i \neq 1, k$ is shorthand, see footnote 10.

\(^{26}\)Recall that the notation $i = 1, k$ and $i \neq 1, k$ is shorthand, see footnote 10.
\[
\frac{\partial^3}{\partial \theta^3} \log p_i(\theta) = 2 \left( \frac{\partial}{\partial \theta} \frac{p_i(\theta)}{p_i(\theta)} \right)^3 - 3 \frac{\partial^2}{\partial \theta^2} \frac{p_i(\theta)}{p_i(\theta)^2} \frac{\partial^2}{\partial \theta^2} \frac{p_i(\theta)}{p_i(\theta)} + \frac{\partial^3}{\partial \theta^3} \frac{p_i(\theta)}{p_i(\theta)}
\] (F.27)

so that by an argument identical to that in the proof of Result F.3 (again notice that the bin widths, \(\delta_i\), in the numerators and denominators all cancel)

\[
= 2 \left( \frac{f^{(1)}(\xi_i^{2\theta})}{f(\xi_i^{1\theta})} \right)^3 - 3 \frac{f^{(1)}(\xi_i^{2\theta}) f^{(2)}(\xi_i^{3\theta})}{f(\xi_i^{1\theta})^2} + \frac{f^{(3)}(\xi_i^{4\theta})}{f(\xi_i^{1\theta})} .
\] (F.28)

The assumptions about continuity over \([a, b] \times N(\theta_0)\) allow us to apply Theorem 5.1, Trench [66], p. 576, to get that the necessary interchanges are valid. Further, these same continuity assumptions allow us to apply the Mean Value Theorem for Integrals as needed. Unlike in Result F.3, the \(\xi_i^{\theta}\)'s in (F.28) depend upon \(\theta\). This (additional) dependence is indicated by our notation. Continuing, from (F.27) and (F.28) it follows that

\[
\sup_{\theta \in N(\theta_0)} \left| \sum_{i \neq 1, k} \frac{N_i}{n} \frac{\partial^3}{\partial \theta^3} \log p_i(\theta) \right| \quad \text{a.s.}
\]

\[
\sup_{\theta \in N(\theta_0)} \left| \sum_{i \neq 1, k} \frac{N_i}{n} \left\{ 2 \left( \frac{f^{(1)}(\xi_i^{2\theta})}{f(\xi_i^{1\theta})} \right)^3 - 3 \frac{f^{(1)}(\xi_i^{2\theta}) f^{(2)}(\xi_i^{3\theta})}{f(\xi_i^{1\theta})^2} + \frac{f^{(3)}(\xi_i^{4\theta})}{f(\xi_i^{1\theta})} \right\} \right| .
\] (F.29)

We make the abbreviations

\[
A_i(\theta) := 2 \left( \frac{f^{(1)}(\xi_i^{2\theta})}{f(\xi_i^{1\theta})} \right)^3 , \quad B_i(\theta) := -3 \frac{f^{(1)}(\xi_i^{2\theta}) f^{(2)}(\xi_i^{3\theta})}{f(\xi_i^{1\theta})^2} , \quad C_i(\theta) := \frac{f^{(3)}(\xi_i^{4\theta})}{f(\xi_i^{1\theta})}
\]

so that (F.29) is
\[
= \sup_{\theta \in N(\theta_0)} \left| \sum_{i \neq 1, k} \frac{N_i}{n} \left\{ A_i(\theta) + B_i(\theta) + C_i(\theta) \right\} \right|
\]
\[
\leq \sum_{i \neq 1, k} \frac{N_i}{n} \left\{ \sup_{\theta \in N(\theta_0)} |A_i(\theta)| + \sup_{\theta \in N(\theta_0)} |B_i(\theta)| + \sup_{\theta \in N(\theta_0)} |C_i(\theta)| \right\} \quad .
\]

(\text{F.30})

Now, \(^{27}\)

\[
\sup_{\theta} |A_i(\theta)| = 2 \sup_{\theta} \left| \frac{f^{(1)}(\xi_i^{2\theta})}{f(\xi_i^{\theta})} \right|^3 \leq 2 \left( \sup_{\theta} \left| \frac{f^{(1)}(\xi_i^{2\theta})}{f^{(1)}(\xi_i^{\theta})} \right| \right)^3
\]

(\text{F.31})

where the inequality follows from (repeated application of) Lemma F.1. Continuing, the constants \(K_{1r}, \ r = 1, 2, 3,\) exist by the continuity and compactness assumptions.

In addition, we assumed that \(K_{10}\) exists, so (\text{F.31}) is

\[
\leq 2 \left( \sup_{\theta} \sup_{x,y \in [a,b]} \left| \frac{f^{(1)}(x,\theta)}{f(y,\theta)} \right| \right)^3 \leq 2 \left( \frac{K_{11}}{K_{10}} \right)^3
\]

Similarly,

\[
\sup_{\theta} |C_i(\theta)| = \sup_{\theta} \left| \frac{f^{(3)}(\xi_i^{4\theta})}{f(\xi_i^{\theta})} \right| \leq \sup_{\theta} \sup_{x,y \in [a,b]} \left| \frac{f^{(3)}(x,\theta)}{f(y,\theta)} \right| \leq \frac{K_{13}}{K_{10}}
\]

And in a similar fashion,

\[
\sup_{\theta} |B_i(\theta)| = 3 \sup_{\theta} \left| \frac{f^{(1)}(\xi_i^{2\theta}) f^{(2)}(\xi_i^{3\theta})}{f(\xi_i^{\theta}) f(\xi_i^{\theta})} \right| \leq 3 \sup_{\theta} \left| \frac{f^{(1)}(\xi_i^{2\theta})}{f(\xi_i^{\theta})} \right| \sup_{\theta} \left| \frac{f^{(2)}(\xi_i^{3\theta})}{f(\xi_i^{\theta})} \right| \quad (\text{F.32})
\]

\(^{27}\)Abbreviating \(\sup_{\theta \in N(\theta_0)}\) with \(\sup_{\theta}.\)
\[ \leq 3 \left( \sup_{\theta} \sup_{x,y \in [a,b]} \left| \frac{f^{(1)}(x, \theta)}{f(y, \theta)} \right| \right) \left( \sup_{\theta} \sup_{x,y \in [a,b]} \left| \frac{f^{(2)}(x, \theta)}{f(y, \theta)} \right| \right) \leq 3 \frac{K_{11} K_{12}}{K_{10} K_{10}}, \]

where the inequality in (F.32) follows from Lemma F.1. Combining all these inequalities for \( A_i(\theta), B_i(\theta) \) and \( C_i(\theta) \), (F.30) is

\[ a.s. \sum_{i \neq 1,k} \frac{N_i}{n} \left\{ 2 \left( \frac{K_{11}}{K_{10}} \right)^3 + \frac{K_{13}}{K_{10}} + 3 \frac{K_{11} K_{12}}{K_{10}^2} \right\} \leq 2 \left( \frac{K_{11}}{K_{10}} \right)^3 + \frac{K_{13}}{K_{10}} + 3 \frac{K_{11} K_{12}}{K_{10}^2} \]

and the result follows. □□□

**Result F.5** Suppose that the infinite bins remain fixed (see footnote 18). Further, suppose that there is a neighborhood \( N(\theta_0) \) such that \( \forall \theta \in N(\theta_0) \) and for any choice of \( \alpha \in \mathbb{R} \) (\( \alpha \) finite)

\[ \frac{\partial^r}{\partial \theta^r} \int_{-\infty}^{\alpha} f(x, \theta) dx = \int_{-\infty}^{\alpha} \frac{\partial^r}{\partial \theta^r} f(x, \theta) dx, \quad r = 1, 2, 3, \quad (F.33) \]

and similarly for \( [\alpha, \infty) \). Also suppose that there exist finite constants \( K_{2r}, r = 0, 1, 2, 3, \) such that for \( r = 1, 2, \) and 3,

\[ \sup_{\theta \in N(\theta_0)} \left| \int_{-\infty}^{\alpha} f^{(r)}(x, \theta) dx \right| \leq K_{2r} \quad \text{and} \quad \sup_{\theta \in N(\theta_0)} \left| \int_{\alpha}^{\infty} f^{(r)}(x, \theta) dx \right| \leq K_{2r}, \]

and

\[ \inf_{\theta \in N(\theta_0)} \int_{-\infty}^{\alpha} f(x, \theta) dx \geq K_{20} > 0 \quad \text{and} \quad \inf_{\theta \in N(\theta_0)} \int_{\alpha}^{\infty} f(x, \theta) dx \geq K_{20} > 0, \]

where \( f^{(r)}(x, \theta) := \frac{\partial^r}{\partial \theta^r} f(x, \theta) \). Then\(^{28}\)

\(^{28}\)Recall that the notation \( i = 1, k \) and \( i \neq 1, k \) is shorthand, see footnote 10.
\[
\sup_{\theta \in \mathcal{N}(\theta_0)} \left| \sum_{i=1, k} \frac{N_i}{n} \frac{\partial^3}{\partial \theta^3} \log p_i(\theta) \right| \quad \text{is eventually bounded a.s.}
\]

**Proof.** Let \( s_i(\theta) := \frac{\partial^3}{\partial \theta^3} \log p_i(\theta) \). We consider the case of \((\infty, a]\),\(^{29}\) the other case being similar. First\(^{30}\)

\[
\sup_{\theta} \left| \sum_{i=1, k} s_i(\theta) \frac{N_i}{n} \right|^{a.s.} \leq \sum_{i=1, k} N_i \sup_{\theta} |s_i(\theta)|.
\]

As in the proof of Result F.4, it follows by rules of differentiation that

\[
\sup_{\theta} |s_1(\theta)| = \sup_{\theta} \left| 2 \left( \frac{\partial}{\partial \theta} p_1(\theta) \right)^3 - 3 \frac{\partial^2}{\partial \theta^2} p_1(\theta) \frac{\partial^3}{\partial \theta^3} p_1(\theta) p_1(\theta)^2 + \frac{\partial^3}{\partial \theta^3} p_1(\theta) \frac{\partial^3}{\partial \theta^3} p_1(\theta) \right|
\]

\[
= \sup_{\theta} |E_1(\theta) + F_1(\theta) + G_1(\theta)| \leq \sup_{\theta} |E_1(\theta)| + \sup_{\theta} |F_1(\theta)| + \sup_{\theta} |G_1(\theta)|, \quad (F.34)
\]

where \(E_1(\theta), F_1(\theta),\) and \(G_1(\theta)\) are defined in the obvious fashion. Now

\[
\sup_{\theta} |E_1(\theta)| = \sup_{\theta} \left| \frac{\partial}{\partial \theta} p_1(\theta) \right|^3 \leq \left( \sup_{\theta} \left| \frac{\partial}{\partial \theta} p_1(\theta) \right| \right)^3
\]

\[
= \left( \sup_{\theta} \left| \frac{\int_{-\infty}^{a} \frac{\partial}{\partial \theta} f(x, \theta) dx}{\int_{-\infty}^{a} f(x, \theta) dx} \right| \right)^3 \leq \left( \frac{K_21}{K_20} \right)^3,
\]

\(^{29}\)See footnote 10 for explanation of the notation \( i = 1, I_1 = (\infty, a], \) etc.

\(^{30}\)We continue to abbreviate \( \sup_{\theta \in \mathcal{N}(\theta_0)} \) with \( \sup_{\theta} \).
where the first inequality follows by Lemma F.1 and the last equality follows by (F.33). Thus
\[ \frac{N_1}{n} \sup_{\theta} |E_1(\theta)| \overset{a.s.}{\leq} \left( \frac{K_{21}}{K_{20}} \right)^3. \]

Similar arguments work for \( \sup_{\theta} |F_1(\theta)| \) and \( \sup_{\theta} |G_1(\theta)| \). Thus \( \exists K \) such that
\[ \frac{N_1}{n} \left( \sup_{\theta} |E_1(\theta)| + \sup_{\theta} |F_1(\theta)| + \sup_{\theta} |G_1(\theta)| \right) \overset{a.s.}{\leq} K. \]

This establishes the result. \( \Box \)

**Result F.6** If the assumptions of Results F.4 and F.5 hold, then
\[ \sup_{\theta \in N(\theta_0)} \left| \frac{1}{n} g'''(\theta) \right| = \sup_{\theta \in N(\theta_0)} \left| \sum_{i=1}^{k} \frac{N_i}{n} \frac{\partial^3}{\partial \theta^3} \log p_i(\theta) \right| \text{ is eventually bounded a.s.} \]

**Proof.** This is obvious. \( \Box \)

**Result F.7** Suppose that the assumptions in Results F.1, F.4 and F.5 hold. Then (using the notation from section F.2) it follows that \( \frac{1}{n} g'''(\theta_n^*) \) is eventually bounded in probability.

**Proof.** Under the assumptions of Results F.4 and F.5, we may apply Result F.6 to obtain that (take \( N(\theta_0) \) as the largest neighborhood that will satisfy all the assumptions)
\[
\sup_{\theta \in N(\theta_0)} \left| \frac{1}{n} g'''(\theta) \right| \quad (F.35)
\]

is eventually bounded a.s. Restating this formally for later use, if (F.35) is eventually bounded a.s., then \( \exists K \) and an \( n_1 \) such that except possibly on a set of 0 probability

\[
\sup_{\theta \in N(\theta_0)} \left| \frac{1}{n} g'''(\theta) \right| \leq K \quad \forall n \geq n_1 \quad . \quad (F.36)
\]

Under the assumptions of Result F.1, we established in Section F.3 that \( \hat{\theta}_n \xrightarrow{p} \theta_0 \).

Pick any \( \epsilon_0 > 0 \) such that \( B(\theta_0, \epsilon_0) \), the ball centered at \( \theta_0 \) of radius \( \epsilon_0 \), is contained in \( N(\theta_0) \). Then

\[
\left\{ \hat{\theta}_n \in B(\theta_0, \epsilon_0) \right\} \subseteq \left\{ \hat{\theta}_n \in N(\theta_0) \right\}
\]

which implies that

\[
P \left\{ \hat{\theta}_n \in B(\theta_0, \epsilon_0) \right\} \leq P \left\{ \hat{\theta}_n \in N(\theta_0) \right\} \quad . \quad (F.37)
\]

Choose \( \eta > 0 \). There is an \( n_\eta \) such that

\[
P \left\{ \hat{\theta}_n \in B(\theta_0, \epsilon_0) \right\} \geq 1 - \eta \quad , \quad \forall n \geq n_\eta
\]

hence by (F.37)

\[
P \left\{ \hat{\theta}_n \in N(\theta_0) \right\} \geq 1 - \eta \quad , \quad \forall n \geq n_\eta \quad . \quad (F.38)
\]

Continuing in this formal manner, since \( \theta_n^* \in (\theta_0, \hat{\theta}_n) \),
\[ \{ \hat{\theta}_n \in N(\theta_0) \} \subseteq \left\{ \frac{1}{n} g''''(\hat{\theta}_n^*) \leq \sup_{\theta \in N(\theta_0)} \left| \frac{1}{n} g''''(\theta) \right| \right\} \]

hence it follows that

\[ P \left\{ \hat{\theta}_n \in N(\theta_0) \right\} \leq P \left\{ \frac{1}{n} g''''(\hat{\theta}_n^*) \leq \sup_{\theta \in N(\theta_0)} \left| \frac{1}{n} g''''(\theta) \right| \right\} , \quad (F.39) \]

so that

\[ P \left\{ \left| \frac{1}{n} g''''(\hat{\theta}_n^*) \right| \leq \sup_{\theta \in N(\theta_0)} \left| \frac{1}{n} g''''(\theta) \right| \right\} \geq 1 - \eta \quad \forall \ n \geq n_\eta \quad (F.40) \]

follows from (F.38) and (F.39). Let \( n'_\eta := \max \{ n_1, n_\eta \} \) (notice that \( n_1 \) does not depend upon \( \eta \)). Putting (F.36) and (F.40) together, this gives

\[ P \left\{ \left| \frac{1}{n} g''''(\hat{\theta}_n^*) \right| \leq K \right\} \geq 1 - \eta \quad \forall \ n \geq n'_\eta , \]

as desired. \( \square \)

**Result F.8** Suppose that the assumptions in Results F.1, F.4 and F.5 hold. Then (using the notation in section F.2) it follows that

\[ \frac{1}{2n} \left( \hat{\theta}_n - \theta_0 \right) g''''(\theta_0^*) = o_p(1) . \]

**Proof.** From Result F.7, we have that \( \hat{\theta}_n \xrightarrow{p} \theta_0 \) and that \( \frac{1}{n} g''''(\theta_0^*) \) is eventually bounded in probability. The result follows. \( \square \)
F.7 Proof of the identity $\sum_{i=1}^{k} g_i^2 p_i = - \sum_{i=1}^{k} h_i p_i$

Result F.9

$$\sum_{i=1}^{k} \left( \frac{\partial}{\partial \theta} \log p_i \right)^2 p_i = - \sum_{i=1}^{k} p_i \frac{\partial^2}{\partial \theta^2} \log p_i$$ \hspace{1cm} (F.41)

where\(^{31}\) we abbreviate $p_i(\theta_0)$ with $p_i$.

Proof. This argument is similar to the classical argument which establishes that

$$E \left[ \left( \frac{\partial}{\partial \theta} \log f(X, \theta) \right)^2 \right] = - E \left[ \frac{\partial^2}{\partial \theta^2} \log f(X, \theta_0) \right]$$ \hspace{1cm} (F.42)

provided that the necessary interchanges are valid. Recall from (F.13) that

$$\sum_{i=1}^{k} \frac{\partial}{\partial \theta} p_i = 0$$ \hspace{1cm} (F.43)

Also recall that this fact does not require any interchanges to be valid (unlike (F.14)).

Now (F.43) is equivalent to

$$\sum_{i=1}^{k} \left( \frac{1}{p_i} \frac{\partial}{\partial \theta} p_i \right) p_i = 0$$ \hspace{1cm} (F.44)

and taking another derivative of (F.44),

$$\Rightarrow \sum_{i=1}^{k} \frac{\partial}{\partial \theta} \left( \left( \frac{1}{p_i} \frac{\partial}{\partial \theta} p_i \right) p_i \right) = 0$$

$$\iff \sum_{i=1}^{k} \left\{ \frac{1}{p_i} \frac{\partial}{\partial \theta} p_i \frac{\partial}{\partial \theta} p_i + \frac{\partial}{\partial \theta} \left( \frac{1}{p_i} \frac{\partial}{\partial \theta} p_i \right) p_i \right\} = 0$$

\(^{31}\)Recall that in our notation (see Section F.2) (F.41) is equivalent to $\sum_{i=1}^{k} g_i^2 p_i = - \sum_{i=1}^{k} h_i p_i$.\m
\[ \iff \sum_{i=1}^{k} \left\{ \frac{1}{p_i} \left( \frac{\partial}{\partial \theta} p_i \right)^2 + p_i \frac{\partial^2}{\partial \theta^2} \log p_i \right\} = 0 \]

\[ \iff \sum_{i=1}^{k} \left\{ \left( \frac{\partial}{\partial \theta} \log p_i \right)^2 p_i + \sum_{i=1}^{k} p_i \frac{\partial^2}{\partial \theta^2} \log p_i \right\} = 0 \]

\[ \square \square \square \]

**F.8** \[ \sum_{i=1}^{k} g_i^2 p_i \] is bounded below

**Result F.10** Suppose that the infinite bins remain fixed (see footnote 18). In addition, suppose at least one of the following holds

1. For any choice of \( \alpha \in \mathbb{R} \) (\( \alpha \) finite),

\[ \frac{\partial}{\partial \theta} \int_{-\infty}^{\alpha} f(x, \theta_0) dx = \int_{-\infty}^{\alpha} \frac{\partial}{\partial \theta} f(x, \theta_0) dx , \quad (F.45) \]

and similarly for \([\alpha, \infty)\). Either

\[ \left| \int_{-\infty}^{\alpha} \frac{\partial}{\partial \theta} f(x, \theta_0) dx \right| > 0 \quad \text{and} \quad \int_{-\infty}^{\alpha} f(x, \theta_0) dx > 0 , \quad \text{or} \quad (F.46) \]

\[ \left| \int_{b}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta_0) dx \right| > 0 \quad \text{and} \quad \int_{b}^{\infty} f(x, \theta_0) dx > 0 . \quad (F.47) \]
2. \( f(x, \theta_0) \) and \( \frac{\partial}{\partial \theta} f(x, \theta_0) \) are continuous, \( \int_a^b f(x, \theta_0) dx > 0 \) and there exists \( K_{41} > 0 \) such that

\[
\left| \frac{\partial}{\partial \theta} f(x, \theta_0) \right| \geq K_{41} , \ \forall x \in [a, b] .
\]

Then there exists a \( K_{40} < \infty \) such that

\[
f(x, \theta_0) \leq K_{40} , \ \forall x \in [a, b]
\]

and

\[
\sum_{i=1}^k g_i^2 p_i := \sum_{i=1}^k \left( \frac{\partial}{\partial \theta} \log p_i \right)^2 p_i \text{ is bounded below.} \quad (F.48)
\]

**Proof.** If 1. holds, then either (F.46) or (F.47) holds. Suppose (F.46) holds (if (F.47) holds, the argument is identical) then

\[
g_i^2 p_i = \left( \frac{\partial}{\partial \theta} p_i \right) p_i = \left( \frac{\int_{-\infty}^a \frac{\partial}{\partial \theta} f(x, \theta_0) dx}{p_i} \right)^2 p_i \geq c > 0
\]

for some \( c \) independently of \( k \), since the infinite bins remain fixed,\(^{32}\) where the interchange is valid by (F.45). Thus

\[
\sum_{i=1}^k g_i^2 p_i \geq g_1^2 p_1 \geq c > 0 . \quad (F.49)
\]

\(^{32}\)For notation here, see footnotes 10 and 18.
and (F.48) follows.

On the other hand, if 2. holds then

\[
\sum_{i \neq 1,k} g_i^2 p_i = \sum_{i \neq 1,k} \left( \frac{\partial f(x_i)}{\partial \theta} \right)^2 p_i = \sum_{i \neq 1,k} \left( \frac{\partial f(x_i)}{\partial \theta} \frac{f(x_i^{(2)})}{f(x_i^{(1)})} \right)^2 p_i 
\]

\[
\geq \left( \frac{K_{41}}{K_{40}} \right)^2 \sum_{i \neq 1,k} p_i = \left( \frac{K_{41}}{K_{40}} \right)^2 \int_a^b f(x, \theta_0) dx > 0
\]

which holds independently of \( k \). Notice that the interchange is valid by continuity of \( f(x, \theta_0) \) and \( \frac{\partial}{\partial \theta} f(x, \theta_0) \) (cf. Theorem 5.1, Trench [66], p. 576). Also the existence of the \( \eta_i^{(r)} \)'s, \( r = 1, 2 \), is guaranteed by the MVT for Integrals (we have made this same argument repeatedly). Thus (F.48) also follows from 2.

\[\square\square\square\]

**Remark:** While we first believed that 2. would be the simpler assumption to use (in our main theorems later on), it so happens that we cannot expect 2. to hold in great generality. Apparently 1. will hold in greater generality, hence that will be the assumption we use.

**F.9 Main Results (Lemmas)**

**Lemma F.2** Suppose that\(^{33}\)

1. *(Identifiability)* \( \forall \delta > 0, \exists k_0 \) and \( \eta_1, \eta_2 > 0 \) small enough that

\[
\sum_{i=1}^k (p_i(\theta_0) - \eta_i) \log \frac{p_i(\theta_0)}{p_i(\theta)} \geq \eta_2 \quad \forall k \geq k_0, \forall \theta \in \{ |\theta - \theta_0| \geq \delta \}
\]

\(^{33}\)We must also require Assumptions F.1 - F.3.
2. The infinite bins remain fixed (see footnotes 10 and 18),

3. \( n, k \to \infty \) and

(a) \( n \min_i p_i(\theta_0) \to \infty \),

(b) \( p_i(\theta), i = 1, \ldots, k \), are continuous functions of \( \theta \),

(c) \( k/n \max_i p_i(\theta_0) \to 0 \), and

(d)

\[
\frac{k}{\sqrt{n}} \max_i \frac{1}{\sqrt{p_i(\theta_0)}} \to 0 ,
\]

4. There is a neighborhood\(^{34}\) of \( \theta_0 \), say \( N(\theta_0) \), such that \( f(x, \theta) \) and \( f^{(r)}(x, \theta) \), \( r = 1, 2, 3 \), are continuous on \([a, b] \times N(\theta_0)\)^{35} (where \( f^{(r)}(x, \theta) = \frac{\partial^r}{\partial \theta^r} f(x, \theta) \)) and for any choice of \( \alpha \in \mathbb{R} \) (finite) and \( \forall \theta \in N(\theta_0) \)

\[
\frac{\partial^r}{\partial \theta^r} \int_{-\infty}^{\alpha} f(x, \theta)dx = \int_{-\infty}^{\alpha} \frac{\partial^r}{\partial \theta^r} f(x, \theta)dx \quad r = 1, 2, 3 ,
\]

and similarly for \([\alpha, \infty)\),

5. (a) There is a neighborhood of \( \theta_0 \), say \( N(\theta_0) \), and a finite constant \( K_{10} \) such that

\[
f(x, \theta) \geq K_{10} > 0 \quad \forall \theta \in N(\theta_0) \quad \forall x \in [a, b] . \tag{F.50}
\]

\(^{34}\)In this assumption it is implicitly assumed that Assumption F.3, i.e. one of the technical assumptions, holds. This was mentioned in Section F.1.

\(^{35}\)See footnote 24 about the notation "\( N(\theta_0) \)".
(b) There is a neighborhood of \( \theta_0 \), say \( N(\theta_0) \), and finite constants \( K_{2r} \), \( r = 0, 1, 2, 3 \), such that for \( r = 1, 2, 3 \),

\[
\sup_{\theta \in N(\theta_0)} \left| \int_{-\infty}^{a} f^{(r)}(x, \theta) \, dx \right| \leq K_{2r}, \quad (F.51)
\]

\[
\sup_{\theta \in N(\theta_0)} \left| \int_{b}^{\infty} f^{(r)}(x, \theta) \, dx \right| \leq K_{2r}, \quad (F.52)
\]

and

\[
\inf_{\theta \in N(\theta_0)} \int_{-\infty}^{a} f(x, \theta) \, dx \geq K_{20} > 0, \quad (F.53)
\]

\[
\inf_{\theta \in N(\theta_0)} \int_{b}^{\infty} f(x, \theta) \, dx \geq K_{20} > 0, \quad (F.54)
\]

6. Either

\[
\left| \int_{-\infty}^{a} \frac{\partial}{\partial \theta} f(x, \theta_0) \, dx \right| > 0 \quad \text{and} \quad \int_{-\infty}^{a} f(x, \theta_0) \, dx > 0, \quad (F.55)
\]

or

\[
\left| \int_{b}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta_0) \, dx \right| > 0 \quad \text{and} \quad \int_{b}^{\infty} f(x, \theta_0) \, dx > 0. \quad (F.56)
\]

then
\[
\sqrt{\sum_{i=1}^{k} g_i^2 p_i \sqrt{n} \left( \theta_n - \theta_0 \right)} = \sqrt{\sum_{i=1}^{k} \left( \frac{\partial \log p_i(\theta_0)}{\partial \theta} \right)^2 p_i(\theta_0) \sqrt{n} \left( \hat{\theta}_n - \theta_0 \right)} \xrightarrow{\mathcal{D}} N(0,1) ,
\]

where
\[
\hat{\theta}_n = \arg \max_{\theta \in \Theta} \sum_{i=1}^{k} \frac{N_i}{n} \log p_i(\theta)
\]

**Proof.** Recall from the discussion in Section F.2 that

\[
\sqrt{n} \left( \hat{\theta}_n - \theta_0 \right) = \frac{\frac{1}{\sqrt{n}} g'(\theta_0)}{-\frac{1}{n} g''(\theta_0) - \frac{1}{2n} \left( \hat{\theta}_n - \theta_0 \right) g'''(\theta^*_n)}
\]

(F.57)

where \( \theta^*_n \in (\theta_0, \hat{\theta}_n) \) and \( g(\theta) := \sum_{i=1}^{k} N_i \log p_i(\theta) \). By Lemma F.2, assumption 3a., we may apply Result F.2 to obtain

\[
\left( \sum_{i=1}^{k} g_i^2 p_i \right)^{-1/2} \frac{1}{\sqrt{n}} g'(\theta_0) \xrightarrow{\mathcal{D}} N(0,1) .
\]

(F.58)

Also recall from Section F.2 that

\[-\frac{1}{n} g''(\theta_0) = -\sum_{i=1}^{k} h_i p_i + \left( \sum_{i=1}^{k} h_i p_i - \frac{1}{n} g''(\theta_0) \right) .\]

By Lemma F.2, assumptions 3d., 4., 5a., and 5b., we may apply Result F.3 to get that \( \sum_{i=1}^{k} h_i p_i - \frac{1}{n} g''(\theta_0) \xrightarrow{\mathcal{L}} 0 \), hence

\[
\sum_{i=1}^{k} h_i p_i - \frac{1}{n} g''(\theta_0) = o_p(1) .
\]
By Result F.9, \(- \sum_{i=1}^{k} h_i p_i = \sum_{i=1}^{k} g_i^2 p_i\), so that

\[- \frac{1}{n} g''(\theta_0) = \sum_{i=1}^{k} g_i^2 p_i + o_p(1) \, .\]

Further, by Lemma F.2, assumptions 1., 2., 3b., 3c., 4., 5a., and 5b., we may apply Result F.8 to get

\[- \frac{1}{2n} \left( \hat{\theta}_n - \theta_0 \right) g'''(\theta_n^*) = o_p(1) \, .\]

By Lemma F.2, assumption 6., apply Result F.10 to get that \(\left( \sum_{i=1}^{k} g_i^2 p_i \right)^{-1}\) is bounded, hence

\[
\frac{1}{\sum_{i=1}^{k} g_i^2 p_i} \left\{ \frac{1}{n} g''(\theta_0) - \frac{1}{2n} \left( \hat{\theta}_n - \theta_0 \right) g'''(\theta_n^*) \right\} =
\]

\[
\frac{1}{\sum_{i=1}^{k} g_i^2 p_i} \left\{ \sum_{i=1}^{k} g_i^2 p_i + o_p(1) \right\} = 1 + o_p(1) \, . \tag{F.59}
\]

Finally, multiplying both sides of (F.57) by \(\sqrt{\sum_{i=1}^{k} g_i^2 p_i}\), we get

\[
\frac{1}{\sqrt{\sum_{i=1}^{k} g_i^2 p_i}} \frac{1}{\sqrt{n}} \left( \hat{\theta}_n - \theta_0 \right) = \frac{1}{\sqrt{\sum_{i=1}^{k} g_i^2 p_i}} \frac{1}{\sqrt{n}} g'(\theta_0)
\]

\[
\left\{ \frac{1}{\sum_{i=1}^{k} g_i^2 p_i} \left\{ \frac{1}{n} g''(\theta_0) - \frac{1}{2n} \left( \hat{\theta}_n - \theta_0 \right) g'''(\theta_n^*) \right\} \right\} =
\]

\[
\{ 1 + o_p(1) \}^{-1} \frac{1}{\sqrt{\sum_{i=1}^{k} g_i^2 p_i}} \frac{1}{\sqrt{n}} g'(\theta_0) \overset{D}{\to} N(0, 1) \, .
\]

By (F.58) and Slutsky’s Theorem.

\[\square\square\square\]

It appears to be to our advantage to assume a little more than we did for Lemma F.2 in order to simplify some of the assumptions. We now prove Results F.11 and F.12,
which will help us to establish Lemma F.3, a somewhat simpler version of Lemma F.2.

**Result F.11** Suppose that there is a neighborhood $N(\theta_0)$ of $\theta_0$ such that $f(x, \theta)$ and $\frac{\partial^r}{\partial \theta^r} f(x, \theta)$, $r = 1, 2, 3$ are all continuous on $\mathbb{R} \times \overline{N(\theta_0)}$. Suppose further that there exist functions $H_r(x)$, $r = 1, 2, 3$, such that

$$\left| \frac{\partial^r}{\partial \theta^r} f(x, \theta) \right| \leq H_r(x) , \quad \forall \theta \in \overline{N(\theta_0)} , \quad r = 1, 2, 3 , \quad \text{and} \quad \int H_r(x) dx < \infty ,$$

then it follows that

1. $$\frac{\partial^r}{\partial \theta^r} \int_D f(x, \theta) dx = \int_D \frac{\partial^r}{\partial \theta^r} f(x, \theta) dx , \quad \forall \theta \in \overline{N(\theta_0)} , \quad r = 1, 2, 3 ,$$

where $D$ is any interval of one of the following forms (where $c$, $d$ and $\alpha$ are finite):

$$[c, d] , \ (\alpha, \alpha] , \ [\alpha, \infty) , \ \text{or} \ (\alpha, \infty) , \quad (F.60)$$

2. There are finite constants $K_{2r}$, $r = 0, 1, 2, 3$, such that

$$\sup \left\{ \int_{-\infty}^a f^{(r)}(x, \theta) dx \right\} \leq K_{2r} , \quad r = 1, 2, 3 ,$$

and

---

36See footnote 24 about the notation $\overline{N(\theta_0)}$.
\[
\sup_{\theta \in N(\theta_0)} \left| \int_{b}^{\infty} f^{(r)}(x, \theta) dx \right| \leq K_{2r}, \quad r = 1, 2, 3,
\]

where \( f^{(r)}(x, \theta) = \frac{\partial^r}{\partial \theta^r} f(x, \theta) \).

**Proof.** We consider the case of \( r = 1 \). The other cases are similar.\(^{37}\) If \( D = [c, d] \) (\( c, d \) finite) is a closed interval in \( \mathbb{R} \), then the interchange follows immediately by the continuity assumptions (apply Theorem 5.1, Trench [66], p. 576).

Suppose \( D = [\alpha, \infty) \), some finite \( \alpha \). We proceed now to verify that all the assumptions of Theorem 5.6, Trench [66], p. 586, hold in our situation, our immediate goal being to apply that theorem in order to establish (F.62). Clearly, \( \int_{\alpha}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta) dx \) exists and is finite for some \( \theta \in N(\theta_0) \) (because \( \int_{\alpha}^{\infty} \left| \frac{\partial}{\partial \theta} f(x, \theta) \right| dx \leq \int H_1(x) dx < \infty \), \( \forall \theta \in N(\theta_0) \)). We recall the definition of the improper integral

**Definition F.1** (Trench [66], p. 157) Suppose \( f \) is locally integrable on \([r, s)\), then

\(^{37}\)This proof consists mostly of verifying that the hypotheses in several advanced calculus theorems are satisfied.

\(^{38}\)Without all the details, one important difference for the other cases is that, e.g. for \( r = 2 \), \( I_1 = [\alpha, \infty) \), we must be able to show that

\[
\int_{\alpha}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta) dx \tag{F.61}
\]

converges for some \( \theta \in \overline{N(\theta_0)} \), which does not follow as automatically as it does for the \( r = 1 \) case. When

\[
\int_{\alpha}^{\infty} f(x, \theta) dx
\]

must converge because \( f(x, \theta) \) is a density function. However, the convergence in (F.61) follows readily from the existence of \( H_1(x) \).
\[
\int_r^s f(x)dx = \lim_{t \to s^-} \int_r^t f(x)dx
\]

if this limit exists and is finite.

And one may recall that \( f \) is locally integrable on \([\alpha, \infty)\) if \( f \) is integrable on every finite closed subinterval of \([\alpha, \infty)\). Since we already know that \( \int_\alpha^\infty \frac{\partial}{\partial \theta} f(x, \theta)dx \) exists (and is finite), this implies that \( \frac{\partial}{\partial \theta} f(x, \theta) \) is locally integrable on \([\alpha, \infty)\) (since otherwise the improper integral is either (i) not finite or (ii) does not exist). Having demonstrated this, we observe further that \( H_1(x) \) is exactly the function we need in order to successfully apply the Weierstrass Convergence Test (Trench [66], Theorem 5.2, p. 581). Apply that theorem to get

\[
\int_\alpha^\infty \frac{\partial}{\partial \theta} f(x, \theta)dx \quad \text{converges uniformly on } \overline{N(\theta_0)} .
\]

Since \( f(x, \theta) \) and \( f^{(1)}(x, \theta) \) are continuous on \( \mathbb{R} \times \overline{N(\theta_0)} \), they are continuous on \([\alpha, \infty) \times \overline{N(\theta_0)}\). The integral \( \int_\alpha^\infty f(x, \theta)dx \) converges for all \( \theta \in \overline{N(\theta_0)} \), even though it need only converge for one such \( \theta \) for us to use Theorem 5.6, Trench [66]. Hence we have verified that all the assumptions of Theorem 5.6, Trench [66], p. 586, hold.

Apply that theorem to get

\[
\frac{\partial}{\partial \theta} \int_\alpha^\infty f(x, \theta)dx = \int_\alpha^\infty \frac{\partial}{\partial \theta} f(x, \theta)dx .
\]

\[ (F.62) \]

\[ 39 \text{Of course, it converges } \forall \theta \in \Theta, \text{ but we don't need that much here.} \]
This completes the argument for $[\alpha, \infty)$. One may use a similar argument for $(-\infty, \alpha]$.

Also, $\forall \theta \in \overline{N(\theta_0)}$

$$\frac{\partial}{\partial \theta} \int_{-\infty}^{\alpha} f(x, \theta)dx = \frac{\partial}{\partial \theta} \left\{ \int_{-\infty}^{\alpha} f(x, \theta)dx + \int_{\alpha}^{\infty} f(x, \theta)dx \right\}$$

$$= \frac{\partial}{\partial \theta} \int_{-\infty}^{\alpha} f(x, \theta)dx + \frac{\partial}{\partial \theta} \int_{\alpha}^{\infty} f(x, \theta)dx$$  \hspace{1cm} (F.63)

where the first equality follows by the definition of $f_{-\infty}^{\alpha}$ (i.e. $\alpha$ is finite, but arbitrary) and the second equality follows by the usual sum rule for differentiation\(^{40}\) Continuing, by (F.62), (F.63) is

$$= \int_{-\infty}^{\alpha} \frac{\partial}{\partial \theta} f(x, \theta)dx + \int_{-\infty}^{\alpha} \frac{\partial}{\partial \theta} f(x, \theta)dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta)dx.$$  \hspace{1cm} .

This establishes conclusion 1. of Result F.11.

Lastly,

$$\left| \int_{-\infty}^{\alpha} f^{(r)}(x, \theta)dx \right| \leq \int_{-\infty}^{\alpha} \left| f^{(r)}(x, \theta)dx \right| dx \leq \int_{-\infty}^{\alpha} H_r(x)dx < \infty$$

so that there are finite constants $K_{2r}$, $r = 0, 1, 2, 3, \ldots$\(^{41}\) such that

$$\sup_{\theta \in \overline{N(\theta_0)}} \left| \int_{-\infty}^{\alpha} f^{(r)}(x, \theta)dx \right| \leq K_{2r}$$

and similarly for $\int_{\alpha}^{\infty}$. This establishes 2. of Result F.11.

\(^{40}\)Note that this step would not be valid unless we had proved differentiability at (F.62) first. Since we have demonstrated that $\int_{-\infty}^{\alpha} f(x, \theta)dx$ and $\int_{\alpha}^{\infty} f(x, \theta)dx$ are differentiable, applying the sum rule for derivatives is legal.

\(^{41}\)Namely, $K_{2r} = \int_{-\infty}^{\alpha} H_r(x)dx$. 
Result F.12 Suppose that the assumptions of Result F.11 hold. In addition, suppose that each $H_r(x)$ is bounded. Then

1.

$$
\frac{\partial^r}{\partial \theta^r} \int_D f(x, \theta) dx = \int_D \frac{\partial^r}{\partial \theta^r} f(x, \theta) dx \quad \forall \theta \in N(\theta_0) \quad , \quad r = 1, 2, 3 ,
$$

where $D$ is any interval of the form described in (F.60).

2. There are finite constants $K_{2r}$, $r = 0, 1, 2, 3$, such that for $r = 1, 2, 3$,

$$
\sup_{\theta \in N(\theta_0)} \left| \int_{-\infty}^{a} f^{(r)}(x, \theta) dx \right| \leq K_{2r} \quad \text{and} \quad \sup_{\theta \in N(\theta_0)} \left| \int_{b}^{\infty} f^{(r)}(x, \theta) dx \right| \leq K_{2r}
$$

where $f^{(r)}(x, \theta) = \frac{\partial^r}{\partial \theta^r} f(x, \theta)$,

3. There are finite constants $K_{1r}$, $r = 1, 2, 3$, such that

$$
\left| f^{(r)}(x, \theta) \right| \leq K_{1r} \quad , \quad \forall \theta \in N(\theta_0) \quad , \quad \forall x \in [a, b] \quad , \quad r = 1, 2, 3
$$

Proof. This is obvious.

Lemma F.3 Suppose that\textsuperscript{42}

1. (Identifiability) $\forall \delta > 0$, $\exists k_0$ and $\eta_1, \eta_2 > 0$ small enough that

$$
\sum_{i=1}^{k} (p_i(\theta_0) - \eta_1) \log \frac{p_i(\theta_0)}{p_i(\theta)} \geq \eta_2 \quad , \quad \forall k \geq k_0 \quad , \quad \forall \theta \in \{ |\theta - \theta_0| \geq \delta \}
$$

\textsuperscript{42}We must also require Assumptions F.1 and F.2.
2. The infinite bins remain fixed (see footnotes 10 and 18),

3. \( n, k \to \infty \) and

(a) \( n \min_i p_i(\theta_0) \to \infty \),

(b) \( p_i(\theta) \), \( i = 1, \ldots, k \), are continuous functions of \( \theta \),

(c) \( k/n \max_i p_i(\theta_0) \to 0 \), and

(d)

\[
\frac{k}{\sqrt{n}} \max_i \frac{1}{\sqrt{p_i(\theta_0)}} \to 0 ,
\]

4. Suppose that there is a neighborhood \( N(\theta_0) \) of \( \theta_0 \) such that \( f(x, \theta) \) and \( \frac{\partial^r}{\partial \theta^r} f(x, \theta) \), \( r = 1, 2, 3 \), are all continuous on \( \mathbb{R} \times \overline{N(\theta_0)} \).\(^{43}\) Suppose further that there exist functions \( H_r(x) \), \( r = 1, 2, 3 \), such that

\[
\left| \frac{\partial^r}{\partial \theta^r} f(x, \theta) \right| \leq H_r(x) , \ \forall \theta \in \overline{N(\theta_0)} , \ r = 1, 2, 3 \ \text{and} \ \int H_r(x)dx < \infty ,
\]

and that each \( H_r(x) \) is bounded,

5. (a) There exists a nbd \( N(\theta_0) \) and a constant \( K_{10} \) such that

\[
f(x, \theta) \geq K_{10} > 0 \ , \ \forall \theta \in N(\theta_0) \ , \ \forall x \in [a, b]
\]

(b) There exists a nbd \( N(\theta_0) \) and a constant \( K_{20} \) such that

\[
\inf_{\theta \in N(\theta_0)} \int_{-\infty}^{a} f(x, \theta)dx \geq K_{20} > 0 \ \text{and} \ \inf_{\theta \in N(\theta_0)} \int_{b}^{\infty} f(x, \theta)dx \geq K_{20} > 0
\]

\(^{43}\)See footnote 24 about the notation \( \overline{N(\theta_0)} \).
(c) Either
\[ \left| \int_{-\infty}^{a} \frac{\partial}{\partial \theta} f(x, \theta_0) dx \right| > 0 \quad \text{and} \quad \int_{-\infty}^{a} f(x, \theta_0) dx > 0, \quad (F.64) \]

or
\[ \left| \int_{b}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta_0) dx \right| > 0 \quad \text{and} \quad \int_{b}^{\infty} f(x, \theta_0) dx > 0. \quad (F.65) \]

then
\[ \sqrt{\sum_{i=1}^{k} g_i^2 p_i \sqrt{n}} (\hat{\theta}_n - \theta_0) = \sqrt{\sum_{i=1}^{k} \left( \frac{\partial}{\partial \theta} \log p_i(\theta_0) \right)^2 p_i(\theta_0) \sqrt{n}} (\hat{\theta}_n - \theta_0) \overset{D}{\rightarrow} N(0, 1), \]

where
\[ \hat{\theta}_n = \arg\max_{\theta \in \Theta} \sum_{i=1}^{k} \frac{N_i}{n} \log p_i(\theta) \]

Proof. It suffices to show that the assumptions of Lemma F.3 imply the assumptions of Lemma F.2. The assumptions 1., 2., and 3. in Lemma F.3 are unchanged from the corresponding assumptions 1., 2., and 3. of Lemma F.2, hence it is enough to show that assumptions 4. and 5. of Lemma F.3 imply assumptions 4.-6. of Lemma F.2.

If Lemma F.3, assumption 4. holds, then the hypotheses of Result F.12 are satisfied, hence we may apply that result. The conclusion of Result F.12 ensures that assumption 4. of Lemma F.2 is satisfied, along with (F.51) and (F.52). Hence it
remains to require that (F.50) and (F.53) and (F.54) hold, as well Lemma F.2 assumption 6. But that is exactly what we require in Lemma F.3, assumption 5. □□□

F.10 A Simplification (Our Main Theorem)

One may observe that in Lemmas F.2 and F.3, asking for conditions on rates involving \( p_i(\theta_0)'s \) ties our hands. The difficulty is not so much that we don't know the \( p_i(\theta_0)'s \), since a ready estimate is available via the observed proportions. However, if it is necessary for us to use the data to approximately satisfy the conditions on the \( p_i(\theta_0)'s \), it seems clear that we might destroy our (assumed) multinomial structure in the process.

As it turns out, we need make no additional assumptions beyond those we have already used in Lemma F.2 and Lemma F.3 in order to free ourselves from the difficulties just mentioned. We will state a new theorem, Theorem F.1, which avoids treating the \( p_i(\theta_0)'s \) directly. But first it is necessary to establish another result.

Result F.13 Suppose that the infinite bins remain fixed (see footnote 18), \( \exists K_{10} \) such that \( f(x, \theta_0) \geq K_{10} > 0 \) on \([a, b] \), and \( f(x, \theta_0) \) is continuous. Denote the bin width for \( I_i \) by \( \delta_i \), \( i \neq 1, k \). Suppose that \( \max_{i \neq 1, k} \delta_i \to 0 \),

\[ n \min_{i \neq 1, k} \delta_i \to \infty \quad \text{(F.66)} \]

and
\[
\frac{k}{\sqrt{n}} \max_{i \neq 1, k} \frac{1}{\sqrt{\delta_i}} \to 0
\] 

then \( n \min_i p_i(\theta_0) \to \infty \) and

\[
\frac{k}{\sqrt{n}} \max_i \frac{1}{\sqrt{p_i(\theta_0)}} \to 0.
\]

In addition, if we assume that \( k/n \to 0 \), then

\[
\frac{k}{n} \max_i p_i(\theta_0) \to 0.
\]

**Proof.** Observe that for \( i \neq 1, k \),

\[
p_i(\theta_0) = \int_{I_i} f(x, \theta_0) \, dx = f(\zeta_i, \theta_0) \delta_i \geq \delta_i K_{10}
\]

where \( \zeta_i \) lies in \( I_i \). Observe that (F.68) implies that

\[
\min_{i \neq 1, k} p_i(\theta_0) \geq K_{10} \min_{i \neq 1, k} \delta_i
\]

which implies that

\[
\frac{1}{\min_{i \neq 1, k} \sqrt{p_i(\theta_0)}} \leq \frac{1}{\sqrt{K_{10} \min_{i \neq 1, k} \sqrt{\delta_i}}}
\]

so

---

\(^{44}\)Recall that the notation \( i = 1, k \) and \( i \neq 1, k \) is shorthand, see footnote 10.
by (F.67). The infinite bins do not matter here because beyond a certain \( k \) and \( n \), the smallest \( p_i(\theta_0) \) must be one of the finite bins (due to the assumption that \( \max_{i \neq 1, k} \delta_i \to 0 \)). This is so because \( p_1(\theta_0) \) and \( p_k(\theta_0) \) remain constant (under the fixed-infinite bins assumption).

By (F.69)

\[
\frac{k}{\sqrt{n}} \max_{i \neq 1, k} \frac{1}{p_i(\theta_0)} \leq \frac{k}{\sqrt{n}} \frac{1}{\sqrt{K}} \max_{i \neq 1, k} \frac{1}{\sqrt{\delta_i}} \to 0 \tag{F.70}
\]

and the right hand side of (F.71) goes to infinity by (F.66). Notice again that we ignore the infinite bins in (F.71) thanks to the assumption that \( \max_{i \neq 1, k} \delta_i \to 0 \).

Finally, if \( \max_{i \neq 1, k} \delta_i \to 0 \), then for \( k \) and \( n \) large enough the largest \( p_i(\theta_0) \) will be the larger of \( p_1(\theta_0) \) or \( p_k(\theta_0) \), and if \( k/n \to 0 \) then

\[
\frac{k}{n} \max_i p_i(\theta_0) = \frac{k}{n} \max_{i=1,k} p_i(\theta_0) \to 0
\]

where the equality holds for \( k \) and \( n \) large enough. \( \square \)

**Theorem F.1** Suppose that\(^{45}\)

1. (Identifiability) \( \forall \delta > 0, \exists k_0 \) and \( \eta_1, \eta_2 > 0 \) small enough that

\(^{45}\)We must also require Assumptions F.1 and F.2.
\[
\sum_{i=1}^{k} \left( p_i(\theta_0) - \eta_i \right) \log \frac{p_i(\theta_0)}{p_i(\theta)} \geq \eta_2 , \quad \forall \ k \geq k_0 , \forall \ \theta \in \{|\theta - \theta_0| \geq \delta\}
\]

2. The infinite bins remain fixed (see footnotes 10 and 18),

3. Let \( \delta_i, i \neq 1, k \), denote the \( i \)th bin width (for the \( k - 2 \) finite bins). Assume

(a) \( p_i(\theta), i = 1, \ldots, k \) are continuous functions of \( \theta \),

(b) \( k/n \to 0 \) and \( \max_{i \neq 1, k} \delta_i \to 0 \),

(c) \( n \min_{i \neq 1, k} \delta_i \to \infty \), and

(d) \[
\frac{k}{\sqrt{n}} \max_{i \neq 1, k} \frac{1}{\sqrt{\delta_i}} \to 0
\]

4. Suppose that there is a neighborhood \( N(\theta_0) \) of \( \theta_0 \) such that \( f(x, \theta) \) and \( \frac{\partial^r}{\partial \theta^r} f(x, \theta) \), \( r = 1, 2, 3 \), are all continuous on \( \mathbb{R} \times N(\theta_0) \).\(^{46}\) Suppose further that there exist functions \( H_r(x) \), \( r = 1, 2, 3 \), such that

\[
\left| \frac{\partial^r}{\partial \theta^r} f(x, \theta) \right| \leq H_r(x) , \quad \forall \ \theta \in N(\theta_0) , \quad r = 1, 2, 3 \quad \text{and} \quad \int H_r(x)dx < \infty \ ,
\]

and that each \( H_r(x) \) is bounded,

5. (a) There exists a nbd \( N(\theta_0) \) and a constant \( K_{10} \) such that

\[
f(x, \theta) \geq K_{10} > 0 , \quad \forall \ \theta \in N(\theta_0) , \quad \forall x \in [a, b]
\]

\(^{46}\)See footnote 24 about the notation "\( \overline{N(\theta_0)} \)".
(b) There exists a nbd $N(\theta_0)$ and a constant $K_{20}$ such that

$$\inf_{\theta \in N(\theta_0)} \int_{-\infty}^{a} f(x, \theta) dx \geq K_{20} > 0 \quad \text{and} \quad \inf_{\theta \in N(\theta_0)} \int_{b}^{\infty} f(x, \theta) dx \geq K_{20} > 0$$

(c) Either

$$\left| \int_{-\infty}^{a} \frac{\partial}{\partial \theta} f(x, \theta_0) dx \right| > 0 \quad \text{and} \quad \int_{-\infty}^{a} f(x, \theta_0) dx > 0 , \quad \text{(F.72)}$$

or

$$\left| \int_{b}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta_0) dx \right| > 0 \quad \text{and} \quad \int_{b}^{\infty} f(x, \theta_0) dx > 0 . \quad \text{(F.73)}$$

then

$$\sqrt{\sum_{i=1}^{k} g_i^2 p_i \sqrt{n} (\hat{\theta}_n - \theta_0)} = \sqrt{\sum_{i=1}^{k} \left( \frac{\partial}{\partial \theta} \log p_i(\theta_0) \right)^2 p_i(\theta_0) \sqrt{n} (\hat{\theta}_n - \theta_0) \overset{\mathcal{D}}{\rightarrow} N(0,1) ,$$

where

$$\hat{\theta}_n = \arg \max_{\theta \in \Theta} \sum_{i=1}^{k} \frac{N_i}{n} \log p_i(\theta)$$

Proof. This theorem is identical to Lemma F.3 except that assumption 3. is changed.

We may use assumption 3. of Theorem F.1 to apply Result F.13. Result F.13 allows us to conclude that Lemma F.3, assumption 3. holds. Apply Lemma F.3. QED