NORTHWESTERN UNIVERSITY

SEMIPARAMETRIC EFFICIENCY BOUNDS AND TESTING IN MODELS WITH SHAPE RESTRICTIONS

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Economics

By

Gautam Tripathi

EVANSTON, ILLINOIS

June 1997

UMI Number: 9731355

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

To My Parents

ABSTRACT

Semiparametric estimation of finite dimensional parameters, when the unknown functions have to satisfy some smoothness properties, has been studied extensively. However, when dealing with economic data these unknown functions often have to satisfy other properties besides smoothness. These properties are actually the restrictions that economic theory imposes upon unknown functional forms. Typically, these restrictions turn out to be shape restrictions such as concavity, linear homogeneity and monotonicity. Though extensively studied in economic theory, there has only been a limited use of these restrictions in econometric practice notwithstanding their tremendous usefulness.

In this dissertation, I compute efficiency bounds for finite dimensional parameters, when the unknown function in the model is either concave, or homogeneous of degree r. I also construct estimators that actually achieve these bounds and show that homogeneity of the unknown function can lead to dramatic gains in efficiency for estimating finite dimensional parameters. As a subsidiary result I have developed a kernel estimator for homogeneous functions which, as far as I know, is new to the current econometric literature. I use this estimator to develop an asymptotically consistent test for homogeneity of functional form. Furthermore, I also show that if we restrict attention to the class of all regular estimators with square root asymptotics, then concavity of the unknown function does not help in estimating the finite dimensional parameters more efficiently.

In conclusion, this dissertation fulfills a twofold objective. Firstly, it enlarges the class of models that applied economists can deal with efficiently, and provides them with new techniques to efficiently estimate the finite dimensional parameters in semiparametric models with shape restrictions. Secondly, it eliminates a long standing lacuna in existing theoretical literature on efficient estimation, which has so far confined its attention to models where restrictions have been placed on the distribution of the "error" term while the unknown functional form, apart from some smoothness conditions, has been left virtually unrestricted. To the best of my knowledge, this attempt is the first of its kind to develop efficiency bounds for models where the shape restrictions are imposed on the unknown functional form rather than on the distribution of the error terms, which is assumed to be Gaussian.

ACKNOWLEDGEMENTS

I wish to thank my dissertation committee comprising Professors Ian Domowitz, Rosa Matzkin and Tom Severini for their constant encouragement and advice during the course of this research. Without their help this dissertation would not have been written. Needless to say, all errors still remain my responsibility. I am especially indebted to Professor Rosa Matzkin, the chair of my dissertation committee, and to Professor Tom Severini for introducing me to the wonderful world of shape restrictions and semiparametric estimation. Their easy accessibility and willingness to cheerfully answer my frequent, and often tiresome, questions made all the difference.

I am also grateful to the Alfred P. Sloan foundation for providing financial support in the form of a Dissertation Fellowship in Economics.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The objective of this dissertation is to determine how finite dimensional parameters can be efficiently estimated for an important class of economic models. In these models, some of the functions are known up to a finite dimensional parameter while the other functions are known only to possess some shape properties such as concavity, linear homogeneity and monotonicity. Such "semiparametric models with shape restrictions" are frequently encountered in microeconometrics. The use of the word semiparametric here highlights the fact that some components of these models are unknown functions, while the others are specified up to a finite dimensional parameter.

In this dissertation, we will calculate the minimum asymptotic variance, hereafter called the efficiency bound, that any estimator of the finite dimensional parameter can achieve in a semiparametric model when the unknown function is either homogeneous of degree r, or is concave. We will also construct an estimator that actually attains the efficiency bound, when the unknown function is homogeneous of degree r.

Previous techniques used in the determination of efficiency bounds applied only to models where either all functions were parametric or where the unknown functions were not restricted to possess any shape property. However, by confining our attenιion only to such cases, we exclude a large class of models that are of prime interest to applied economists. In the models that we will consider, the imposition of shape restrictions usually leads to a reduction in the variance of estimators without harming other limiting properties such as consistency. Omitting shape restrictions, when economic theory demands otherwise, would therefore lead to inefficient estimation procedures thus reducing the power of subsequent statistical analysis. This could have important policy implications, since semiparametric models are being increasingly used to answer policy related questions. However, inclusion of shape restrictions complicates estimation because such restrictions generate constraints that are infinite dimensional in nature. To deal with these infinite dimensional constraints, I use in this dissertation certain techniques borrowed from nonlinear analysis.

This dissertation limits itself to the analysis of i.i.d. observations, which are commonly generated by cross-sectional models for data. We look at the asymptotic variance because it is a widely used criterion in econometrics and statistics to rank estimators. Furthermore, in order to exclude pathological behavior such as superefficiency, all estimators are assumed to satisfy certain regularity conditions. Henceforth, unless specified otherwise the word estimator refers to a regular estimator.

The organization of this dissertation is as follows. In Chapter 1 we construct some typical examples of shape restricted models, and also provide a brief review of efficiency bounds for the parametric case. We then extend these concepts to the case when the nuisance parameter is infinite dimensional.

Chapter 2 introduces a partially linear shape restricted model. We begin by studying the identification issues and large sample properties for this model. We then compute efficiency bounds for the finite dimensional parameters when the unknown function is homogeneous of degree r, and also show how to construct an estimator that achieves these efficiency bounds. We also do the same for the case when the unknown function is concave. The chapter ends with a brief summary of the results obtained.

In Chapter 3, we use the estimator developed in Chapter 2 to estimate homogeneous functions. This estimator is then used to construct a test for detecting the homogeneity of functional forms. Results of a small simulation experiment, conducted to study the finite sample properties of this test, are also presented.

To enhance readability, the proofs of all major theorems and allied results have been confined to the appendices.

CHAPTER 2

SHAPE RESTRICTIONS AND EFFICIENCY BOUNDS

2.1. Models with Shape Restrictions

Let us begin by constructing some examples of shape restricted models that may typically occur in microeconometrics. In these examples, we are interested in estimating the finite dimensional parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\phi}_0$, when the only information we have about the unknown function is that it belongs to a set of functions whose elements satisfy certain properties. We let $\mathcal{F}$ denote this set of functions. All observed data is i.i.d., and ε is the unobserved error component which is assumed independent of the covariates. We also assume that ε is a random term drawn from a Normal distribution with zero mean and finite variance. In these models, any procedure used to obtain additional information about the finite dimensional parameters, apart from the fact that they are elements of a well defined parameter space, is called a semiparametric procedure.

EXAMPLE 2.1.1 (TWO INDEX MODEL). Consider n similar firms in equilibrium, which are geographically dispersed. The firms are competitive, have access to a CRS technology, and produce a single good. Therefore, for each firm the cost function $c(q, \mathbf{w}^*) = qc(\mathbf{w}^*, 1)$. Furthermore, for each firm, the factor prices $\mathbf{w}^*$ are unobserved, while the cost per unit

(c_i) is observable. Notice that $c_i = c(\mathbf{w}_i^*)$, and $c(\cdot)$ is continuous, nondecreasing, concave, and homogeneous of degree one in its arguments.

For simplicity, assume that there are only two factors of production. Now let $w_1^* = h_1(\mathbf{x}, \boldsymbol{\theta}_0)\, e^{\zeta_1}$ and, $w_2^* = h_2(\mathbf{z}, \boldsymbol{\phi}_0)\, e^{\zeta_2}$. Here, h_1, h_2 are known functions functions of exogenous variables, that determine the factor prices, and ζ_1, ζ_2 are error terms. Therefore, $\mathbb{E}(c|h_1, h_2) = \int c(h_1 e^{\zeta_1}, h_2 e^{\zeta_2})\, dF_{\zeta_1 \times \zeta_2} = f^*(h_1, h_2)$, and letting $\varepsilon = c - \mathbb{E}(c|h_1, h_2)$, we obtain the canonical regression model $c_i = f^*(h_1(\mathbf{x}_i, \boldsymbol{\theta}_0), h_2(\mathbf{z}_i, \boldsymbol{\phi}_0)) + \varepsilon_i$. Note that f^* also has the same properties as $c(\cdot)$. That is, f^* is continuous, nondecreasing, concave and homogeneous of degree one in its arguments. □

EXAMPLE 2.1.2 (PARTIALLY LINEAR MODEL). Consider a firm producing two different goods with production functions F_1 and F_2. That is, $y_1 = F_1(\mathbf{x})$, and $y_2 = F_2(\mathbf{z})$, with $(\mathbf{x} \times \mathbf{z}) \in \mathbb{R}^n \times \mathbb{R}^m$. The firm maximizes total profits $p_1 y_1 - \mathbf{w}_1'\mathbf{x} + p_2 y_2 - \mathbf{w}_2'\mathbf{z}$. The maximized profit can be written as $\pi_1(\mathbf{u}) + \pi_2(\mathbf{v})$, where $\mathbf{u} = (p_1, \mathbf{w}_1)$, and $\mathbf{v} = (p_2, \mathbf{w}_2)$.

Now suppose that the econometrician has sufficient information about the first good to parameterize the first profit function as $\pi_1(\mathbf{u}) = \mathbf{u}'\boldsymbol{\theta}_0$. Then the observed profit $\pi_i = \mathbf{u}_i'\boldsymbol{\theta}_0 + \pi_2(\mathbf{v}_i) + \varepsilon_i$, where π_2 is monotone, convex, linearly homogeneous and continuous in its arguments. □

EXAMPLE 2.1.3 (ANOTHER PARTIALLY LINEAR MODEL). Again, suppose we have n similar but geographically dispersed firms with the same profit function. This could happen if, for instance, these firms had access to similar technology. Now suppose that the observed profit depends not only upon the price vector, but also on a linear index of exogenous variables. That is, $\pi_i = \mathbf{x}_i'\boldsymbol{\theta}_0 + \pi^*(p_1^i, \ldots, p_k^i) + \varepsilon_i$, where the profit function π^* is continuous, monotone, convex, and homogeneous of degree one in its arguments. □

REMARK 2.1.1. As pointed out by Robinson (1988), partially linear models can serve as a first approximation to situations with "qualitative uneveness in prior information." □

2.2. Shape Restrictions and Semiparametric Estimation

Semiparametric estimation of the finite dimensional parameters, when $\mathcal{F}$ is just a set of functions satisfying some smoothness properties, has been studied extensively. However, when dealing with economic data the functions in $\mathcal{F}$ often have to satisfy other properties besides smoothness. These properties are actually the restrictions that economic theory imposes upon the unknown function. As the examples given above indicate, these are typically shape restrictions such as concavity, linear homogeneity and monotonicity. It is by now known that these properties provide powerful means for developing new estimation and testing techniques. Though extensively studied in economic theory, there has only been a limited use of these restrictions in econometric practice notwithstanding their tremendous usefulness. As Matzkin (1994) points out, these restrictions can be utilized "to reduce the variance of estimators, to falsify theories, and to extrapolate beyond the support of the data". Moreover, "economic restrictions can be used to guarantee the identification of some nonparametric models and the consistency of some nonparametric estimators".

As is well known, the move from a parametric approach to a semiparametric one is usually accompanied by a loss of efficiency. When restrictions implied by economic theory are imposed on the the semiparametric model, this efficiency loss may be mitigated due to a decrease in the variance of estimators. This problem of variance reduction is most critical since the quality of subsequent analysis depends upon the quality of current inference. By variance reduction I mean not only the computation of the smallest asymptotic variance of any estimator of the parameter of interest, but also the construction of estimators which actually possess this variance. As mentioned before, this minimum variance is called an efficiency bound, and an estimator which attains this bound is termed efficient.

Since shape restrictions can be utilized to reduce the variance of estimators, efficiency bounds are of fundamental importance in semiparametric models with shape restrictions. These bounds can be used to judge the efficiency of a proposed semiparametric estimator and to help develop new estimation techniques. They can also be used to provide a measure of efficiency loss in the move from a purely parametric approach to a semiparametric one. Moreover, in some cases these bounds also help in ruiing out the existence of certain types of estimators (Newey 1990).

The extension of efficiency bounds from a purely parametric to the semiparametric case was first proposed by Stein (1956) and subsequently developed in the statistical works cited in Bickel, Klassen, Ritov, and Wellner (1993). Attracted by the elegance of the semiparametric approach and its wide applicability to economics, several econometricians mentioned in Newey (1990)'s excellent survey article have also made valuable contributions to this area in recent years.

However, most of the research to date has concentrated upon developing efficiency bounds for distribution free models, i.e. models in which the distribution of the error term is unknown (Chamberlain 1986, Cosslett 1987). Where shape restrictions have been involved, they have been imposed on the error distribution (Newey 1988), rather than on the unknown function. Newey (1991) does discuss computing the efficiency bounds for a partially linear model, but here too the unknown functional form has only smoothness restrictions, and no shape restrictions, imposed upon it. But these cases form too narrow a class, since they exclude models with shape restrictions which arise regularly in microeconometrics, i.e. at the firm or the consumer level. In fact, as far as I know, this attempt is the first of its kind to develop efficiency bounds for models where the shape restrictions are imposed on the unknown functional form rather than on the distribution of the error term, which is assumed to be Gaussian. This research therefore, extends the class of models that econometricians can deal with efficiently. It should be of particular interest to any applied practitioner in the field because it provides new insights into incorporating shape restrictions in estimation procedures.

2.3. Asymptotic Efficiency and Lower Bounds

For each n let t_n be an estimate (based on n iid observations) of a real valued parameter β_0. Suppose that, for each β_0, $\sqrt{n}(t_n - \beta_0) \xrightarrow{d} N(0, v(\beta_0))$. Then according to Fisher, $v(\beta_0) \geq i_{\beta_0}^{-1}$, where i_{β_0} is the information contained in a single observation, and $v(\beta_0)$ is called the asymptotic variance of t_n. However, in the absence of suitable regularity conditions, this relationship does not necessarily hold as is indicated by the canonical example of a superefficient estimator given in LeCam (1953). Therefore, to make sure that the information inequality holds, we only consider regular estimators.[1] Regularity conditions which are typically imposed on an estimator sequence to rule out superefficiency, may be found in Bahadur (1964) or van der Vaart (1989).

In his seminal paper, Stein (1956), first proposed the idea of computing nonparametric efficiency bounds by using parametric submodels. The basic idea is as follows. Let $\beta_0 \in \mathbb{R}$ be the parameter of interest and η_0 be a finite dimensional nuisance parameter. The objective is to compute efficiency bounds for β_0, when η_0 is estimated nonparametrically. Now as Stein (1956) points out,

> ... a nonparametric problem is at least as difficult as any of the parametric problems obtained by assuming we have enough knowledge of the unknown state of nature to restrict it to a finite dimensional set.

[1] For the definition of a regular estimator, see Definition 3.4.1.

In fact we can restrict the unknown state of nature to a one dimensional set. It is instructive to see how this is done since the procedure generalizes in a straightforward manner to the case when β_0 is multidimensional and η_0 an infinite dimensional nuisance parameter. We consider the example in Severini (1987).

Let $X_1, \ldots, X_n$ be i.i.d. observation from a pdf $g(x, \beta_{10}, \ldots, \beta_{p0})$. Suppose that the parameter of interest is β_{10} while the nuisance parameter is $\eta_0 = (\beta_{20}, \ldots, \beta_{p0}) \in \mathbb{R}^{p-1}$. The vector of parameters to be estimated is therefore $\binom{\beta_{10}}{\eta_0}$, and the vector of scores is $\binom{S_{\beta_1}}{S_\eta}$. The information matrix for this p dimensional parameter can therefore be partitioned in the usual manner as

$$\begin{pmatrix} \mathbb{E}\, S^2_{\beta_1\beta_1} & \mathbb{E}\, S_{\beta_1} S'_\eta \\ \mathbb{E}\, S'_\eta S_{\beta_1} & \mathbb{E}\, S_\eta S'_\eta \end{pmatrix} = \begin{pmatrix} I_{\beta_1} & I_{\beta_1\eta} \\ I_{\eta\beta_1} & I_{\eta\eta} \end{pmatrix}.$$

Using the partitioned inverse formula, the Fisher information for a regular estimator of β_1 is found to be

$$I_{\beta_1} - I_{\beta_1\eta} I_{\eta\eta}^{-1} I_{\eta\beta_1}. \tag{2.3.1}$$

As Stein pointed out, the same result can be obtained if we look at an appropriate one dimensional parameterization of the nuisance parameter. This may be shown as follows.

Let $t \in [0, 1]$ and define $\boldsymbol{\beta}_t = \boldsymbol{\beta}_0 + t\boldsymbol{\delta}$, for any $\boldsymbol{\delta} \in \mathbb{R}^p$. Then the parameter of interest is $\beta_{1t} = \beta_{10} + t\delta_1$, and the nuisance parameter is, $\eta_t = (\beta_{20} + t\delta_2, \ldots, \beta_{p0} + t\delta_p)$. Notice that with this parameterization, estimating t is equivalent to estimating $\boldsymbol{\beta}_t$.

With this one dimensional parameterization, the loglikelihood function for estimating t from a single observation can be written as $\ell(\boldsymbol{\beta}_t; \mathbf{X}) = \ell(\beta_{10} + t\delta_1, \eta_t; \mathbf{X})$. Thus the score for estimating t is given by

$$\begin{aligned} S_t|_{t=0} &= \frac{d}{dt}\ell(\beta_{1t},\ldots,\beta_{pt};\mathbf{X})\Big|_{t=0} \\ &= \sum_{i=1}^{p}\frac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\beta_i}\delta_i \\ &= \delta_1\left[\frac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\beta_1} + S'_\eta\boldsymbol{\delta}_{-1}\right], \end{aligned}$$

where $\boldsymbol{\delta}_{-1} = (\delta_2/\delta_1,\ldots,\delta_p/\delta_1) \in \mathbb{R}^{p-1}$. The Fisher's information for estimating β_1 (denoted by $\tilde{i}_{\beta_{10}}$) is then given by

$$\begin{aligned} \tilde{i}_{\beta_{10}} = \mathbb{E}\,S^2_{\beta_{10}} &= \frac{\mathbb{E}\,(S_t|_{t=0})^2}{\delta_1^2} \\ &= \mathbb{E}\left(\frac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\beta_1} + S'_\eta\boldsymbol{\delta}_{-1}\right)^2. \end{aligned}$$

Now find the direction $\boldsymbol{\delta}_{-1} \in \mathbb{R}^{p-1}$ which gives the least possible information for estimating β_1, and denote this least information by $i_{\beta_{10}}$. [2] Then letting $S_{\beta_1} = \frac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\beta_1}$,

$$\begin{aligned} i_{\beta_{10}} &= \inf_{\boldsymbol{\delta}_{-1}\in\mathbb{R}^{p-1}} \tilde{i}_{\beta_{10}} \\ &= \inf_{\boldsymbol{\delta}_{-1}\in\mathbb{R}^{p-1}} \mathbb{E}\left(\frac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\beta_1} + S'_\eta\boldsymbol{\delta}_{-1}\right)^2, \quad \text{implying that,} \\ -\boldsymbol{\delta}^*_{-1} &= \text{proj}(S_{\beta_1}|\text{column space of } S_\eta). \end{aligned}$$

That is, the optimal value of $\boldsymbol{\delta}_{-1}$ is given by $\boldsymbol{\delta}^*_{-1} = \mathbb{E}\,(S_\eta S'_\eta)^{-1}\mathbb{E}\,S_\eta S_{\beta_1}$ which after a little algebra yields,

$$\begin{aligned} i_{\beta_{10}} &= \mathbb{E}\,S^2_{\beta_1} - \left(\mathbb{E}\,S_{\beta_1}S'_\eta\right)\left[\mathbb{E}\,S_\eta S'_\eta\right]^{-1}\left(\mathbb{E}\,S'_\eta S_{\beta_1}\right) \\ &= I_{\beta_1} - I_{\beta_1\eta}I^{-1}_{\eta\eta}I_{\eta\beta_1}, \end{aligned}$$

[2] In the literature, $i_{\beta_{10}}$ is referred to as the marginal Fisher information for β_1.

and this is the same as (2.3.1).

2.4. Extension to the Semiparametric Case

We will now extend the one dimensional parameterization, given in the previous section, to the case where the nuisance parameter lies in an an infinite dimensional cone. As will be seen later, the cone structure is useful as it incorporates the shape restrictions imposed on the nuisance parameter. In order to deal with the conical structure of the nuisance parameter space, we need to define the following terms.

Definition 2.4.1 (Cone). Let X be a vector space over $\mathbb{R}$. A subset C of X is called a cone iff for any $c \in C$, and any $\lambda \geq 0$ we have $\lambda c \in C$. □

Definition 2.4.2 (Fréchet Derivative). Let T be a transformation defined on an open domain U in a normed space X and having range in a normed space Y. If for fixed $x \in U$ and each $h \in X$ there exists a linear and continuous operator $L \in \mathcal{L}(X, Y)$ such that

$$\lim_{\|h\| \to 0} \frac{\|T(x+h) - T(x) - Lh\|}{\|h\|} = 0,$$

then T is said to be Fréchet differentiable at x. The operator L, often denoted by $T'(x)$, is called the Fréchet derivative of T at x. □

Remark 2.4.1. Note that since the limit is taken as $\|h\| \to 0$, we only have to consider arbitrarily small perturbations $h \in X$. □

Definition 2.4.3 (Tangent Vector). Let M be a subset of a Banach space X. A vector $x \in X$ is said to be tangent to the set M at a point x_0 if there exist an $\epsilon_0 > 0$ and a mapping $t \mapsto r(t)$ of the interval $(0, \epsilon_0)$ into X such that

$$x_0 + tx + r(t) \in M, \quad \text{for all } t \in (0, \epsilon_0), \text{ and,}$$
$$\frac{\|r(t)\|}{t} \to 0 \quad \text{as } t \to 0. \quad \square$$

DEFINITION 2.4.4 (ADMISSIBLE CURVE). In the above definition, let $\gamma(t) = x_0 + tx + r(t)$. Then γ is said to be an admissible curve in M through x_0. It has the property that $\gamma(0) = x_0$, and $\gamma'(0) = x$ is the tangent vector to M at x_0. □

DEFINITION 2.4.5 (TANGENT CONE AND TANGENT SPACE). The set of vectors which are tangent to a set M at the point x_0, is denoted by $T(M, x_0)$, and, is a closed non-empty cone. This cone is called the tangent cone to M at x_0. If this cone is a subspace, then it is called the tangent space to M at x_0. □

REMARK 2.4.2. An equivalent characterization of tangent vectors and tangent cones is given in Appendix A. This appendix also contains several useful results about tangent cones that are used subsequently. □

We now return to our original problem. So let β_0 be a real valued parameter of interest in an open set $\mathrm{B} \subset \mathbb{R}$, and f^* be the true value of the nuisance parameter. Furthermore, let $f^* \in \mathcal{F}$, where $\mathcal{F}$ is a convex cone in a Banach space $\mathcal{H}$ and assume that the parameter space $\mathrm{B} \times \mathcal{F}$ is parameterized by the curve $\beta \mapsto (\beta, \eta_\beta)$ such that $\eta_\beta|_{\beta=\beta_0} = f^*$. Now let $t \mapsto \beta_t$ be an admissible curve in B through β_0. Any point in the parameter space $\mathrm{B} \times \mathcal{F}$ then has coordinates $(\beta_t, \eta_{\beta_t})$. Again note that with this parameterization, estimating t is equivalent to estimating $(\beta_t, \eta_{\beta_t})$. Now consider the following assumption.

ASSUMPTION 2.4.1. *Let the score functions be elements of the Hilbert space $\mathcal{L}^2(\mathbb{D})$, where $\mathbb{D}$ is the probability measure induced by the data.* □

This is very useful since the geometry of Hilbert spaces facilitates the solution of projection problems which we shall soon encounter. The Fisher information for estimating t is then given by,

$$\mathbb{E}\left[\frac{d}{dt}\ell(\beta_t,\eta_{\beta_t})\Big|_{t=0}\right]^2 = \left(\frac{d\beta_t}{dt}\Big|_{t=0}\right)^2 \mathbb{E}\left[\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\frac{d}{d\beta}\eta_{\beta_0})\right]^2.$$

Let $\bar{i}_\beta$ be the Fisher information for estimating β. Then since β is a function of t, an application of the chain rule gives,

$$\bar{i}_\beta = \frac{\mathbb{E}\left[\frac{d}{dt}\ell(\beta_t,\eta_{\beta_t})\Big|_{t=0}\right]^2}{\left(\frac{d\beta_t}{dt}\Big|_{t=0}\right)^2} = \mathbb{E}\left[\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\beta}\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\frac{d}{d\beta}\eta_{\beta_0})\right]^2.$$

Since $\bar{i}_\beta$ depends upon η_β only through the tangent vector $\frac{d}{d\beta}\eta_{\beta_0}$, the tangent vector δ^* which gives the least information for estimating β is

$$\delta^* = \operatorname*{argmin}_{\frac{d}{d\beta}\eta_{\beta_0}\in T(\mathcal{F},f^*)} \mathbb{E}\left[\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\frac{d}{d\beta}\eta_{\beta_0})\right]^2.$$

That is, $\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\delta^*)$ is the projection of $\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\beta}$ onto $\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(T(\mathcal{F},f^*))$. Therefore, since we are optimizing in a Hilbert space, δ^* is characterized by the necessary and sufficient conditions given in Theorem H.3. That is,

(i) $\mathbb{E}[\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\delta^*)]\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\delta^*) = 0$, and,

(ii) $\mathbb{E}[\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\delta^*)]\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\delta) \geq 0$, for all $\delta \in T(\mathcal{F},f^*)$.

Now let,

$$i_{\beta_0} = \mathbb{E}\left[\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\delta^*)\right]^2.$$

Then following Stein (1956), $i_{\beta_0}^{-1}$ is a lower bound for the asymptotic variance of any regular estimator of β. This is verified in Severini (1987) for the case when the tangent cone is actually a linear space. But what happens when $T(\mathcal{F},f^*)$ is a proper cone, i.e. when $T(\mathcal{F},f^*)$ is not a linear space? To answer this question, let $\overline{lin\,T(\mathcal{F},f^*)}$ denote the smallest closed linear space containing $T(\mathcal{F},f^*)$. Also, let

$$i_c = \underset{\frac{d}{d\beta}\eta_{\beta_0} \in T(\mathcal{F}, f^*)}{\text{argmin}} \ \mathbb{E}\left[\frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\frac{d}{d\beta}\eta_{\beta_0})\right]^2$$

$$i_l = \underset{\frac{d}{d\beta}\eta_{\beta_0} \in \overline{lin\, T(\mathcal{F}, f^*)}}{\text{argmin}} \ \mathbb{E}\left[\frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\frac{d}{d\beta}\eta_{\beta_0})\right]^2 .$$

Then since $T(\mathcal{F}, f^*) \subset \overline{lin\, T(\mathcal{F}, f^*)}$, $i_c^{-1} \leq i_l^{-1}$. That is, a projection on the tangent cone $T(\mathcal{F}, f^*)$ seems to yield a better lower bound as compared to the one obtained by projecting the parametric scores on $\overline{lin\, T(\mathcal{F}, f^*)}$.

However, with the help of two parametric examples in the next section we will show that a projection on the tangent cone $T(\mathcal{F}, f^*)$ leads to a lower bound which is either

(i) too optimistic for the m.l.e. of β_0, or

(ii) which is actually beaten by the m.l.e..

But if the projection is taken on $\overline{lin\, T(\mathcal{F}, f^*)}$, not only is the efficiency bound so obtained a valid lower bound, but we will also be able to construct regular estimators that actually achieve this bound. Hence the space on which the parametric scores should be projected to obtain the efficiency bounds is $\overline{lin\, T(\mathcal{F}, f^*)}$, and not $T(\mathcal{F}, f^*)$. These results will be extended to the semiparametric case in Section 3.3.

Notice that if we project the parametric scores on $\overline{lin\, T(\mathcal{F}, f^*)}$, the projection δ^* is given by the necessary and sufficient conditions of Theorem H.2. That is, for all $\delta \in \overline{lin\, T(\mathcal{F}, f^*)}$,

$$\mathbb{E}\left[\frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\delta^*)\right]\frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\delta) = 0.$$

2.5. Shape Restrictions in Simple Linear Regression

In the following examples, we impose monotonicity and convexity in the framework of simple linear regression to see how the imposition of such a shape restriction affects the efficiency bounds for the parameter of interest.

EXAMPLE 2.5.1 (MONOTONICITY). Consider the following linear regression,

$$y_i = \theta_0 + \lambda_0 z_i + \varepsilon_i, \qquad i = 1, \dots, n. \tag{2.5.1}$$

Here $\varepsilon \overset{d}{=} \text{NIID}(0,1)$, and z is a random variable with positive variance that is independent of ε. The parameter of interest is $\theta_0 \in (-\infty, \infty)$, and the nuisance parameter is $\lambda_0 \in \Lambda = [0, \infty)$. That is, we want to fit an increasing line to the i.i.d. observations (y, z). Then following the procedure in Section 2.4, we obtain Table (2.5.1).

TABLE (2.5.1). Lower Bounds for Estimating θ_0

Nuisance Parameter	$T(\Lambda, \lambda_0)$	Lower Bound
$\lambda_0 = 0$	$[0, \infty)$	$1/\mathbb{E}\,[1 - Z\min\{0, (\mathbb{E}\,Z)/\mathbb{E}\,Z^2\}]^2$
$\lambda_0 > 0$	$(-\infty, \infty)$	$\mathbb{E}\,Z^2/(\operatorname{Var} Z)$

Notice that the efficiency bound depends upon the true value of the nuisance parameter λ_0. Let us now see if the m.l.e. of θ_0 achieves these bounds.

So define $S_{\mathbf{zz}} = \sum_{i=1}^n (z_i - \bar{z}_n)^2$, $\bar{z}_n = \frac{1}{n}\sum_{i=1}^n z_i$, $\mathbf{z}'\mathbf{z} = \sum_{i=1}^n z_i^2$, and let $\bar{\theta}_n$ denote the m.l.e. of θ_0. Then,

$$\bar{\theta}_n = \begin{cases} \bar{y}_n - \hat{\lambda}_n \bar{z}_n & \text{if } \hat{\lambda}_n \geq 0 \\ \bar{y}_n & \text{if } \hat{\lambda}_n < 0, \end{cases}$$

with $\hat{\lambda}_n = \sum_{i=1}^n (z_i - \bar{z}_n) y_i / S_{\mathbf{zz}}$. Now letting $\phi(\cdot)$ denote the p.d.f., and $\Phi(\cdot)$ the c.d.f. of a standard normal random variable, it can be shown that conditional on observing $z_1, \dots, z_n$,

$$\Pr\{n^{1/2}(\bar{\theta}_n - \theta_0) \leq t\} = \int_{u=-\infty}^{t\sqrt{S_{\mathbf{zz}}/\mathbf{z}'\mathbf{z}}} \frac{e^{-u^2/2}}{\sqrt{2\pi}} \Phi(n^{1/2}\lambda_0\sqrt{\mathbf{z}'\mathbf{z}/n} - \bar{z}_n\sqrt{n/S_{\mathbf{zz}}}u)\,du$$
$$+ \Phi(-n^{1/2}\lambda_0\sqrt{S_{\mathbf{zz}}/n})\,\Phi(t - n^{1/2}\lambda_0\bar{z}_n).$$

Then letting $F(w) = \int_{u=-\infty}^{w\sqrt{\frac{\operatorname{Var} Z}{\mathbb{E} Z^2}}} \phi(u)\Phi(-u\frac{\mathbb{E} Z}{\sqrt{\operatorname{Var} Z}})\, du$, we can show that

$$\Pr\{n^{1/2}(\tilde{\theta}_n - \theta_0) \leq t\} \to \begin{cases} \frac{1}{2}\Phi(t) + F(t) & \text{if } \lambda_0 = 0 \\ \Phi(t\sqrt{\frac{\operatorname{Var} Z}{\mathbb{E} Z^2}}) & \text{if } \lambda_0 > 0, \end{cases}$$

and that the asymptotic variance

$$\operatorname{AsVar}(n^{1/2}(\tilde{\theta}_n - \theta_0)) = \begin{cases} \frac{1}{2}(1 + \frac{\mathbb{E} Z^2}{\operatorname{Var} Z}) - \frac{1}{2\pi}\frac{(\mathbb{E} Z)^2}{\operatorname{Var} Z} & \text{if } \lambda_0 = 0 \\ \frac{\mathbb{E} Z^2}{\operatorname{Var} Z} & \text{if } \lambda_0 > 0. \end{cases} \tag{2.5.2}$$

Remark 2.5.1. (i) When $\mathbb{E} Z \neq 0$,

$$1 < \frac{1}{2}(1 + \frac{\mathbb{E} Z^2}{\operatorname{Var} Z}) - \frac{1}{2\pi}\frac{(\mathbb{E} Z)^2}{\operatorname{Var} Z} < \frac{\mathbb{E} Z^2}{\operatorname{Var} Z},$$

and the asymptotic variance is not continuous at $\lambda_0 = 0$.

(ii) Also notice that when $\lambda_0 = 0$, the asymptotic distribution of the m.l.e. is not normal. □

The results obtained above are presented in tabular form below.

Table (2.5.2). Imposing Monotonicity in Linear Regression

$\mathbb{E} Z$	λ_0	Lower Bound	$\tilde{\theta}_n$
$\mathbb{E} Z = 0$	$\lambda_0 = 0$	Attained by the m.l.e.	Not Regular
$\mathbb{E} Z > 0$	$\lambda_0 = 0$	Not attained by the m.l.e.	Not Regular
$\mathbb{E} Z < 0$	$\lambda_0 = 0$	Beaten by the m.l.e.	Not Regular
$\mathbb{E} Z = 0$	$\lambda_0 > 0$	Attained by the m.l.e.	Regular
$\mathbb{E} Z > 0$	$\lambda_0 > 0$	Attained by the m.l.e.	Regular
$\mathbb{E} Z < 0$	$\lambda_0 > 0$	Attained by the m.l.e.	Regular

Remark 2.5.2. A brief description of the results in Table (2.5.2) follows.

(i) When $\lambda_0 = 0$, the efficiency bounds are attained only when $\mathbb{E} Z = 0$. When $\mathbb{E} Z \neq 0$, the bound is either not attained (when $\mathbb{E} Z > 0$), or is actually beaten by $\tilde{\theta}_n$ (when $\mathbb{E} Z < 0$). However, it may be shown that in all these cases the estimator $\tilde{\theta}_n$ is not regular.

(ii) When $\lambda_0 > 0$, not only is $\tilde{\theta}_n$ regular, but it also attains the lower bounds. Now the tangent cone $T(\Lambda, \lambda_0)$, when $\lambda_0 > 0$ is $(-\infty, \infty)$. But $(-\infty, \infty)$ is also the smallest closed linear space containing $[0, \infty)$. Hence, if we restrict ourselves to the class of regular estimators, the space on which the projection is taken to obtain the efficiency bound should be $\overline{lin\, T(\Lambda, \lambda_0)}$, rather than the tangent cone itself. Projection on this larger space will lead to bounds that can be attained by regular estimators. To do any better, we would have to use an estimator that is not regular. □

EXAMPLE 2.5.2 (CONVEXITY). We now impose convexity in linear regression. This is easily done by substituting $Z = X^2$ in the previous example. Thus the shape restricted regression now becomes,

$$y_i = \theta_0 + \lambda_0 x_i^2 + \varepsilon_i, \qquad i = 1, \dots, n,$$

under the same conditions as before. Notice that imposing the restriction $\lambda_0 \geq 0$, now implies that we are fitting a convex function to the data. We now have the following results, which are stronger than those obtained in the previous example.

TABLE (2.5.3). Imposing Convexity in Linear Regression

λ_0	Lower Bound	$\tilde{\theta}_n$
$\lambda_0 = 0$	Not attained by the m.l.e.	Not Regular
$\lambda_0 > 0$	Attained by the m.l.e.	Regular

As before, when $\lambda_0 > 0$, the efficiency bound is attained by $\tilde{\theta}_n$. However, when $\lambda_0 = 0$, the efficiency bound is not attained. These results once again show that to obtain efficiency bounds which are attainable by regular estimators, the projection must be taken on $\overline{lin\, T(\Lambda, \lambda_0)}$, rather than the tangent cone $T(\mathcal{F}, f^*)$ itself. □

2.6. Scalar Parameter of Interest

With the previous section in mind, we are now in a position to deal with scalar parameters of interest. So let β_0 be an element of an open set $\mathrm{B} \subset \mathbb{R}$. $f^* \in \mathcal{F}$ is the true nuisance parameter, where $\mathcal{F}$ is a convex cone in a Banach space $\mathcal{H}$. Assume that the parameter space $\mathrm{B} \times \mathcal{F}$ is parameterized by a smooth curve $\beta \mapsto (\beta, \eta_\beta)$ such that $\eta_\beta|_{\beta=\beta_0} = f^*$, and let

$$\delta^* = \underset{\frac{d}{d\beta}\eta_{\beta_0} \in \overline{lin\,T(\mathcal{F},f^*)}}{\operatorname{argmin}} \mathbb{E}\left[\frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\frac{d}{d\beta}\eta_{\beta_0})\right]^2 .$$

Remark 2.6.1. The direction δ^* is the projection of the score of the parameters of interest onto $\overline{lin\,T(\mathcal{F},f^*)}$, and is called the least favorable direction for estimating β_0. A curve η_β which gives rise to this tangent vector is called a least favorable curve. However, η_β may not necessarily lie in the cone $\mathcal{F}$. For instance, let $\mathcal{F}$ be the set of all $\mathrm{C}^2(\mathbf{Z})$ - concave functions, and let f^* be affine. Then from Section 3.7 we have $\overline{lin\,T(\mathcal{F},f^*)} = \mathrm{C}^2(\mathbf{Z})$, and therefore suppose that $\delta^* \in \overline{lin\,T(\mathcal{F},f^*)}$ is a strictly convex function. Then the curve $\lambda_t = f^* + t\delta^*$ has δ^* as the tangent vector and $\lambda_0 = f^*$, but λ_t being strictly convex does not lie in $\mathcal{F}$. But since $\overline{lin\,T(\mathcal{F},f^*)} \subset \overline{lin\mathcal{F}}$, we can always find a curve in $\overline{lin\mathcal{F}}$ which gives rise tc the least favorable direction δ^*. We call this curve, a least favorable curve. Notice that while Theorem H.2 implies that the least favorable direction is unique, no such implication holds for the least favorable curve. For instance, the curves $t \mapsto f^* + t\delta^*$ and $t \mapsto f^* + t(t+1)\delta^*$ give rise to the same least favorable direction at $t = 0$.

Hence, we have the following definition.

Definition 2.6.1 (Least Favorable Curve and Direction). Let,

$$\mathrm{B} \times \overline{lin\mathcal{F}}$$

be parameterized by a smooth curve $\beta \mapsto (\beta, \eta_\beta)$ such that, $\eta_\beta|_{\beta=\beta_0} = f^*$. Then $\eta_\beta \in \overline{lin\mathcal{F}}$ is said to be a least favorable curve for estimating β_0, if $\frac{d}{d\beta}\eta_{\beta_0} \in \overline{lin\,T(\mathcal{F},f^*)}$ minimizes

$$\mathbb{E}[\frac{d\ell(\beta_0, \eta_{\beta_0})}{d\beta}]^2 = \mathbb{E}[\frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\frac{d}{d\beta}\eta_{\beta_0})]^2.$$

Moreover, the direction $\frac{d}{d\beta}\eta_{\beta_0} \in \overline{lin T(\mathcal{F}, f^*)}$ is called the least favorable direction for estimating β_0. □

REMARK 2.6.2. (i) Let $\frac{d}{d\beta}\eta_{\beta_0}$ be the least favorable direction for estimating β_0, and define

$$i_{\beta_0} = \mathbb{E}[\frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\frac{d}{d\beta}\eta_{\beta_0})]^2.$$

Then i_{β_0} is called the semiparametric information for β_0, and it may be shown that $i_{\beta_0}^{-1}$ is a lower bound for the asymptotic variance of any regular estimator of β_0. See for instance, van der Vaart (1989).

(ii) If $\mathcal{F}$ is a linear space, then $T(\mathcal{F}, f^*) = \overline{lin T(\mathcal{F}, f^*)} = \mathcal{F}$.

(iii) As mentioned above, $\overline{lin T(\mathcal{F}, f^*)} \subset \overline{lin \mathcal{F}}$. A sufficient condition for the converse to hold is that $f^* \in T(\mathcal{F}, f^*)$. This is so, because f^* need not always be an element of $T(\mathcal{F}, f^*)$. To see this, let $A = \{1\}$. Then $T(A, 1) = \{0\}$, but $1 \notin T(A, 1)$. However, since this condition always holds for the cases which interest us, we assume the following. □

ASSUMPTION 2.6.1. *Let $\mathcal{F}$ be a convex cone, and let $f^* \in \mathcal{F}$. Then, $f^* \in T(\mathcal{F}, f^*)$.* □

REMARK 2.6.3. (i) Actually, this assumption holds whenever $\mathcal{F}$ is a cone. To see this let $\mathcal{F}$ be a cone, and let $f^* \in \mathcal{F}$. Now consider the curve $\gamma(t) = f^* + tf^*$, for $t > 0$. Since $\mathcal{F}$ is a cone, $\gamma(t) \in \mathcal{F}$ for all $t > 0$. Also, $\gamma'(0) = f^*$. Hence, $\gamma(t)$ is a curve in $\mathcal{F}$ with f^* as the tangent vector. That is, $f^* \in T(\mathcal{F}, f^*)$.

(ii) Furthermore, convexity of $\mathcal{F}$ implies that $\mathcal{F} \subset T(\mathcal{F}, f^*)$. This may be seen as follows. Since the tangent cone $T(\mathcal{F}, f^*)$ is the smallest closed cone containing $\mathcal{F} - f^*$, we have that $\mathcal{F} - f^* \subset T(\mathcal{F}, f^*)$ or $\mathcal{F} \subset T(\mathcal{F}, f^*) + f^*$. But if $f^* \in T(\mathcal{F}, f^*)$, then $\mathcal{F} \subset T(\mathcal{F}, f^*)$, since $T(\mathcal{F}, f^*)$ is also convex due to the convexity of $\mathcal{F}$. □

Therefore, under this assumption we can show that whenever $\mathcal{F}$ is a convex cone, LEMMA 2.6.1. $\overline{lin T(\mathcal{F}, f^{\bullet})} = \overline{lin \mathcal{F}}$.

PROOF. See Appendix H. □

2.7. Multidimensional Parameter of Interest

Now suppose that $\beta_0 \in \mathbf{B}^{\circ} \subset \mathbb{R}^p$. The nuisance parameter $f^{\bullet}$ is once again an element of $\mathcal{F}$, a convex cone in a Banach space $\mathcal{H}$. Assume that $\mathbf{B} \times \overline{lin\mathcal{F}}$ is parameterized by a smooth curve $\boldsymbol{\beta} \mapsto (\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}})$ such that $\eta|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = f^{\bullet}$. We then have the following definition.

DEFINITION 2.7.1 (LEAST FAVORABLE SURFACE). Let $t \mapsto \boldsymbol{\beta}_t$ be any admissible curve in $\mathbf{B}$ through $\boldsymbol{\beta}_0$. Then $\eta_{\boldsymbol{\beta}}$ is called a least favorable surface for estimating $\boldsymbol{\beta}_0$, if $\eta_{\boldsymbol{\beta}_t}$ is a least favorable curve for estimating t. That is, $\eta_{\boldsymbol{\beta}_t}$ minimizes $\mathbb{E}\,(\frac{d\ell(\boldsymbol{\beta}_t, \eta_{\boldsymbol{\beta}_t})}{dt}|_{t=0})^2$. □

Using this definition we can obtain the following theorems of Severini (1987). These theorems therefore extend Severini's results for the case when the nuisance parameter is restricted to lie in a cone. These results are also of interest because they show that dealing with a p dimensional parameter of interest is equivalent to to solving p individual optimization problems. The proofs of the following theorems are provided in Appendix B for the sake of completeness.

THEOREM 2.7.1. *Let $\eta_{\boldsymbol{\beta}}$ be a least favorable surface for estimating $\boldsymbol{\beta}$. Denote the least favorable direction* [3] *by $\boldsymbol{\delta}^{\bullet} = (\frac{d}{d\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$. That is, $\delta_i^{\bullet} = (\frac{d}{d\beta_i}\eta_{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ for $i = i, \ldots, p$. Then, $\boldsymbol{\delta}^{\bullet} = (\delta_1^{\bullet}, \ldots, \delta_p^{\bullet})$ satisfies*

$$\mathbb{E}\left[\frac{\partial\ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial\beta_i} + \frac{\partial\ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial\eta}(\delta_i^{\bullet})\right]\frac{\partial\ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial\eta}(\delta) = 0,$$

for all $\delta \in \overline{lin T(\mathcal{F}, f^{\bullet})}$, and $i = 1, \ldots, p$.

[3] Note that the least favorable direction is now an element of $\times_{i=1}^{p}\overline{lin T(\mathcal{F}, f^{\bullet})}$.

THEOREM 2.7.2. *η_β is a least favorable surface for estimating β, if and only if for all $\delta_i \in \overline{lin T(\mathcal{F}, f^*)}$, and $i = 1, \dots, p$,*

$$\sum_{i=1}^{p} \mathbb{E}\left[\frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \beta_i} + \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\frac{d}{d\beta_i}\eta_{\beta_0})\right] \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\delta) = 0.$$

THEOREM 2.7.3. *Let,*

$$I = \mathbb{E}\left[\frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\frac{d}{d\beta}\eta_{\beta_0})\right]\left[\frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\frac{d}{d\beta}\eta_{\beta_0})\right]'$$

be the information matrix when η_β is a curve in $\overline{lin\mathcal{F}}$, through f^*. Then there exists a matrix I_{β_0} such that $\alpha'(I_{\beta_0} - I)\alpha \leq 0$ for all $\alpha \in \mathbb{R}^p$, iff I_{β_0} corresponds to the information matrix when η_β is a least favorable surface.

REMARK 2.7.1. (i) This theorem shows that a least favorable surface η_β, minimizes Fisher's information I_{β_0} in the usual sense. That is, for any other information matrix I, the matrix difference $I_{\beta_0} - I$ is always negative semi-definite. Following van der Vaart (1989), $I_{\beta_0}^{-1}$ remains a valid lower bound for the asymptotic variance of regular estimators of β_0.

(ii) Furthermore, Theorem 2.7.1 and Theorem 2.7.3 together imply that to find the least favorable direction in the multiparameter case, we simply find the least favorable direction corresponding to each component of the parameter vector.

CHAPTER 3

A GENERAL SHAPE RESTRICTED MODEL

3.1. Introduction

Consider the regression $y = \mathbf{x}\beta_0 + f^*(\mathbf{z}) + \varepsilon$, with $\varepsilon \overset{d}{=} \mathrm{N}(0,1)$. In this chapter we show how to efficiently estimate β_0, when the only information we have about f^* is that it has a certain shape. That is, we efficiently estimate β_0 when the only thing we know about f^* is that it lies in $\mathcal{F}$, where $\mathcal{F}$ is a compact set of functions with certain shape properties. These shape properties are such that $\mathcal{F}$ is a convex cone. We examine two cases in detail. In the first case $\mathcal{F}$ is the set of C^2 - homogeneous functions of degree r, while in the second case the elements of $\mathcal{F}$ are C^2 - concave functions. In each case we compute the efficiency bounds for estimating β_0, and also propose an estimator that attains these bounds. The efficiency bounds are shown to be determined by a projection onto $\overline{lin\,T(\mathcal{F}, f^*)}$, the smallest closed linear space containing the tangent cone to $\mathcal{F}$ at f^*. This tangent cone, denoted by $T(\mathcal{F}, f^*)$, seems at first sight to be the natural space to determine the efficiency bounds. However, we prove an "impossibility" result showing that projecting onto $T(\mathcal{F}, f^*)$ yields bounds that are not attainable by any $n^{1/2}$ consistent, regular estimator of β_0. This "impossibility" result is used to show that in the class of all $n^{1/2}$ consistent regular estimators of β_0, homogeneity of f^* can lead to dramatic efficiency gains in estimating

$\boldsymbol{\beta}_0$, while concavity of f^* does not help in estimating $\boldsymbol{\beta}_0$ more efficiently.

We now begin our study of a general shape restricted semiparametric model by analyzing the partially linear model. For $i = 1, \ldots, n$, consider the regression

$$y_i = x_{1i}\beta_{10} + x_{2i}\beta_{20} + \ldots + x_{pi}\beta_{p0} + f^*(z_{1i}, z_{2i}) + \varepsilon.$$

Assumption 3.1.1. *In the above model let,*

(i) $\varepsilon \overset{\mathrm{d}}{=} N(0, \sigma_0^2)$, *where* σ_0^2 *is known;*

(ii) $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{p0}) \in \mathbf{B}^\circ$, *where* $\mathbf{B}$ *is a compact subset of* $\mathbb{R}^p$*;*

(iii) $\mathbf{x}$ *comes from a distribution with compact support* $\mathbf{X}$ *in* $\mathbb{R}^p$*. Similarly,* $\mathbf{z} = (z_1, z_2)$ *comes from a distribution with compact support* $\mathbf{Z} = Z_1 \times Z_2$ *in* $\mathbb{R}^2$*. Furthermore,* $\mathbf{x}, \mathbf{z}$ *have a joint density* $g_0(\cdot, \cdot)$*, which induces a probability measure* $\mathbb{G}$ *on support* $S_{\mathbb{G}}$*;*

(iv) *let* $\mathcal{H}$ *denote the set of all* C^2 *functions on* $\mathbf{Z}$ *with uniformly bounded values, gradients, and Hessians. Then* $f^* \in \mathcal{F} \subseteq \mathcal{H}$*, where* $\mathcal{F}$ *is a closed, convex cone in* $\mathcal{H}$*, and consists of functions that satisfy certain shape properties;*

(v) ε *and* $(\mathbf{x}, \mathbf{z})$ *are independent, and we observe* $(\mathbf{x}, y, \mathbf{z})$. □

Remark 3.1.1. (i) The assumption that σ_0^2 is known, is w.l.o.g. since it can be shown that the efficiency bound is not affected by the knowledge of σ_0^2. Therefore, we choose $\sigma_0^2 = 1$.

(ii) Since $\mathcal{H}$ is a compact subset of $\mathrm{C}^2(\mathbf{Z})$ w.r.t. the C^2 norm and $\mathcal{F}$ is a closed subset of $\mathcal{H}$, $\mathcal{F}$ is also compact w.r.t. the C^2 norm. □

Notation 3.1.1. Unless otherwise specified, $||\cdot||$ represents the sup norm in function spaces, and the Euclidean norm in finite dimensional spaces. □

3.2. Identification

The question of identification, as has been pointed out many authors, is logically prior to that of estimation. There is no sense in estimating parameters which are not identified. In this section we provide sufficient conditions under which the parameters (f^*, β) are identified. As we shall soon see, parameters in the partially linear model are identified under fairly weak conditions.

Since we observe only the 3-tuple $(\mathbf{x}, y, \mathbf{z})$, the most that can be obtained from the data is the joint density of $(\mathbf{x}, y, \mathbf{z})$. The question of identification then reduces to that of recovering the true parameter values (f^*, β_0) from this joint density.

Assumption 3.2.1 (Identification). *Let,*

(i) *the vector* β_0 *be without an intercept term, and*

(ii) *let the elements of the vector* $\varphi(\mathbf{x}, \mathbf{z}) = \mathbf{x} - \mathbb{E}(\mathbf{x}|\mathbf{z})$ *be linearly* $\mathbb{G}$ *- independent. That is, if* $\mathbf{a}'\varphi(\mathbf{x}, \mathbf{z}) = 0$ *for* $\mathbb{G}$ *- a.a.* $(\mathbf{x}, \mathbf{z})$, *then* $\mathbf{a} = \mathbf{0}$. □

Remark 3.2.1. We exclude intercept terms because they cannot be identified in the partially linear model. □

We then have the following result, which was first obtained by Robinson (1988). For the sake of completeness, we provide a proof of this result.

Theorem 3.2.1 (Robinson). *Let the partially linear model satisfy Assumption* 3.2.1. *Then,* (β_0, f^*) *is identified in* $(\mathbf{B}, \mathcal{F})$.

Proof. First notice that if the true parameter values are replaced by $(\beta, f) \in \mathbf{B} \times \mathcal{F}$, then

$$y - \mathbb{E}(y|\mathbf{z}; \beta, f) = \varphi(\mathbf{x}, \mathbf{z}) \cdot \beta + \varepsilon.$$

So to show that (β_0, f^*) is identified, let (β_1, f_1) and (β_2, f_2) be two values of the true parameter (β_0, f^*) and let

$$\mathbb{E}(y|\mathbf{z};\beta_1, f_1) = \mathbb{E}(y|\mathbf{z};\beta_2, f_2).$$

But this implies that $\varphi(\mathbf{x},\mathbf{z})\cdot(\beta_1 - \beta_2) = 0$. And since the elements of $\varphi(\mathbf{x},\mathbf{z})$ are linearly independent by assumption, we have $\beta_1 = \beta_2$ which also implies that $f_1 = f_2$. Therefore, (β_0, f^*) is identified. □

3.3. Efficiency Bounds for the Partially Linear Model

Once we know that our model is identified, we can proceed with its asymptotic analysis. Now since the data generating process is

$$y_i = \mathbf{x}_i\beta_0 + f^*(\mathbf{z}_i) + \varepsilon_i \quad i = 1, \ldots, n, \tag{3.3.1}$$

the loglikelihood for a single observation is given by

$$\ell(\beta_0, f^*|\mathbf{x}, y, \mathbf{z}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}[y - \mathbf{x}\beta_0 - f^*(\mathbf{z})]^2 + \log g_0(\mathbf{x}, \mathbf{z}).$$

Now let $\beta \in \mathbf{B}$ represent the parameter of interest, and let $\beta \mapsto (\beta, \eta_\beta)$ be a smooth curve in $\mathbf{B} \times \overline{lin\mathcal{F}}$, such that $\eta_\beta|_{\beta=\beta_0} = f^*$. The vector (β, η_β) is called a parametric submodel. The word parametric here refers to the fact that since the nonparametric part is now indexed by β, the estimation problem is restricted to finite dimensional or parametric space. The term submodel simply means that (β, η_β) is just one of the several parameterizations that may be chosen. Notice that the parametric submodel passes through the truth when $\beta = \beta_0$.

Assuming that the data is generated by this parametric submodel, the loglikelihood function becomes

$$\ell(\beta, \eta_\beta|\mathbf{x}, y, \mathbf{z}) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}[y - \mathbf{x}\beta - \eta_\beta(\mathbf{z})]^2 + \log g_0(\mathbf{x}, \mathbf{z}). \tag{3.3.2}$$

Note that at β_0, this loglikelihood function equals the true likelihood. Hence the score function for β is,

$$\begin{aligned} S_{\beta_0} &= \frac{d\ell(\beta_0, \eta_{\beta_0})}{d\beta} \\ &= \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \eta_{\beta_0})}{\partial \eta}(\frac{d}{d\beta}\eta_{\beta_0}). \end{aligned}$$

Notice that even though S_{β_0} is a $(p \times 1)$ vector, from the discussion in Section 2.7 we know that it suffices to look at the componentwise scores. Therefore, for $i = 1, \ldots, p$, $S_{\beta_{i0}} = -\varepsilon[x_i + (\frac{d}{d\beta_i}\eta_{\beta_0})]$, with least favorable direction δ_i^* given by

$$\delta_i^* = \underset{\xi \in \overline{lin\, T(\mathcal{F}, f^*)}}{\operatorname{argmin}} \mathbb{E}\,[x_i + \xi]^2. \tag{3.3.3}$$

Hence the efficient score $\vec{S}$, for computing the semiparametric efficiency bounds of β_0 is

$$\vec{S} = \varepsilon \begin{pmatrix} x_1 - \delta_1^* \\ \vdots \\ x_p - \delta_p^* \end{pmatrix}.$$

The matrix $(\mathbb{E}\vec{S}\vec{S}')^{-1}$, then gives the semiparametric efficiency bounds for regular estimators of β_0.

However, merely knowing the bounds is not enough. To be of any use, the bounds must be attainable. We now discuss the construction of estimators that achieve these efficiency bounds.

So let $L_n(\beta, \eta_\beta) = \sum_{i=1}^n \ell(\beta, \eta_\beta | \mathbf{x}_i, y_i, \mathbf{z}_i)$ denote the empirical loglikelihood function for the data generated by the parametric submodel.

Now if we knew η_β, maximizing $L_n(\beta, \eta_\beta)$ would lead to an estimate of β, with asymptotic variance depending on η_β. And since m.l.e. is efficient in the parametric case, the asymptotic variance of this estimator would just be the inverse of the Fisher information for β. However, the fact of the matter is that we do not know η_β, and so straightforward maximization of the likelihood is infeasible. To get feasible estimates we adopt the approach given in Severini and Wong (1992).

So fix β, and let $\hat{\eta}_\beta$ denote a consistent estimator of η_β. Then $L_n(\beta, \hat{\eta}_\beta)$ is called the profile (or concentrated) likelihood for β. We now show that maximizing $L_n(\beta, \hat{\eta}_\beta)$ leads to an efficient estimate of β, provided that $\hat{\eta}_\beta$ is an estimator of a least favorable curve.

Since we want $\hat{\beta}_n$ — obtained by maximizing $L_n(\beta, \hat{\eta}_\beta)$ — to have the same asymptotic distribution as the estimator obtained by maximizing $L_n(\beta, \eta_\beta)$, we require that $L_n(\beta, \hat{\eta}_\beta)$ and $L_n(\beta, \eta_\beta)$ have the same local behavior at $\beta = \beta_0$. In particular, and this is evident from the standard way of proving asymptotic normality, we require $[n^{-1/2}\frac{dL_n(\beta_0, \hat{\eta}_{\beta_0})}{d\beta} - n^{-1/2}\frac{dL_n(\beta_0, \eta_{\beta_0})}{d\beta}] = o_p(1)$. But,

$$\begin{aligned}
n^{-1/2}\left[\frac{dL_n(\beta_0, \hat{\eta}_{\beta_0})}{d\beta} - \frac{dL_n(\beta_0, \eta_{\beta_0})}{d\beta}\right] &= n^{-1/2}\frac{d}{d\beta}\left[L_n(\beta_0, \hat{\eta}_{\beta_0}) - L_n(\beta, \eta_\beta)\right] \\
&\approx n^{-1/2}\frac{d}{d\beta}\left[\frac{\partial L_n(\beta_0, \eta_{\beta_0})}{\partial \eta}(\hat{\eta}_{\beta_0} - \eta_{\beta_0})\right] \\
&= \underbrace{n^{-1/2}\frac{d}{d\beta}\frac{\partial L_n(\beta_0, \eta_{\beta_0})}{\partial \eta}(\hat{\eta}_{\beta_0} - \eta_{\beta_0})}_{\text{Term I}} \\
&\quad + \underbrace{n^{-1/2}\frac{\partial L_n(\beta_0, \eta_{\beta_0})}{\partial \eta}\left(\frac{d}{d\beta}\hat{\eta}_{\beta_0} - \frac{d}{d\beta}\eta_{\beta_0}\right)}_{\text{Term II}}.
\end{aligned}$$

Now the Fréchet derivative

$$\frac{\partial \ell(\beta, \eta_\beta; \mathbf{x}_i, y_i, \mathbf{z}_i)}{\partial \eta} = y_i - \mathbf{x}_i\beta - \eta_\beta(\mathbf{z}_i),$$

and therefore

$$\left[\frac{d}{d\beta}\frac{\partial L_n(\beta,\eta_\beta)}{\partial\eta}\right](\hat{\eta}_{\beta_0}-\eta_{\beta_0}) = \sum_{i=1}^{n}\left[\frac{d}{d\beta}\{y_i - \mathbf{x}_i\beta - \eta_\beta(\mathbf{z}_i)\}\right](\hat{\eta}_{\beta_0}(\mathbf{z}_i)-\eta_{\beta_0}(\mathbf{z}_i))$$

$$= -\sum_{i=1}^{n}\left[\mathbf{x}_i + \frac{d}{d\beta}\eta_\beta(\mathbf{z}_i)\right](\hat{\eta}_{\beta_0}(\mathbf{z}_i)-\eta_{\beta_0}(\mathbf{z}_i)).$$

Thus evaluated at $\boldsymbol{\beta}_0$,

$$n^{-1/2}\frac{d}{d\beta}\frac{\partial L_n(\beta_0,\eta_{\beta_0})}{\partial\eta} = -n^{-1/2}\sum_{i=1}^{n}[\mathbf{x}_i + \frac{d}{d\beta}\eta_{\beta_0}(\mathbf{z}_i)](\hat{\eta}_{\beta_0}(\mathbf{z}_i)-\eta_{\beta_0}(\mathbf{z}_i)).$$

Now suppose that we can show that $||\hat{\eta}_{\beta_0} - \eta_{\beta_0}|| = o_p(n^{-\alpha_1})$ for some $\alpha_1 > 0$. So consider the sum

$$n^{-1/2}\sum_{i=1}^{n}(\mathbf{x}_i + \frac{d}{d\beta}\eta_{\beta_0})n^{\alpha_1}(\hat{\eta}_{\beta_0}-\eta_{\beta_0}).$$

If η_β is chosen to be a least favorable curve, then by Assumption 2.6.1 and the least favorable curve property of η_β

$$\mathbb{E}\,(\mathbf{x}_i + \frac{d}{d\beta}\eta_{\beta_0})n^{\alpha_1}(\hat{\eta}_{\beta_0}-\eta_{\beta_0}) = \mathbf{0},$$

and the terms in the above mentioned sum are centered around their means. Therefore, by applying a uniform CLT valid in function spaces

$$n^{-1/2}\sum_{i=1}^{n}(\mathbf{x}_i + \frac{d}{d\beta}\eta_{\beta_0})n^{\alpha_1}(\hat{\eta}_{\beta_0}-\eta_{\beta_0}) = O_p(1),$$

and this implies that

$$n^{-1/2}\sum_{i=1}^{n}(\mathbf{x}_i + \frac{d}{d\beta}\eta_{\beta_0})(\hat{\eta}_{\beta_0}-\eta_{\beta_0}) = o_p(1).$$

Hence Term I is $o_p(1)$. Now let us look at Term II. We again show that if $||\frac{d}{d\beta_0}\hat{\eta}_{\beta_0} - \frac{d}{d\beta_0}\eta_{\beta_0}|| = o_p(n^{-\alpha_2})$ for some $\alpha_2 > 0$ then Term II is also $o_p(1)$. To see this, first notice that by Remark 2.6.1 for any $\xi \in \overline{lin\,T(\mathcal{F},f^*)}$ there exists a curve $\eta_\beta \in \overline{lin\mathcal{F}}$ with tangent ξ, such that $\eta_{\beta_0} = f^*$. Then by using the unbiased property of the score functions

$$
\begin{aligned}
0 &= \mathbb{E}\frac{d}{d\beta}\ell(\beta_0,\eta_{\beta_0}) \\
&= \mathbb{E}[\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\xi)] \\
&= \mathbb{E}\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\xi),
\end{aligned}
$$

since $\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\beta}$ is just the parametric score function. Therefore, the terms in

$$n^{-1/2}\frac{\partial L_n(\beta_0,\eta_{\beta_0})}{\partial\eta}n^{\alpha_2}\left(\frac{d}{d\beta}\hat{\eta}_{\beta_0} - \frac{d}{d\beta}\eta_{\beta_0}\right)$$

are also centered around their means, and by applying a functional CLT we can again show that

$$n^{-1/2}\frac{\partial L_n(\beta_0,\eta_{\beta_0})}{\partial\eta}n^{\alpha_2}\left(\frac{d}{d\beta}\hat{\eta}_{\beta_0} - \frac{d}{d\beta}\eta_{\beta_0}\right) = O_p(1),$$

implying that

$$n^{-1/2}\frac{\partial L_n(\beta_0,\eta_{\beta_0})}{\partial\eta}\left(\frac{d}{d\beta}\hat{\eta}_{\beta_0} - \frac{d}{d\beta}\eta_{\beta_0}\right) = o_p(1).$$

Hence Term II is also $o_p(1)$, and

$$\left[n^{-1/2}\frac{dL_n(\beta_0,\hat{\eta}_{\beta_0})}{d\beta}\right] = \left[n^{-1/2}\frac{dL_n(\beta_0,\eta_{\beta_0})}{d\beta}\right] + o_p(1). \tag{3.3.4}$$

Therefore, if $\hat{\eta}_\beta$ is an estimate of a least favorable curve, maximizing $L_n(\beta,\hat{\eta}_\beta)$ yields the same asymptotic result as maximizing $L_n(\beta,\eta_\beta)$, and under certain regularity conditions we can show that $\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} \mathrm{N}(0, I_{\beta_0}^{-1})$, where $I_{\beta_0}^{-1}$ is the semiparametric information for β_0. Thus, in order to do feasible maximum likelihood estimation, η_β must be a least favorable curve otherwise the terms in Term I and Term II will not be centered around their means. Hence we would not be able to apply a CLT, and this approach would fail.

REMARK 3.3.1. Notice that this argument also indicates that to obtain efficient estimators of β_0, we need estimators of the nonparametric part (and their derivatives) that are consistent with rates of convergence faster than n^0. We will obtain such convergence rates by using kernel estimators. □

3.4. An Impossibility Theorem

As the parametric examples in Section 2.5 demonstrated, when $T(\mathcal{F}, f^*)$ is a proper cone projection on $T(\mathcal{F}, f^*)$ led to efficiency bounds that were unattainable by the m.l.e.. In this section we make that argument rigorous for semiparametric models. That is, we will now show that if efficiency bounds for estimating β_0 are obtained by projecting the parametric scores on a proper tangent cone, then no regular $n^{1/2}$ - consistent estimator of β_0 can achieve these bounds.

Notation 3.4.1. Let λ_β be a curve in $\mathcal{F}$ such that $\lambda_{\beta_0} = f^*$, and let δ be any vector in $\mathbb{R}^p$. □

The above mentioned result will be shown to hold under the following condition.

Assumption 3.4.1 (Proper Tangent Cone). *Let the matrix $I_1 - I_2$ be negative definite, where,*

$$I_1 =$$

$$\inf_{\xi \in \times_{i=1}^{p} lin\, T(\mathcal{F}, f^*)} \mathbb{E}\left[\frac{\partial \ell(\beta_0, \lambda_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \lambda_{\beta_0})}{\partial \lambda}(\xi)\right]\left[\frac{\partial \ell(\beta_0, \lambda_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \lambda_{\beta_0})}{\partial \lambda}(\xi)\right]'$$

and,

$$I_2 =$$

$$\inf_{\xi \in \times_{i=1}^{p} T(\mathcal{F}, f^*)} \mathbb{E}\left[\frac{\partial \ell(\beta_0, \lambda_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \lambda_{\beta_0})}{\partial \lambda}(\xi)\right]\left[\frac{\partial \ell(\beta_0, \lambda_{\beta_0})}{\partial \beta} + \frac{\partial \ell(\beta_0, \lambda_{\beta_0})}{\partial \lambda}(\xi)\right]' .$$

□

Remark 3.4.1. (i) In the above assumption, the infima are taken w.r.t. the usual order on the space of all $p \times p$ matrices. From Theorem 2.7.3, I_1 and I_2 exist if λ_{β_0} is a least favorable curve in $\overline{lin\mathcal{F}}$ and $\mathcal{F}$, respectively. Therefore, a necessary condition for this assumption to hold is that the least favorable directions in $\times_{i=1}^{p}\overline{lin\,T(\mathcal{F},f^{*})}$ and $\times_{i=1}^{p}T(\mathcal{F},f^{*})$ be different.

(ii) This condition certainly holds for the parametric examples provided before. It also holds for the semiparametric models under study. For instance, let $\mathcal{F}$ be the set of concave functions in $\mathcal{H}$, and let f^{*} be affine. Then as can be seen from Section 3.7, I_2 is achieved by a tangent vector which is a concave function in $\mathcal{H}$, while I_1 is achieved by a tangent vector which is just a conditional expectation subject to smoothness conditions. □

Before we state the main result of this section, we define what we mean by a "regular" sequence of estimators.

DEFINITION 3.4.1 (REGULAR ESTIMATOR). Let $\beta_n = \beta_0 + n^{-1/2}\delta$. Then a sequence of estimators $\hat{\beta}_n$ is said to be regular if $n^{1/2}(\hat{\beta}_n - \beta_n)$ converges in distribution, under β_n, to a limiting distribution that does not depend upon δ. □

REMARK 3.4.2. Let ξ_β be any curve in $C^2(\mathbf{Z})$ through f^{*}. That is, $\xi_\beta \in C^2(\mathbf{Z})$ for all $\beta \in \mathbf{B}$, and $\xi_{\beta_0} = f^{*}$. Now perturb β such that the perturbed value, denoted by β_n, lies in a $n^{-1/2}$ neighborhood. That is, for any $\delta \in \mathbb{R}^p$, $\beta_n = \beta_0 + n^{-1/2}\delta$. Then by the mean value theorem

$$\xi_{\beta_n} = f^{*} + n^{-1/2}\delta'\frac{d}{d\beta}\xi_{\tilde{\beta}_n},$$

where, $\tilde{\beta}_n$ lies between β_0 and β_n. Therefore, $n^{-1/2}$ perturbations of the finite dimensional parameter generate $n^{-1/2}$ perturbations of the infinite dimensional parameter. However, it is clear that if $f^{*} \in \mathcal{F} \subset C^2(\mathbf{Z})$, the perturbation ξ_{β_n} need not lie in $\mathcal{F}$. □

The main result of this section is the following theorem.

THEOREM 3.4.1. *Let $\mathcal{F}$ be a convex cone in $\mathcal{H}$, and let $f^* \in \partial\mathcal{F}$ be such that the tangent cone $T(\mathcal{F}, f^*)$ is a proper cone. Then under Assumption* 3.4.1, *no regular $n^{1/2}$ consistent estimator of β_0 can achieve the efficiency bounds obtained by projecting the parametric scores onto the tangent cone $T(\mathcal{F}, f^*)$.*

PROOF. See Appendix C. □

REMARK 3.4.3. The utility of this result will become evident in Section 3.7, when we impose concavity upon f^*. In Section 3.7

$$\mathcal{F} = \{f \in \mathcal{H} : f \text{ is concave}\},$$

and $\mathcal{F}$ is a proper cone, i.e. it is not a linear space. Now suppose that f^* lies on the boundary of $\mathcal{F}$, say for instance f^* is affine. As can be seen from Secref 3.7, when f^* is affine $T(\mathcal{F}, f^*) = \mathcal{F}$. Hence by Theorem 3.4.1, projection of the parametric scores onto $T(\mathcal{F}, f^*)$ will lead to efficiency bounds that are unattainable by any regular $n^{1/2}$ - consistent estimator of β_0. To obtain attainable efficiency bounds, we have to project onto $\overline{lin\, T(\mathcal{F}, f^*)}$ which is just $\mathcal{H}$. Hence, concavity of f^* does not help us in estimating β_0 more efficiently. □

3.5. Consistency and Asymptotic Normality of $\hat{\beta}_n$

Closely following Severini and Wong (1992) in this section, we will show that the estimator $\hat{\beta}_n$ obtained by maximizing the profile likelihood is consistent and asymptotically normal.

NOTATION 3.5.1. With $L_n(\beta, \eta_\beta)$ as defined before, let,

$$L_n(\hat{\beta}_n, \hat{\eta}_{\hat{\beta}_n}) = \sup_{\beta \in B} L_n(\beta, \hat{\eta}_\beta). \quad \square$$

REMARK 3.5.1. Unless otherwise specified, all expectations are taken under the truth i.e. under (β_0, f^*). □

Assumption 3.5.1 (Identification). *Let*

$$\mathbb{K}(\beta,\beta_0) = \mathbb{E}\,\ell(\beta,\eta_\beta;\mathbf{x},y,\mathbf{z}) - \mathbb{E}\,\ell(\beta_0,f^*;\mathbf{x},y,\mathbf{z}).$$

Then,

(i) $\mathbb{K}(\beta,\beta_0) < 0$, *if* $\beta \neq \beta_0$.

(ii) β_0 *is the unique global maximum of* $\mathbb{K}(\beta,\beta_0)$. *That is,*

$$\sup_{\beta\in B} \mathbb{E}\,\ell(\beta,\eta_\beta;\mathbf{x},y,\mathbf{z}) = \mathbb{E}\,\ell(\beta_0,f^*;\mathbf{x},y,\mathbf{z}). \quad \Box$$

Remark 3.5.2. Because we had shown that β_0 was identified in Section 3.2, this assumption holds for the partially linear model. □

Assumption 3.5.2 (Smoothness). *For* $i,j = 1,\dots,p$, *let*

(i) $\mathbb{E}\left\{\sup_{\beta\in\mathbf{B}}\sup_{\eta\in\overline{lin\mathcal{F}}}|\frac{\partial\ell(\beta,\eta;\mathbf{x},y,\mathbf{z})}{\partial\eta}|^2\right\} < \infty$,

(ii) $\mathbb{E}\left\{\sup_{\beta\in\mathbf{B}}\sup_{\eta\in\overline{lin\mathcal{F}}}|\frac{\partial^2\ell(\beta,\eta;\mathbf{x},y,\mathbf{z})}{\partial\beta_i\partial\eta}|^2\right\} < \infty$,

(iii) $\mathbb{E}\left\{\sup_{\beta\in\mathbf{B}}\sup_{\eta\in\overline{lin\mathcal{F}}}|\frac{\partial^3\ell(\beta,\eta;\mathbf{x},y,\mathbf{z})}{\partial\beta_i\partial\beta_j\partial\eta}|^2\right\} < \infty$,

(iv) $\mathbb{E}\left\{\sup_{\beta\in\mathbf{B}}\sup_{\eta\in\overline{lin\mathcal{F}}}|\frac{\partial^2\ell(\beta,\eta;\mathbf{x},y,\mathbf{z})}{\partial\eta^2}|^2\right\} < \infty$,

(v) $\mathbb{E}\left\{\sup_{\beta\in\mathbf{B}}\sup_{\eta\in\overline{lin\mathcal{F}}}|\frac{\partial^3\ell(\beta,\eta;\mathbf{x},y,\mathbf{z})}{\partial\beta_i\partial\eta^2}|^2\right\} < \infty$. □

Remark 3.5.3. Let us verify (i). Other conditions can be similarly checked. Since the Fréchet derivative $\frac{\partial\ell(\beta,\eta_\beta;\mathbf{x},y,\mathbf{z})}{\partial\eta} = y - \mathbf{x}\beta - \eta_\beta(\mathbf{z})$,

$$\begin{aligned}\left|\frac{\partial\ell(\beta,\eta_\beta;\mathbf{x},y,\mathbf{z})}{\partial\eta}\right|^2 &\leq 2|y-\mathbf{x}\beta|^2 + 2|\eta_\beta(\mathbf{z})|^2\\ &\leq 4|y|^2 + 4p^2\|x\|\|\beta\| + 2|\eta_\beta(\mathbf{z})|^2.\end{aligned}$$

Hence, for all $\mathbf{z}$

$$\mathbb{E}\left\{\sup_{\boldsymbol{\beta}\in\mathbf{B}}\sup_{\eta\in\mathcal{H}}\left|\frac{\partial\ell(\boldsymbol{\beta},\eta_{\boldsymbol{\beta}};\mathbf{x},y,\mathbf{z})}{\partial\eta}\right|^2\right\} \leq 4\mathbb{E}\,|y|^2 + 4p^2\sup_{\mathbf{x}\in\mathbf{X}}\|\mathbf{x}\|\sup_{\boldsymbol{\beta}\in\mathbf{B}}\|\boldsymbol{\beta}\| + 2\sup_{\eta\in\mathcal{H}}|\eta|^2$$
$$< \infty,$$

since $(\mathbf{x},\mathbf{z},\boldsymbol{\beta},\eta)$ all come from compact sets. □

Assumption 3.5.3 (Nuisance Parameters). *Let a least favorable curve be given by $\eta_{\boldsymbol{\beta}}$, and let $\hat{\eta}_{\boldsymbol{\beta}}$ be a consistent estimator of $\eta_{\boldsymbol{\beta}}$. Then for some $\alpha_1, \alpha_2 \geq \frac{1}{4}$, $\delta > 0$, and $i = 1, \ldots, p$, assume that*

(i) $\sup_{u,v}|\hat{\eta}_{\boldsymbol{\beta}_0}(u,v) - \eta_{\boldsymbol{\beta}_0}(u,v)| = o(n^{-\alpha_1})$,

(ii) $\sup_{u,v}|\frac{d}{d\beta_i}\hat{\eta}_{\boldsymbol{\beta}_0}(u,v) - \frac{d}{d\beta_i}\eta_{\boldsymbol{\beta}_0}(u,v)| = o(n^{-\alpha_2})$,

(iii) $\sup_{\boldsymbol{\beta}\in\mathbf{B}}\sup_{u,v\in\mathbf{Z}}|\hat{\eta}_{\boldsymbol{\beta}}(u,v) - \eta_{\boldsymbol{\beta}}(u,v)| = o_p(1)$,

(iv) $\sup_{\boldsymbol{\beta}\in\mathbf{B}}\sup_{u,v\in\mathbf{Z}}|\frac{d}{d\beta_i}\hat{\eta}_{\boldsymbol{\beta}}(u,v) - \frac{d}{d\beta_i}\eta_{\boldsymbol{\beta}}(u,v)| = o_p(1)$,

(v) $\sup_{\boldsymbol{\beta}\in\mathbf{B}}\sup_{u,v\in\mathbf{Z}}|\frac{d^2}{d\beta_i^2}\hat{\eta}_{\boldsymbol{\beta}}(u,v) - \frac{d^2}{d\beta_i^2}\eta_{\boldsymbol{\beta}}(u,v)| = o_p(1)$,

(vi) $\sup_{\boldsymbol{\beta}\in\mathbf{B}}\sup_{u,v\in\mathbf{Z}}|\frac{\partial}{\partial\xi}\hat{\eta}_{\boldsymbol{\beta}}(u,v) - \frac{\partial}{\partial\xi}\eta_{\boldsymbol{\beta}}(u,v)| = o_p(n^{-\delta})$, $\xi \in \{u,v\}$, *and*

(vii) $\sup_{\boldsymbol{\beta}\in\mathbf{B}}\sup_{u,v\in\mathbf{Z}}|\frac{\partial}{\partial\xi}\frac{d}{d\beta_i}\hat{\eta}_{\boldsymbol{\beta}}(u,v) - \frac{\partial}{\partial\xi}\frac{d}{d\beta_i}\eta_{\boldsymbol{\beta}}(u,v)| = o_p(n^{-\delta})$, $\xi \in \{u,v\}$. □

We then have the following results.

Theorem 3.5.1. *$\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$, as $n \to \infty$.*

Proof. See Appendix D. □

Theorem 3.5.2. *$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(0, I_{\boldsymbol{\beta}_0}^{-1})$.*

Proof. See Appendix D. □

Theorem 3.5.3. *$\hat{\boldsymbol{\beta}}_n$ is regular.*

Proof. See Appendix D. □

3.6. The Case of Homogeneity

Let us now examine how certain shape restrictions on the unknown function, that are important in economic theory, influence the efficiency bounds for finite dimensional parameters. We begin with the case when f^* is a homogeneous function of degree r.

So let $\mathcal{F}$ be the set of functions in $\mathcal{H}$ which are homogeneous of degree r, i.e.

$$\mathcal{F} = \{f \in \mathcal{H} : f(\lambda \mathbf{z}) = \lambda^r f(\mathbf{z}), \lambda \geq 0\}.$$

Then the solution to (3.3.3) is a projection of x_i onto $\overline{lin\, T(\mathcal{F}, f^*)}$. Since $\mathcal{F}$ is a closed linear space, from Corollary A.1 we have $T(\mathcal{F}, f^*) = \mathcal{F}$ which implies that $\overline{lin\, T(\mathcal{F}, f^*)} = \mathcal{F}$. Therefore, we simply project x_i onto $\mathcal{F}$. The solution to this projection problem is given by the following theorem.

THEOREM 3.6.1. *The projection of x_i onto $\mathcal{F}$ is the function*

$$\delta^*(u,v) = -\frac{v^r\, \mathbb{E}(x_i z_2^r | \frac{z_1}{z_2} = \frac{u}{v})}{\mathbb{E}(z_2^{2r} | \frac{z_1}{z_2} = \frac{u}{v})}.$$

PROOF. See Appendix E. □

REMARK 3.6.1. Notice that if we had taken the projection as [1] $-\frac{u^r\, \mathbb{E}(x_i z_1^r | \frac{z_1}{z_2} = \frac{u}{v})}{\mathbb{E}(z_1^{2r} | \frac{z_1}{z_2} = \frac{u}{v})}$, the uniqueness of the projections (postulated by the classical projection theorem) would imply that

$$-\frac{u^r\, \mathbb{E}(x_i z_1^r | \frac{z_1}{z_2} = \frac{u}{v})}{\mathbb{E}(z_1^{2r} | \frac{z_1}{z_2} = \frac{u}{v})} = -\frac{v^r\, \mathbb{E}(x_i z_2^r | \frac{z_1}{z_2} = \frac{u}{v})}{\mathbb{E}(z_2^{2r} | \frac{z_1}{z_2} = \frac{u}{v})}.$$

That this is indeed the case, can be seen as follows.

[1] That this is a valid projection can be seen immediately, since this also satisfies the necessary and sufficient conditions of the classical projection theorem.

$$\frac{u^r\,\mathbb{E}\left(x_i z_1^r \middle| \frac{z_1}{z_2} = \frac{u}{v}\right)}{\mathbb{E}\left(z_1^{2r} \middle| \frac{z_1}{z_2} = \frac{u}{v}\right)} = \frac{u^r\,\mathbb{E}\left(z_1^r \frac{z_2^r}{z_2^r} x_i \middle| \frac{z_1}{z_2} = \frac{u}{v}\right)}{\mathbb{E}\left(z_1^{2r} \frac{z_2^{2r}}{z_2^{2r}} \middle| \frac{z_1}{z_2} = \frac{u}{v}\right)}$$

$$= \frac{u^r \frac{u^r}{v^r}\,\mathbb{E}\left(x_i z_2^r \middle| \frac{z_1}{z_2} = \frac{u}{v}\right)}{\frac{u^{2r}}{v^{2r}}\mathbb{E}\left(z_2^{2r} \middle| \frac{z_1}{z_2} = \frac{u}{v}\right)}$$

$$= \frac{v^r\,\mathbb{E}\left(x_i z_2^r \middle| \frac{z_1}{z_2} = \frac{u}{v}\right)}{\mathbb{E}\left(z_2^{2r} \middle| \frac{z_1}{z_2} = \frac{u}{v}\right)}.$$

This is a nice test of the validity of the result obtained in Theorem 3.6.1. □

From Theorem 3.6.1 the efficient score $\vec{S}$ for computing the semiparametric efficiency bounds of β_0 is

$$\vec{S} = \varepsilon \begin{pmatrix} x_1 - \frac{z_2^r\,\mathbb{E}(x_1 z_2^r | \frac{z_1}{z_2})}{\mathbb{E}(z_2^{2r} | \frac{z_1}{z_2})} \\ \vdots \\ x_p - \frac{z_2^r\,\mathbb{E}(x_p z_2^r | \frac{z_1}{z_2})}{\mathbb{E}(z_2^{2r} | \frac{z_1}{z_2})} \end{pmatrix}.$$

The matrix $(\mathbb{E}\vec{S}\vec{S}')^{-1}$, then gives the semiparametric efficiency bounds for β_0 when the true function f^* is homogeneous of degree r. A natural question at this point is to inquire about the gain in efficiency obtained by imposing the shape restriction of homogeneity. The following example provides an interesting case in point.

Example 3.6.1 (Homogeneity Increases Efficiency). Let $\beta \in \mathbb{R}^2$. Our model is then $y = x_1\beta_1 + x_2\beta_2 + f^*(z_1, z_2) + \varepsilon$, with $\varepsilon \overset{d}{=} \mathrm{N}(0,1)$.

To simplify matters even further, let $z_1 = z_2 = z$, and let $\mathbf{x} = (x_1, x_2)$ be completely predictable by z. Say for instance, $x_1 = z^2$ and $x_2 = z^3$. The model then reduces to

$$y = z^2\beta_1 + z^3\beta_2 + f^*(z,z) + \varepsilon.$$

Now consider the following two cases.

Case I: <u>No shape restrictions on f^*.</u>

That is, just assume that $f^* \in \mathcal{H}$. We claim that in this case, (β_1, β_2, f^*) is not identified. To see this, let g be the joint density of (y, z) and define

$$\begin{aligned} S_1 &\equiv (\beta_1, \beta_2, f^*) \\ S_2 &\equiv (\alpha_1, \alpha_2, h^*). \end{aligned}$$

Then if we can show that there exist structures S_1, S_2 with $S_1 \neq S_2$ such that $g(y, z; S_1) = g(y, z; S_2)$, (β_1, β_2, f^*) is not identified. So let $S_1 = (1, 0, f^*(z, z))$ and $S_2 = (1, 1, f^*(z, z) - z^3)$. Clearly $S_1 \neq S_2$, and since $g(y, z) = g(y|z)g(z)$ we have,

$$\begin{aligned} g(y, z; S_1) &= (2\pi)^{-1/2} \exp\{-\frac{1}{2}(y - z^2 - z^3 \cdot 0 - f^*(z, z))^2\} g(z) \\ &= (2\pi)^{-1/2} \exp\{-\frac{1}{2}(y - z^2 - f^*(z, z))^2\} g(z) \quad \text{and,} \\ g(y, z; S_2) &= (2\pi)^{-1/2} \exp\{-\frac{1}{2}(y - z^2 - z^3 - f^*(z, z) + z^3)^2\} g(z) \\ &= (2\pi)^{-1/2} \exp\{-\frac{1}{2}(y - z^2 - f^*(z, z))^2\} g(z). \end{aligned}$$

Therefore, $g(y, z; S_1) = g(y, z; S_2)$ and so (β_1, β_2, f^*) is not identified. It is not difficult to see that this also implies that both (β_1, β_2) and f^* are separately not identified. Due to this lack of identification, the lower bound for the variance of any estimator of (β_1, β_2) is (∞, ∞).

Case II: Now let f^* be homogeneous of degree 1.

Since f^* is homogeneous of degree 1,

$$y = z^2\beta_1 + z^3\beta_2 + z f^*(1, 1) + \varepsilon.$$

But this clearly shows that in this case $(\beta_1, \beta_2, f^*(1, 1))$ is identified. Therefore, the efficiency bounds for estimators of β_1 and β_2 are finite.

Hence, by imposing a shape restriction on f^* we can identify the finite dimensional parameters and achieve a dramatic gain in efficiency. □

REMARK 3.6.2. Even though this example has an artificial flavor, it illustrates the potential gains in efficiency that may be obtained by imposing shape restrictions. An interesting exercise here would be to examine what shape restrictions on f^*, besides homogeneity, allow us to identify β. Notice that this example also illustrates the strength of homogeneity as a shape restriction. □

Here is another example demonstrating that large gains in efficiency are possible under homogeneity, even when the parameter of interest is identified.

EXAMPLE 3.6.2 (ANOTHER SIMPLE EXAMPLE). For $(\beta_0, z) \in \mathbb{R} \times \mathbb{R}$, and $\varepsilon \overset{d}{=} \mathrm{N}(0,1)$, let $y = x\beta_0 + f^*(z) + \varepsilon$, where $x, z \overset{d}{=} \mathrm{UIID}(0,1)$, and f^* is linearly homogeneous. Now since f^* is homogeneous of degree one, $y = x\beta_0 + zf^*(1) + \varepsilon$, and the lower bound for the variance of a regular estimator of β_0 can be shown to be 8.4. However, if homogeneity is not imposed upon f^* it is easy to see that β_0 still remains identified, but the lower bound increases to 12. Therefore, the asymptotic relative efficiency of the estimator under homogeneity w.r.t. the estimator when homogeneity is not imposed is $\frac{12}{8.4} = 1.428$. Thus the loss in efficiency by not imposing homogeneity, when f^* is truly homogeneous, is 42.8%. □

To show that the bounds obtained in the beginning of this section are meaningful, we now construct an estimator of β_0 that achieves these bounds. As discussed in Section 3.3, we need $\hat{\eta}_\beta$ (the estimator of a least favorable curve) to efficiently estimate β_0. Once we have $\hat{\eta}_\beta$, we can estimate β_0 by maximizing the empirical profile likelihood. So let $\hat{\eta}_\beta$ be a consistent estimator of a least favorable curve η_β. If

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n &= \arg\max_{\boldsymbol{\beta}} -\frac{1}{2}\sum_{i=1}^{n}[y_i - \mathbf{x}_i\boldsymbol{\beta} - \hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z}_i)]^2 + \text{constant} \\ &= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}[y_i - \mathbf{x}_i\boldsymbol{\beta} - \hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z}_i)]^2,\end{aligned}$$

then as discussed before, $\hat{\boldsymbol{\beta}}_n$ is an efficient estimator of $\boldsymbol{\beta}_0$. We now define a least favorable curve, and also propose an intuitive estimator of the least favorable curve.

Assumption 3.6.1. *Let* $\mathrm{K}(\cdot)$ *be a positive, real valued function on* $\mathbb{R}$ *such that,*

(i) $\mathrm{K}(\cdot)$ *vanishes outside* $[-1, 1]$,

(ii) $\sup_{s\in(-1,1)} |\mathrm{K}'(s)| < \infty$, *and* $\sup_{s\in(-1,1)} |\mathrm{K}''(s)| < \infty$,

(iii) $\int_{-1}^{1} \mathrm{K}(s)\, ds = 1$,

(iv) $\int_{-1}^{1} s\mathrm{K}(s)\, ds = 0$,

(v) $\int_{-1}^{1} s^2\mathrm{K}(s)\, ds < \infty$. □

Assumption 3.6.2. *Let* a_n *be a sequence of positive numbers (the "window width") such that* $a_n \to 0$ *and* $na_n \to \infty$. □

Proposition 3.6.1. *With* $\mathrm{K}(\cdot)$ *and* a_n *as defined, for any* $(u, v) \in Z_1 \times Z_2$ *let*

$$\begin{aligned}\eta_{\boldsymbol{\beta}}(u,v) &= \frac{v^r\mathbb{E}\left[y_j z_{2j}^r | \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} | \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} - \sum_{i=1}^{p}\beta_i \frac{v^r\mathbb{E}\left[x_{ij} z_{2j}^r | \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} | \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} \\ \hat{\eta}_{\boldsymbol{\beta}}(u,v) &= \frac{v^r\sum_{j=1}^{n} y_j z_{2j}^r \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])}{\sum_{j=1}^{n} z_{2j}^{2r}\mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])} - \sum_{i=1}^{p}\beta_i \frac{v^r\sum_{j=1}^{n} x_{ij} z_{2j}^r \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])}{\sum_{j=1}^{n} z_{2j}^{2r}\mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])}.\end{aligned}$$

Then $\eta_{\boldsymbol{\beta}}$ *is a least favorable curve, and* $\hat{\eta}_{\boldsymbol{\beta}}$ *is a consistent estimator of* $\eta_{\boldsymbol{\beta}}$.

Proof. See Appendix E. □

The estimation problem yielding efficient estimates of $\boldsymbol{\beta}_0$ is then

$$\hat{\beta}_n = \operatorname*{argmin}_{\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \left[y_i - \sum_{k=1}^{p} x_{ki}\beta_k - \frac{z_{2i}^r \sum_{j=1}^n y_j z_{2j}^r \mathrm{K}(\frac{1}{a_n}[\frac{z_{1i}}{z_{2i}} - \frac{z_{1j}}{z_{2j}}])}{\sum_{j=1}^n z_{2j}^{2r} \mathrm{K}(\frac{1}{a_n}[\frac{z_{1i}}{z_{2i}} - \frac{z_{1j}}{z_{2j}}])} \right.$$
$$\left. - \sum_{k=1}^{p} \beta_k \frac{z_{2i}^r \sum_{j=1}^n x_{kj} z_{2j}^r \mathrm{K}(\frac{1}{a_n}[\frac{z_{1i}}{z_{2i}} - \frac{z_{1j}}{z_{2j}}])}{\sum_{j=1}^n z_{2j}^{2r} \mathrm{K}(\frac{1}{a_n}[\frac{z_{1i}}{z_{2i}} - \frac{z_{1j}}{z_{2j}}])} \right]^2 .$$

Thus to make sure that the consistency and asymptotic normality results in Section 3.5 still hold, we have only to verify Assumption 3.5.3. This is done using the following result.

THEOREM 3.6.2. *Let the random variable z_1/z_2 have support T and p.d.f. $p(t)$. Also, let*

(i′) $\psi(t) = \mathbb{E}\,[y z_2^r | \frac{z_1}{z_2} = t] p(t)$,

(ii′) $\mu(t) = \mathbb{E}\,[z_2^{2r} | \frac{z_1}{z_2} = t] p(t)$,

(iii′) $\rho_i(t) = \mathbb{E}\,[x_i z_2^r | \frac{z_1}{z_2} = t] p(t)$, *for* $i = 1, \ldots, p$, *and*

let η_β be a least favorable curve consistently estimated by $\hat{\eta}_\beta$ given above. Furihermore, let $\frac{d}{d\beta}\eta_{\beta_0}$ denote the least favorable direction and assume that

(i) $\mathbb{E}\,|y|^q < \infty$, *for some* $q > 2$,

(ii) $\sup_{t\in T} \mathbb{E}\,[y^2|t] < \infty$,

(iii) $\sup_{t\in T} p(t) < \infty$,

(iv) $0 < \inf_{u\in Z_1} |u| \leq \sup_{u\in Z_1} |u| < \infty$,

(v) $0 < \inf_{v\in Z_2} |v| \leq \sup_{v\in Z_2} |v| < \infty$,

(vi) $0 < \sup_{t\in T} |\psi^{(j)}(t)| < \infty$, *for* $j = 0, 1, 2, 3$,

(vii) $0 < \sup_{t\in T} |\rho_i^{(j)}(t)| < \infty$, *for* $j = 0, 1$, *and* $i = 1, \ldots, p$,

(viii) $\inf_{t\in T} |\mu(t)| > 0$, $0 < \sup_{t\in T} |\mu'(t)| < \infty$.

Then for $\lambda > 0$, a sufficient condition to obtain

(1) $\sup_{\beta\in\mathrm{B}} \sup_{u,v\in\mathrm{Z}} |\hat{\eta}_\beta(u,v) - \eta_\beta(u,v)| = o_p(n^{-\lambda})$, *and*

(2) $\sup_{\beta \in \mathrm{B}} \sup_{u,v \in \mathrm{Z}} |\frac{d}{d\beta}\hat{\eta}_\beta(u,v) - \frac{d}{d\beta}\eta_\beta(u,v)| = o_p(n^{-\lambda})$,

is to choose the window width $a_n = n^{-\alpha}$, such that $\alpha > 0$ satisfies

$$\frac{\lambda}{2} \leq \alpha < \frac{(1-2\lambda)(q-1)}{q}.$$

Moreover, if $\xi \in \{u, v\}$, then to obtain

(1′) $\sup_{\beta \in \mathrm{B}} \sup_{u,v \in \mathrm{Z}} |\frac{\partial}{\partial \xi}\hat{\eta}_\beta(u,v) - \frac{\partial}{\partial \xi}\eta_\beta(u,v)| = o_p(n^{-\lambda})$, *and*

(2′) $\sup_{\beta \in \mathrm{B}} \sup_{u,v \in \mathrm{Z}} |\frac{\partial}{\partial \xi}\frac{d}{d\beta}\hat{\eta}_\beta(u,v) - \frac{\partial}{\partial \xi}\frac{d}{d\beta}\eta_\beta(u,v)| = o_p(n^{-\lambda})$,

it is again sufficient to choose $a_n = n^{-\alpha}$, with $\alpha > 0$ satisfying

$$\frac{\lambda}{2} \leq \alpha < \frac{(1-2\lambda)(q-1)}{3q-2}.$$

PROOF. See Appendix F □

3.7. The Case of Concavity

Let us now look at the case when the unknown function f^* is concave. We want to examine the relationship between the efficiency bounds for β_0 and the degree of concavity of f^*. That is, we want to find the efficiency bounds for estimating β_0 when,

(i) f^* is strictly concave, or

(ii) f^* is affine, or

(iii) when f^* is concave but not strictly concave.

We begin with a simpler problem to obtain a geometrical insight into the original problem. So as an illustration, we look at the space of C^2 concave functions on Z, a compact subset of $\mathbb{R}$. Later on we will obtain results for the case when $\mathrm{Z} \subset \mathbb{R}^2$, as in the case of homogeneity. Note that $\mathcal{F}$ is used to denote the set of concave functions on Z, even when Z is a subset of the real line. However, this should not cause any confusion. As usual, before proceeding we define some useful terms.

Definition 3.7.1 (Half-Space). If V is a Banach space and $f \in V^{*}, f \neq 0$, then $V_f^{+} = \{v \in V : f(v) \geq 0\}$ is called the (positive) half-space defined by f. $\partial V_f^{+} = \{v \in V : f(v) = 0\}$ is called the boundary of the half-space.

Definition 3.7.2 (Manifold with boundary). Let H be a half-space in a Banach space V. Then a smooth manifold with boundary (modeled on V) is a Hausdorff space M, such that open sets in M are diffeomorphic to open sets in H.

Remark 3.7.1. For each $u \in Z$, let $H_u = \{f \in \mathrm{C}^2(Z) : f''(u) \leq 0\}$. Then H_u is a closed half-space of $\mathrm{C}^2(Z)$ defined by f'', and $\partial H_u = \{f \in \mathrm{C}^2(Z) : f''(u) = 0\}$. Note that in general the differentiation operator is unbounded, but the use of the C^2 norm here, makes it into a bounded operator. Also note that open sets in H_u are of two types:

(i) those that contain points of ∂H_u (i.e., all $f \in \mathrm{C}^2(Z)$ such that, $f''(u) = 0$),

(ii) and those that do not. □

Now back to the geometry. We begin by noticing that $\mathcal{F}$ is a convex cone imbedded in $\mathrm{C}^2(Z)$. Since $\mathrm{C}^2(Z)$ is a Banach space, it is a smooth manifold. Now if we could somehow show that $\mathcal{F}$ was also a smooth manifold modeled on $\mathrm{C}^2(Z)$, then any point in $\mathcal{F}$, and f^{*} in particular, would have neighborhoods diffeomorphic to open sets in $\mathrm{C}^2(Z)$. This would imply that the tangent space at each point of $\mathcal{F}$ would be $\mathrm{C}^2(Z)$ itself. Therefore, projecting onto $\overline{lin\, T(\mathcal{F}, f^{*})}$ would be equivalent to projecting onto $\mathrm{C}^2(Z)$. Hence irrespective of the degree of concavity of f^{*}, the projection would just be a $\mathrm{C}^2(\mathrm{Z})$ function. In geometrical terms, we would be able to approach f^{*} from any direction and no gains in efficiency would occur.

Unfortunately, $\mathcal{F}$ is not a smooth manifold modeled on $\mathrm{C}^2(Z)$. Heuristically, this may be seen as follows. With H_u as defined in Remark 3.7.1, $\mathcal{F} = \cap_{u \in \mathrm{Z}} H_u$, and since the boundary of each H_u may be represented as a line in $\mathbb{R}^2$, $\mathcal{F}$ has the structure of a wedge. The cutting edge of this wedge (the "kink") is the collection of all linear and constant functions, including the zero function, while each face of this wedge is occupied by functions whose second derivative vanishes at that point. This wedge structure agrees with our intuition, since we know that $\mathcal{F}$ is a cone. Because of this structure, the location of f^* inside $\mathcal{F}$ (i.e. the degree of concavity of f^*) affects the direction from which we can approach it.

Since $\mathcal{F}$ is not a smooth manifold, the tangent space varies from point to point. Therefore, we first have to find $T(\mathcal{F}, f^*)$, the tangent cone to $\mathcal{F}$ at f^*. It is not unreasonable to expect that since $\mathcal{F}$ is a proper cone, $T(\mathcal{F}, f^*)$ will also be a proper cone. But keep in mind that we have to project onto $\overline{lin\, T(\mathcal{F}, f^*)}$, and not onto $T(\mathcal{F}, f^*)$. Once we obtain $T(\mathcal{F}, f^*)$, we will show that $\overline{lin\, T(\mathcal{F}, f^*)}$ is just $\mathcal{H}$. Hence projecting onto $\overline{lin\, T(\mathcal{F}, f^*)}$ is equivalent to projecting onto $\mathcal{H}$, and as far as the finite dimensional parameters are concerned there is no gain in efficiency from the concavity of f^*.

Now back to the original problem where we make this argument rigorous. So let $\mathcal{F}$ be the set of concave functions in $\mathcal{H}$, and $f^* \in \mathcal{F}$. Then to obtain the semiparametric information for the finite dimensional parameters, we have to project onto $\overline{lin\, T(\mathcal{F}, f^*)}$. Let us first determine the nature of this space. To do so, consider the set of functions defined below.

Let $\mathbf{Z}_0$ be a non empty subset of $\mathbf{Z}$, and let, [2]

$$\mathcal{W} = \{f \in \mathcal{H} : \det[\nabla^2 f(\mathbf{u})] = 0, \text{or} \nabla^2 f(\mathbf{u}) \text{is n.d. for all } \mathbf{u} \in \mathbf{Z}_0 \subset \mathbf{Z}.\}$$

REMARK 3.7.2. (i) The reason for defining $\mathcal{W}$ will soon be apparent.

(ii) Since the Hessian of $f \in \mathcal{W}$ is negative semi-definite on $\mathbf{Z}_0$, we can characterize $\mathcal{W}$ as the set of functions in $\mathcal{H}$ which are concave on $\mathbf{Z}_0 \subset \mathbf{Z}$. This implies that $\mathcal{F} \subset \mathcal{W}$. Notice that a function could be strictly convex and still be in $\mathcal{W}$ if the determinant

[2] The abbreviation "n.d." stands for "negative definite."

of its Hessian vanishes on $\mathbf{Z}_0$. For instance, in $\mathbb{R}^2$ the function $(x,y) \mapsto x^4 + y^4$ is strictly convex, but its Hessian is zero at $(0,0)$. Hence if $\mathbf{Z}_0 = \{(0,0)\}$, then $(x,y) \mapsto x^4 + y^4$ is an element of $\mathcal{W}$.

(iii) $\mathcal{W}$ is a closed convex cone and not a linear space, since all strictly concave functions are in $\mathcal{W}$ while some strictly convex functions are not. □

We now have the following results.

THEOREM 3.7.1. *Let $\mathcal{F}$ be the set of concave functions in $\mathcal{H}$, and let $f^* \in \mathcal{F}$. Then with $\mathcal{W}$ as defined above,*

$$T(\mathcal{F}, f^*) = \begin{cases} \mathcal{H} & \textit{if } f^* \textit{ is strictly concave on } \mathbf{Z}, \\ \mathcal{F} & \textit{if } f^* \textit{ is affine on } \mathbf{Z}, \\ \mathcal{W} & \textit{if } f^* \textit{ is concave (but not strictly concave) on } \mathbf{Z}. \end{cases}$$

PROOF. See Appendix G. □

REMARK 3.7.3. (i) Notice that the tangent cone $T(\mathcal{F}, f^*)$ is not unique, but depends upon the degree of concavity of f^*.

(ii) If f^* is concave, but not strictly concave on $\mathbf{Z}$, there exists a nonempty set $\mathbf{Z}_0 \subset \mathbf{Z}$, on which $\det[\nabla^2 f^*]$ vanishes, [3] while on $\mathbf{Z} - \mathbf{Z}_0$ the Hessian matrix $\nabla^2 f^*$ is negative definite. This gives the rationale for defining $\mathcal{W}$. □

THEOREM 3.7.2. $\overline{lin\, T(\mathcal{F}, f^*)} = \mathcal{H}$.

PROOF. See Appendix G. □

As before, for $i = 1, \ldots, p$ the score function for β_i is

$$S_{\beta_i} = \begin{cases} \varepsilon[x_i + \delta(z_1, z_2)],\ \delta \in \mathcal{H} & \text{if } f^* \text{ is strictly concave,} \\ \varepsilon[x_i + \delta(z_1, z_2)],\ \delta \in \mathcal{F} & \text{if } f^* \text{ is affine,} \\ \varepsilon[x_i + \delta(z_1, z_2)],\ \delta \in \mathcal{W} & \text{if } f^* \text{ is concave, but not strictly concave.} \end{cases}$$

[3] Otherwise, by Theorem H.1 $\alpha'[\nabla^2 f^*]\alpha < 0$ for all $\alpha \in \mathbb{R}^2$, and f^* is strictly concave.

The most difficult one-dimensional sub-problem, i.e. the one with the least information, is then obtained by searching for $\delta \in \overline{lin\, T(\mathcal{F}, f^*)}$ such that the Fisher information is minimized. Following Theorem 3.7.2, the semiparametric information for β_{i0} is therefore,

$$i_{\beta_{i0}} = \inf_{\delta \in \mathcal{H}} \mathbb{E}\left[x_i + \delta(z_1, z_2)\right]^2.$$

This implies that the function δ_i^* solving the above optimization problem is

$$\delta_i^* = \operatorname{proj}[x_i|\mathcal{H}],$$

and from the following theorem, a projection on $\mathcal{H}$ is easily obtained.

THEOREM 3.7.3. *Let δ_i^* be the projection of x_i on $\mathcal{H}$. Then,*

$$\delta_i^*(u, v) = -\mathbb{E}(x_i|z_1 = u, z_2 = v).$$

PROOF. By imposing sufficient differentiability on the density functions, $\mathbb{E}(x_i|z_1 = u, z_2 = v) \in \mathcal{H}$. Hence, all that remains is to verify the orthogonality condition of the classical projection theorem. But this is straightforward. □

From the above theorem, the efficient score $\vec{S}$ for computing the semiparametric efficiency bounds of β_0 is

$$\vec{S} = \varepsilon \begin{pmatrix} x_1 - \mathbb{E}(x_1|z_1, z_2) \\ \vdots \\ x_p - \mathbb{E}(x_p|z_1, z_2) \end{pmatrix}.$$

The matrix $(\mathbb{E}\vec{S}\vec{S}')^{-1}$ then gives the required semiparametric efficiency bounds for β_0, when the true function f^* is concave.

Furthermore, let $K : [-1, 1] \times [-1, 1] \to \mathbb{R}$ be a kernel satisfying the multivariate version of Assumption 3.6.1. Then it is easy to see that

$$\eta_\beta(u,v) = \mathbb{E}\left(y_j | z_{1j} = u, z_{2j} = v\right) - \sum_{k=1}^{p} \beta_k \mathbb{E}\left(x_{kj} | z_{1j} = u, z_{2j} = v\right)$$

$$\hat{\eta}_\beta(u,v) = \frac{\sum_{j=1}^{n}(y_j - \sum_{k=1}^{p} x_{kj}\beta_k)\mathrm{K}(\frac{u-z_{1j}}{a_n}, \frac{v-z_{2j}}{a_n})}{\sum_{j=1}^{n} \mathrm{K}(\frac{u-z_{1j}}{a_n}, \frac{v-z_{2j}}{a_n})},$$

where η_β is a least favorable curve, and $\hat{\eta}_\beta$ is a consistent estimator of η_β. As in Section 3.6, it may also be verified that $\hat{\eta}_\beta$ and $\frac{d}{d\beta}\hat{\eta}_\beta$ satisfy all assumptions regarding rates of convergence etc. that an estimator of a least favorable curve has to satisfy.

The estimation problem yielding efficient estimates of β_0 when f^* is concave, is then

$$\hat{\beta}_n = \underset{\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \left[y_i - \sum_{k=1}^{p} x_{kj}\beta_k - \frac{\sum_{j=1}^{n}(y_j - \sum_{k=1}^{p} x_{kj}\beta_k)\mathrm{K}(\frac{z_{1i}-z_{1j}}{a_n}, \frac{z_{2i}-z_{2j}}{a_n})}{\sum_{j=1}^{n} \mathrm{K}(\frac{z_{1i}-z_{1j}}{a_n}, \frac{z_{2i}-z_{2j}}{a_n})} \right]^2 .$$

3.8. Conclusion

Recent trends clearly indicate the growing popularity of semiparametric techniques in econometrics. As econometricians incorporate restrictions of economic theory in these techniques, they will gain even wider acceptance among applied economists. This dissertation is a step in this direction, *viz.*, the integration of economic theory with econometric practice. Hopefully, it will be a stepping stone to the general theory of efficient semiparametric estimation under shape restrictions. Such a theory will be obtained when the class of shape restriction is extended to include all popular restrictions imposed by economic theory on unknown functions. However, in this chapter we have concentrated upon the two basic shape restrictions of homogeneity and concavity.

Under certain regularity conditions, we find that the efficiency bound for any regular estimator of β_0 is determined only by $\overline{lin\, T(\mathcal{F}, f^*)}$, the smallest closed linear space containing the tangent cone $T(\mathcal{F}, f^*)$. In fact, we show that efficiency bounds determined by $T(\mathcal{F}, f^*)$ cannot be attained by any regular $n^{1/2}$ - consistent estimator of β_0. Hence if two different shape restrictions on $\mathcal{F}$ produce the same $\overline{lin\, T(\mathcal{F}, f^*)}$, then the efficiency bound for any regular estimator of β_0 will be the same in both the cases.

In Section 3.6 we computed the efficiency bound for β_0, when the unknown function f^* was a homogeneous function of degree r. In order to do so, we had to obtain orthogonal projections on the space of homogeneous functions to find the least favorable direction. The computation of the efficiency bound also helped us construct an efficient estimator for β_0. This estimator was obtained by maximizing the profile (or the concentrated) likelihood, and is based on an approach of Severini and Wong (1992). The idea is extremely intuitive and is motivated by the fact that in parametric models maximum likelihood is efficient, leading to the possibility of it being efficient in semiparametric models. The construction of this estimator required a two step procedure. In the first step, the unknown function f^* was estimated while the finite dimensional parameter β_0 was kept fixed. In the second step, this estimate of f^* was used to concentrate the likelihood, which was then maximized over the finite dimensional parameter to produce an estimate of β_0. However, just any nonparametric estimate of f^* cannot be used to concentrate the likelihood in the second step. This is so, because approximation by an arbitrary estimator of f^* may introduce a bias in the asymptotic distribution for the estimator of β_0. But this bias disappears if we estimate f^* by a least favorable curve.

We showed that when f^* was a homogeneous function of degree r, the least favorable curve was also another homogeneous function of the same degree. Once this least favorable curve was used to estimate f^*, we demonstrated that maximizing the concentrated likelihood led to an efficient estimator of β_0. More importantly, since the least favorable direction in this case also turned out to be a homogeneous function, there are gains from homogeneity for estimating the finite dimensional parameters. Furthermore, as a by product of this research we have developed kernel estimators of homogeneous functions. Such an estimator is used in Chapter 4 to develop a test for homogeneity of functional form.

In Section 3.7 we computed the efficiency bounds for β_0 when f^* is a concave function, and also proposed an estimator that achieved these bounds. This problem is different from the previous one because unlike the space of homogeneous functions, the space of concave functions is not a linear-space. It is in fact a closed cone with strictly concave functions in its interior, and weakly concave functions on its boundary. This characterization is important, because it implies that the location of f^* inside this cone, i.e. whether f^* is strictly or weakly concave, will influence the efficiency bound for β_0. Moreover, this cone structure also creates some technical problems. For instance, the notion of the derivative as a best linear approximation makes sense only for linear-spaces or, in general for smooth manifolds (spaces which resemble linear spaces at any given point). Unfortunately, a cone is neither a linear space nor a smooth manifold. But since the space on which the projections are obtained is $\overline{lin\, T(\mathcal{F}, f^*)}$, these problems can be overcome without too much difficulty.

We find that when f^* is concave, the least favorable direction, obtained by projecting the scores of the parameter of interest onto $\overline{lin\, T(\mathcal{F}, f^*)}$, is just a twice continuously differentiable function. Hence if we restrict attention to the class of $n^{1/2}$-consistent regular estimators, computing efficiency bounds for β_0 when f^* is concave is equivalent to computing efficiency bounds for β_0 when f^* is just a C^2 - function. That is, we cannot do any better in estimating the finite dimensional parameters when we know that f^* is concave. However, if the least favorable direction is obtained by projecting the parametric scores onto $T(\mathcal{F}, f^*)$, then there is a possibility of gains from concavity, but at the expense of losing regularity.

CHAPTER 4

ESTIMATION AND TESTING OF HOMOGENEOUS FUNCTIONS

4.1. Introduction

Consider the regression model $y = f(\mathbf{x}) + \varepsilon$. In this chapter we obtain a test for the hypothesis that f is homogeneous of degree r, where r is assumed to be known to the economist. Using the approach developed in Chapter 3, we show how to construct the least squares estimator of f under homogeneity. Furthermore, we present the results of a small simulation experiment which was conducted to study the finite sample behavior of our test statistic.

Let us begin by analyzing the canonical regression

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \qquad i = 1, \dots, n.$$

The data $\{y_i, \mathbf{x}_i\}_{i=1}^{n}$ are assumed to be realizations of i.i.d. random variables $(Y, \mathbf{X})$ which take values in $\mathbb{R} \times \mathbb{R}^p$, where $p \geq 2$. Furthermore,

(i) The observations $\mathbf{x}$, come from a distribution with compact support $S_{\mathbf{X}} \subset \mathbb{R}^p$. W.l.o.g. let $S_{\mathbf{X}} = [0, 1]^p$.

(ii) The distribution of $\mathbf{x}$ has a Lebesgue density $p(\cdot)$, which is twice continuously differentiable.

(iii) The error terms ε are assumed to have full support with $\mathbb{E}\varepsilon = 0$, and variance

$$\sigma_\varepsilon^2(\mathbf{t}) = \text{Var}(\varepsilon|\mathbf{X} = \mathbf{t}) < \infty.$$

Moreover, ε is independent of $\mathbf{x}$.

(iv) For some $\alpha > 0$, $\mathbb{E}(|Y|^{2+\alpha}|\mathbf{X} = \mathbf{t})$ is a continuous function.

(v) The functional form of f is not known to the economist.

Thus far, the assumptions have been purely statistical in nature. Now suppose that the data are generated from an economic model that imposes some additional qualitative restrictions on the data generating process. We focus in particular on shape restrictions, and assume that the function f is a C^2-homogeneous function of degree $r \geq 0$.

ASSUMPTION 4.1.1. *The degree of homogeneity r, is known to the economist.*

As such models occur rather frequently in microeconomics, it is important to know if the shape restriction of homogeneity is a valid restriction. This is crucial, since most shape restrictions (including homogeneity) usually arise as a result of some optimization problem that economic agents are assumed to solve. Hence, if we reject the hypothesis that f is homogeneous, we also reject the hypothesis that agents are assumed to be solving an optimization problem that implied homogeneity of functional form. Moreover, since homogeneity is a particularly tractable shape restriction, say as compared to concavity or monotonicity, focusing on homogeneity may often lead to a simplification of econometric analysis. For instance, suppose that f is the cost function for a competitive firm producing a single output with $p-1$ inputs. This implies that f is linearly homogeneous, increasing and concave in factor prices. Therefore, rejection of homogeneity alone is sufficient to reject the hypothesis that the firm is minimizing costs.

In subsequent sections, we show how easy it is to estimate homogeneous functions. We construct an estimator for f, when f is homogeneous of degree r, and show that this estimator is optimal in the sense of being arbitrarily close to the least squares estimator. Moreover, we also develop a fully nonparametric test for the hypothesis that f is indeed homogeneous. Furthermore, we report the results of a simulation experiment which was performed to study the small sample properties of this test. The following notation is used throughout this chapter.

NOTATION 4.1.1. (i) We denote vectors in boldface. Thus, $\mathbf{x} = (x_1, x_2, \dots, x_p)$ and $\mathbf{x}_j = (x_{1,j}, x_{2,j}, \dots, x_{p,j})$ denote the values taken by random variables $\mathbf{X} = (X_1, X_2, \dots, X_p)$ and $\mathbf{X}_j = (X_{1,j}, X_{2,j}, \dots, X_{p,j})$.

(ii) $\mathcal{F}$ is the set of all C^2 functions on $S_{\mathbf{X}}$, which are also homogeneous of degree r. □

Using this notation, the null and alternative hypotheses are:

$$\begin{aligned} H_0 &: f = f^* \quad \text{for some } f^* \in \mathcal{F}, \\ H_1 &: f \neq f^* \quad \text{for all } f^* \in \mathcal{F}. \end{aligned}$$

4.2. Optimal Estimation of Homogeneous Functions

We start our analysis by approximating the least squares estimator of f under H_0. To do so, we determine the function $\pi^* \in \mathcal{F}$ that minimizes the L^2 distance between y and $\mathcal{F}$. The least squares estimator of f under the null, is then arbitrarily close to any consistent estimator of π^*. This technique of estimating homogeneous functions is an extension of the approach taken in Chapter 3, and leads to our first result.

THEOREM 4.2.1. *Let* $\pi^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}\,[y - f(\mathbf{x})]^2$. *Then,*

$$\pi^*(x_1, x_2, \dots, x_p) = \frac{A(x_1, x_2, \dots, x_p)}{B(x_1, x_2, \dots, x_p)}, \quad \textit{where,}$$

$$A(x_1, x_2, \ldots, x_p) = x_p^r\, \mathbb{E}\,(Y X_p^r | \frac{X_1}{X_p} = \frac{x_1}{x_p}, \frac{X_2}{X_p} = \frac{x_2}{x_p}, \ldots, \frac{X_{p-1}}{X_p} = \frac{x_{p-1}}{x_p})$$

$$B(x_1, x_2, \ldots, x_p) = \mathbb{E}\,(X_p^{2r} | \frac{X_1}{X_p} = \frac{x_1}{x_p}, \frac{X_2}{X_p} = \frac{x_2}{x_p}, \ldots, \frac{X_{p-1}}{X_p} = \frac{x_{p-1}}{x_p}).$$

PROOF. By imposing sufficient differentiability on the density functions, we can show that $\pi^*(x_1, \ldots, x_p) \in \mathrm{C}^2(S_{\mathbf{X}})$. But this implies that $\pi^* \in \mathcal{F}$, since it is already homogeneous of degree r by construction. Then using the classical projection theorem, all that remains is the verification of the orthogonality condition. To see that this holds, let g be any element of $\mathcal{F}$. Then, it is easy to see that

$$\begin{aligned} \pi^*(\mathbf{X})g(\mathbf{X}) &= \frac{X_p^r\, \mathbb{E}\,(Y X_p^r | \frac{X_1}{X_p}, \frac{X_2}{X_p}, \ldots, \frac{X_{p-1}}{X_p})}{\mathbb{E}\,(X_p^{2r} | \frac{X_1}{X_p}, \frac{X_2}{X_p}, \ldots, \frac{X_{p-1}}{X_p})} X_p^r g(\frac{X_1}{X_p}, \frac{X_2}{X_p}, \ldots, \frac{X_{p-1}}{X_p}, 1) \\ &= \frac{X_p^{2r}\, \mathbb{E}\,(Y g(X_1, X_2, \ldots, X_p) | \frac{X_1}{X_p}, \frac{X_2}{X_p}, \ldots, \frac{X_{p-1}}{X_p})}{\mathbb{E}\,(X_p^{2r} | \frac{X_1}{X_p}, \frac{X_2}{X_p}, \ldots, \frac{X_{p-1}}{X_p})}. \end{aligned}$$

Now, by using iterated expectations it can be verified that $\mathbb{E}\,[\pi^*(\mathbf{X})g(\mathbf{X})] = \mathbb{E}\,[y g(\mathbf{X})]$. Therefore, $\mathbb{E}\,[y - \pi^*(\mathbf{X})]g(\mathbf{X}) = 0$, and the orthogonality condition holds. □

Notice that under H_0, i.e. when $y = f^*(\mathbf{x}) + \varepsilon$,

$$\begin{aligned} A(x_1, \ldots, x_p) &= x_p^r \mathbb{E}(x_p^r f^*(\mathbf{x}) + x_p^r \varepsilon | \frac{X_1}{X_p} = \frac{x_1}{x_p}, \ldots, \frac{X_{p-1}}{X_p} = \frac{x_{p-1}}{x_p}) \\ &= x_p^r f^*(\frac{x_1}{x_p}, \ldots, \frac{x_{p-1}}{x_p}, 1) \mathbb{E}(x_p^{2r} | \frac{X_1}{X_p} = \frac{x_1}{x_p}, \ldots, \frac{X_{p-1}}{X_p} = \frac{x_{p-1}}{x_p}) \\ &= f^*(\mathbf{x}) B(x_1, \ldots, x_p), \end{aligned}$$

using the fact that ε is also independent of $(\frac{X_1}{X_p}, \frac{X_2}{X_p}, \ldots, \frac{X_{p-1}}{X_p})$. Thus, we have obtained the following result.

LEMMA 4.2.1. *Under* H_0, $\pi^* = f^*$.

Having determined π^*, it is now quite easy to construct a consistent estimator for it. Since π^* is a ratio of conditional expectations, simply replace each conditional expectation by its nonparametric analog *viz.* the kernel estimator. We then have the following result.

LEMMA 4.2.2. *Let,* $\hat{\pi}_n^*(x_1,\dots,x_p) = \dfrac{\hat{A}^*(x_1,\dots,x_p)}{\hat{B}^*(x_1,\dots,x_p)}$ *where,*

$$\hat{A}^*(x_1,\dots,x_p) = \frac{x_p^r}{na_n^{p-1}} \sum_{j=1}^{n} y_j x_{p,j}^r \mathrm{K}\left(\frac{x_1/x_p - x_{1,j}/x_{p,j}}{a_n},\dots,\frac{x_{p-1}/x_p - x_{p-1,j}/x_{p,j}}{a_n}\right)$$

$$\hat{B}^*(x_1,\dots,x_p) = \frac{1}{na_n^{p-1}} \sum_{j=1}^{n} x_{p,j}^{2r} \mathrm{K}\left(\frac{x_1/x_p - x_{1,j}/x_{p,j}}{a_n},\dots,\frac{x_{p-1}/x_p - x_{p-1,j}/x_{p,j}}{a_n}\right).$$

Then, $\hat{\pi}_n^*(x_1,\dots,x_p) \xrightarrow{p} \pi^*(x_1,\dots,x_p)$.

REMARK 4.2.1. (i) Note that even though $\mathbf{x} \in \mathbb{R}^p$, the argument of the kernel $\mathrm{K}(\cdot)$ is an element of $\mathbb{R}^{p-1}$. That is, a homogeneous function on $\mathbb{R}^p$ is estimated after reducing its dimension by one. This step has a profound consequence. As may be seen from the proof of Lemma 4.3.2, it is this reduction in the dimension of the nonparametric estimator that makes the distribution of our test statistic $O_p(1)$ under H_0. For assumptions on the kernel functions used in estimation, see the next section.

(ii) Since under the null hypothesis $\pi^* = f^*$, we can consistently estimate f^* by $\hat{\pi}_n^*$. Henceforth, we denote a consistent estimator of f^* by $\hat{f}_n^*$, where $\hat{f}_n^* = \hat{\pi}_n^*$.

(iii) Furthermore, we can modify the proof in Appendix F to show that with $a_n = O(\{\frac{\log n}{n}\}^{\frac{1}{3+p}})$,

$$\sup_{\mathbf{t}\in S_{\mathbf{x}}} |\hat{f}_n^*(\mathbf{t}) - f^*(\mathbf{t})| = o_p\left(\left\{\frac{\log n}{n}\right\}^{\frac{2}{3+p}}\right) \qquad \text{as } n \to \infty.$$

Note that since $(\frac{\log n}{n})^{\frac{2}{3+p}}$ is the optimal rate of of convergence under H_0, $\hat{f}_n^*$ is asymptotically optimal according to Stone (1982). □

4.3. Constructing the Test Statistic

In this section we construct a sample statistic for testing H_0. This statistic is based on an extension of the approach used in Severini and Staniswalis (1991). In their paper, Severini and Staniswalis developed a statistic for a parametric null hypothesis. However, in our case the null hypothesis is fully nonparametric. To help understand how our test works, we analyze in detail the behavior of the statistic at a fixed point. But let us first describe some notation which will be used subsequently.

NOTATION 4.3.1. Let $\hat{f}_n(\mathbf{t})$ denote the kernel estimator of f at a fixed point $\mathbf{t}$. That is,

$$\hat{f}_n(\mathbf{t}) = \frac{\hat{g}_n(\mathbf{t})}{\hat{p}_n(\mathbf{t})} \quad \text{where,}$$

$$\hat{g}_n(\mathbf{t}) = \frac{1}{nb_n^p} \sum_{j=1}^{n} y_j \mathrm{K}(\frac{\mathbf{t} - \mathbf{x}_j}{b_n})$$

$$\hat{p}_n(\mathbf{t}) = \frac{1}{nb_n^p} \sum_{j=1}^{n} \mathrm{K}(\frac{\mathbf{t} - \mathbf{x}_j}{b_n}). \quad \square$$

ASSUMPTION 4.3.1 (KERNEL). *The kernel function used above belongs to the class of product kernels. That is, for* $\mathbf{t} = (t_1, \ldots, t_p)$, *let* $\mathrm{K}(\mathbf{t}) = \Pi_{i=1}^{p} k(t_i)$ *be a real valued function on* $\mathbb{R}^p$. *Here, each* $k(\cdot)$ *is real valued and satisfies:*

(i) $k(t) = k(-t) \geq 0$,

(ii) $k(\cdot)$ *vanishes outside the interval* $[-1, 1]$,

(iii) $\int_{-1}^{1} k(s)\, ds = 1$,

(iv) $\int_{-1}^{1} sk(s)\, ds = 0$,

(v) $\int_{-1}^{1} s^2 k(s)\, ds < \infty$. □

REMARK 4.3.1. Remember that when we estimate f^*, the kernel $\mathrm{K}(\cdot)$ is defined on $\mathbb{R}^{p-1}$ and not on $\mathbb{R}^p$. This is because, as explained in Remark 4.2.1, we estimate homogeneous functions *after* reducing dimensionality by one. □

Assumption 4.3.2 (Window Width). *Let b_n be a sequence of positive numbers (the "window width") such that $b_n \to 0$ and $nb_n^p \to \infty$.* □

Remark 4.3.2. Similarly, when estimating f^* the window width a_n satisfies $a_n \to 0$, and $na_n^{p-1} \to \infty$. □

Notation 4.3.2. Let $\mathbb{E}_f\,(\cdot)$ denote the expectation of $(\cdot)$ when the unknown function in $(\cdot)$ is f. Since f is not known to the econometrician, so is $\mathbb{E}_f$. We therefore use $\mathbb{E}_{\hat{f}_n}$ to denote the estimator of $\mathbb{E}_f$. This is obtained by replacing f with $\hat{f}_n$ in the expression for $\mathbb{E}_f\,(\cdot)$. □

We are now ready to construct our test statistic. First notice that in our case, the density of the observations $\mathbf{x}_j$ does not change. This coupled with the fact that we are only interested in the functional form of f allows us to base our test statistic on the numerator of its nonparametric estimator, $\hat{f}_n$. This is very helpful since now we do not have to deal with ratios of random variables. So let $\{\mathbf{t}_1, \ldots, \mathbf{t}_m\}$ be m fixed points in $S_{\mathbf{X}}$, and define the test statistic $\Lambda_{m,n}$ as follows.

$$\begin{aligned} \Lambda_{m,n} &= \sum_{j=1}^{m} \frac{(nb_n^p)\,\hat{T}_n^2(\mathbf{t}_j)}{\hat{\sigma}^2(\mathbf{t}_j)} \\ &= \sum_{j=1}^{m} \hat{\Lambda}_{m,n}(\mathbf{t}_j) \end{aligned} \tag{4.3.1}$$

where,

$$\hat{T}_n(\mathbf{t}) = \hat{g}_n(\mathbf{t}) - \mathbb{E}_{\hat{f}_n^*}\,\hat{g}_n(\mathbf{t}),$$

and $\hat{\sigma}^2(\mathbf{t})$ is a consistent estimator of

$$\sigma^2(\mathbf{t}) = \mathbb{E}\,(y^2|\mathbf{t})p(\mathbf{t}) \int_{[-1,1]^p} \mathrm{K}^2(\mathbf{u})\,d\mathbf{u}.$$

REMARK 4.3.3. The notation $\Lambda_{m,n}$ indicates the dependence of the test statistic on the m points at which it is evaluated, and the sample size n. This is important, since m may be allowed to depend upon n. The case of m growing with n is treated in Section 4.6. However, unless mentioned otherwise m is assumed to be some fixed positive integer that does not depend upon n. □

To obtain the asymptotic distribution of $\Lambda_{m,n}$, let us first analyze the asymptotic behavior of $\hat{T}_n$ at a single fixed point $\mathbf{t} \in S_{\mathbf{X}}$. We begin by writing

$$(nb_n^p)^{1/2}\hat{T}_n(\mathbf{t}) = (nb_n^p)^{1/2}[\hat{g}_n(\mathbf{t}) - \mathbb{E}_f\,\hat{g}_n(\mathbf{t})] - B_n(\mathbf{t}), \tag{4.3.2}$$

where,

$$B_n(\mathbf{t}) = (nb_n^p)^{1/2}[\mathbb{E}_{\hat{f}_n^*}\,\hat{g}_n(\mathbf{t}) - \mathbb{E}_f\,\hat{g}_n(\mathbf{t})]$$

represents a bias term. We now have the following result.

LEMMA 4.3.1. *Let* $\sigma^2(\mathbf{t}_j) = \mathbb{E}(y^2|\mathbf{t}_j)p(\mathbf{t}_j)\int_{[-1,1]^p}\mathrm{K}^2(\mathbf{u})\,d\mathbf{u}$, *for* $j = 1,\ldots,m$. *Then,*

$$\begin{pmatrix}(nb_n^p)^{1/2}[\hat{g}_n(\mathbf{t}_1) - \mathbb{E}_f\,\hat{g}_n(\mathbf{t}_1)]\\ \vdots \\ (nb_n^p)^{1/2}[\hat{g}_n(\mathbf{t}_m) - \mathbb{E}_f\,\hat{g}_n(\mathbf{t}_m)]\end{pmatrix} \xrightarrow{d} \mathrm{N}\left(\begin{bmatrix}0\\ \vdots\\ 0\end{bmatrix}, \begin{bmatrix}\sigma^2(\mathbf{t}_1) & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sigma^2(\mathbf{t}_m)\end{bmatrix}\right)$$

PROOF. See Appendix I. □

Furthermore, let δ be any function in $\mathrm{C}^2(S_{\mathbf{X}})$. Then the next lemma is used to determine the asymptotic behavior of $B_n(\mathbf{t})$, both under the null hypothesis H_0, and under a sequence of local alternatives given by

$$\mathrm{H}_{1n} : f(\mathbf{t}) = f^*(\mathbf{t}) + \frac{\delta(\mathbf{t})}{\sqrt{nb_n^p}}.$$

Lemma 4.3.2. *Using the notation given above,*

$$B_n(\mathrm{t}) = \begin{cases} o(1) & \textit{under } \mathrm{H}_0, \\ -p(\mathrm{t})\delta(\mathrm{t}) + o(1) & \textit{under } \mathrm{H}_{1n}. \end{cases} \tag{4.3.3}$$

Proof. See Appendix J. □

Using these two lemmas we immediately obtain the following result.

Theorem 4.3.1. *With $\hat{T}_n(\mathrm{t})$ as defined in* (4.3.2),

$$\sqrt{nb_n^p}\hat{T}_n(\mathrm{t}) \xrightarrow{d} \begin{cases} \mathrm{N}(0,\sigma^2(\mathrm{t})) + o_p(1) & \textit{under } \mathrm{H}_0, \\ \mathrm{N}(p(\mathrm{t})\delta(\mathrm{t}),\sigma^2(\mathrm{t})) + o_p(1) & \textit{under } \mathrm{H}_{1n}. \end{cases}$$

Now let $\hat{\sigma}(\mathrm{t})$ be a consistent estimator of $\sigma(\mathrm{t})$. Then using Slutsky's Theorem and Theorem 4.3.1, we have that

$$\begin{aligned} \frac{\sqrt{nb_n^p}\,\hat{T}_n(\mathrm{t})}{\hat{\sigma}(\mathrm{t})} &= \frac{\sigma(\mathrm{t})}{\hat{\sigma}(\mathrm{t})}\,\frac{\sqrt{nb_n^p}\,\hat{T}_n(\mathrm{t})}{\sigma(\mathrm{t})} \\ &\xrightarrow{d} \begin{cases} \mathrm{N}(0,1) + o_p(1) & \text{under } \mathrm{H}_0, \\ \mathrm{N}(\frac{p(\mathrm{t})\delta(\mathrm{t})}{\sigma(\mathrm{t})},1) + o_p(1) & \text{under } \mathrm{H}_{1n}. \end{cases} \end{aligned}$$

But this implies that for a non-centrality parameter $\nu = \dfrac{p^2(\mathrm{t})\delta^2(\mathrm{t})}{2\sigma^2(\mathrm{t})}$,

$$\frac{(nb_n^p)\,\hat{T}_n^2(\mathrm{t})}{\hat{\sigma}^2(\mathrm{t})} \xrightarrow{d} \begin{cases} \chi_1^2 + o_p(1) & \text{under } \mathrm{H}_0, \\ \chi_1^2(\nu) + o_p(1) & \text{under } \mathrm{H}_{1n}. \end{cases}$$

Hence, we have shown that if $\nu_j = \dfrac{p^2(\mathrm{t}_j)\delta^2(\mathrm{t}_j)}{2\sigma^2(\mathrm{t}_j)}$, where $j = 1, \dots, m$, then

$$\begin{aligned} \hat{\Lambda}_{m,n}(\mathrm{t}_j) &= \frac{(nb_n^p)\,\hat{T}_n^2(\mathrm{t}_j)}{\hat{\sigma}^2(\mathrm{t}_j)} \\ &\xrightarrow{d} \begin{cases} \chi_1^2 + o_p(1) & \text{under } \mathrm{H}_0, \\ \chi_1^2(\nu_j) + o_p(1) & \text{under } \mathrm{H}_{1n}. \end{cases} \end{aligned}$$

But from Lemma 4.3.1 we have that $\hat{\Lambda}_{m,n}(\mathrm{t}_i)$ and $\hat{\Lambda}_{m,n}(\mathrm{t}_j)$ are also asymptotically independent for $i \neq j$. Therefore, we can finally obtain the asymptotic distribution of the sample statistic $\Lambda_{m,n}$ which is given by the following theorem.

THEOREM 4.3.2. *Let* $\nu = \sum_{j=1}^{m} \nu_j$, *where* $\nu_j = \dfrac{p^2(\mathbf{t}_j)\delta^2(\mathbf{t}_j)}{2\sigma^2(\mathbf{t}_j)}$. *Then,*

$$\begin{aligned}\Lambda_{m,n} &= \sum_{j=1}^{m} \hat{\Lambda}_{m,n}(\mathbf{t}_j) = \sum_{j=1}^{m} \frac{(nb_n^p)\,2\sigma^2(\mathbf{t}_j}{\hat{\sigma}^2(\mathbf{t}_j)} \\ &\xrightarrow{d} \begin{cases} \chi_m^2 + o_p(1) & \textit{under } \mathrm{H}_0, \\ \chi_m^2(\nu) + o_p(1) & \textit{under } \mathrm{H}_{1n}. \end{cases}\end{aligned}$$

4.4. Simulation and Computational Procedures

To study the finite sample properties of the proposed test, we performed a small simulation experiment. In order to simplify things, we restricted our attention to $p = 2$, i.e. covariates taking values in $\mathbb{R}^2$. The chosen data generating process was

$$\begin{aligned} y &= f(x_1, x_2) + \varepsilon, \quad \text{where,} \\ f(x_1, x_2) &= \begin{cases} x_1^2 + x_2^2 & \text{under } \mathrm{H}_0 \\ x_1^2 + x_2^2 + \dfrac{e^{x_1^2 + x_2^2}}{\sqrt{nb_n^2}} & \text{under } \mathrm{H}_{1n} \end{cases} \\ x_1, x_2 &\overset{\mathrm{d}}{=} \mathrm{UIID}(1,2) \\ \varepsilon &\overset{\mathrm{d}}{=} \mathrm{N}(0,1). \end{aligned}$$

REMARK 4.4.1. The function $(x_1, x_2) \mapsto x_1^2 + x_2^2$, which is homogeneous of degree 2, was chosen arbitrarily. Similarly, the choice of $\dfrac{e^{x_1^2+x_2^2}}{\sqrt{nb_n^2}}$ as the local perturbation was also arbitrary. □

The statistic $\Lambda_{m,n}$ was evaluated at $m = 25$ points, obtained by constructing a 5×5 grid in $\mathbb{R}^2$. The (x, y)-coordinates of the grid came from the sequence $\{1.1, 1.3, 1.5, 1.7, 1.9\}$. Note that computing $\hat{T}_n(\mathbf{t}) = \hat{g}_n(\mathbf{t}) - \mathbb{E}_{\hat{f}_n} \hat{g}_n(\mathbf{t})$ in its present form is not feasible, since we do not know how to select an optimal bandwidth for $\hat{g}_n$. However, calculating $\hat{T}_n(\mathbf{t})$ becomes simplified if we notice that:

$$\begin{aligned}\hat{T}_n(\mathbf{t}) &= \hat{g}_n(\mathbf{t}) - \mathbb{E}_{\hat{f}_n^*}\hat{g}_n(\mathbf{t})\\ &= \hat{f}_n(\mathbf{t})\hat{p}_n(\mathbf{t}) - \mathbb{E}_{\hat{f}_n^*}\hat{g}_n(\mathbf{t}).\end{aligned}$$

This is easy to compute, since we can now choose optimal bandwidths for $\hat{f}_n$ and $\hat{p}_n$ by cross validation.

REMARK 4.4.2. (i) The estimator of the density, denoted by $\hat{p}_n(\cdot)$, was computed using the Gaussian kernel. The use of a Gaussian kernel simplifies the form of the cross-validation function, which is used to obtain the optimal bandwidth for the density estimator.

(ii) The kernel used to compute the nonparametric estimators $\hat{f}_n$ and $\hat{f}_n^*$, was the Epanechnikov kernel

$$k(u) = \begin{cases} 0.75(1-u^2) & \text{if } -1 \le u \le 1\\ 0 & \text{otherwise,}\end{cases}$$

chosen for its second order optimality properties. Note that for the product Epanechnikov kernel,

$$\begin{aligned}\int_{[-1,1]^p} \mathrm{K}^2(\mathbf{u})\,d\mathbf{u} &= \Pi_{j=1}^p \int_{-1}^1 k^2(u_j)\,du_j\\ &= 0.6^p.\end{aligned}$$

(iii) The window width used in $\hat{f}_n^*$, was also chosen by cross-validation. □

To compute $\mathbb{E}_{\hat{f}_n^*}\hat{g}_n(\mathbf{t})$, notice that

$$
\begin{aligned}
\mathbb{E}_{f^*}\,\hat{g}_n(\mathbf{t}) &= \mathbb{E}_{f^*}\,\{\frac{1}{nb_n^p}\sum_{j=1}^{n} y_j \mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})\} \\
&= \frac{1}{b_n^p}\,\mathbb{E}_{f^*}\,\{y_j \mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})\} \\
&= \frac{1}{b_n^p}\,\mathbb{E}\,\{\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})\mathbb{E}_{f^*}\,(y_j|\mathbf{x}_j)\} \\
&= \frac{1}{b_n^p}\,\mathbb{E}\,\{\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})f^*(\mathbf{x}_j)\}.
\end{aligned}
$$

Hence $\mathbb{E}_{\hat{f}_n^*}\,\hat{g}_n(\mathbf{t}) = \frac{1}{b_n^p}\,\mathbb{E}\,\{\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})\hat{f}_n^*(\mathbf{x}_j)\}$, and was simulated using the algorithm in Table (4.7.1). Furthermore, $\hat{\sigma}^2(\mathbf{t})$ was obtained by utilizing the fact that

$$
\begin{aligned}
\sigma^2(\mathbf{t}) &= \mathbb{E}\,(y^2|\mathbf{t})p(\mathbf{t})\int_{[-1,1]^p} \mathrm{K}^2(\mathbf{u})\,d\mathbf{u} \\
&= \{f^2(\mathbf{t}) + \sigma_\epsilon^2(\mathbf{t})\}p(\mathbf{t})\int_{[-1,1]^p} \mathrm{K}^2(\mathbf{u})\,d\mathbf{u}.
\end{aligned}
$$

And therefore,

$$
\hat{\sigma}^2(\mathbf{t}) = \{\hat{f}_n^2(\mathbf{t}) + \hat{\sigma}_\epsilon^2(\mathbf{t})\}\hat{p}_n(\mathbf{t})\int_{[-1,1]^p} \mathrm{K}^2(\mathbf{u})\,d\mathbf{u}, \quad \text{where,}
$$

$$
\hat{\sigma}_\epsilon^2(\mathbf{t}) = \frac{\frac{1}{nb_n^p}\sum_{j=1}^n \mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})(y_j - \hat{f}_n(\mathbf{x}_j))^2}{\hat{p}_n(\mathbf{t})}.
$$

4.5. Results

The entire code for this simulation was written in GAUSS, and the results for 500 repetitions are presented in Table (4.7.2). As may be seen from this table, the test over rejects under H_0. However, it has excellent power characteristics. Some reasons that may help explain this poor performance under the null are:

(i) Inaccurate choice of m, the number of grid points at which the test statistic is evaluated.

(ii) Poor location of the grid points, which may destroy the independence of the individual terms in $\Lambda_{m,n}$.

The first problem may be resolved by allowing m to be a function of n, as in Severini and Staniswalis (1991). As far as the second problem is concerned, one could pick the m evaluation points and the bandwidth b_n such that, $mb_n^p < 2^{-p}$ for all n. For more on this, see Section 4.6. Of course, there is always a possibility that the test statistic developed in this chapter has poor finite sample performance. If this is indeed the case, then we need to look at modifications of this statistic which yield better finite sample approximations to its asymptotic distribution.

4.6. Letting m Grow with n[1]

In Remark J.1 it was pointed out that local alternatives for which $\delta(\mathbf{t}_i) = 0$, where $i = 1, \ldots, m$, are not detectable. Such local alternatives can be made uninteresting by letting m grow with the sample size n. In this section we develop the asymptotic theory for $\Lambda_{m,n}$, when m depends upon n. But first, some additional notation.

NOTATION 4.6.1. (i) Let m_n denote a sequence of increasing positive integers for $n = 1, 2, \ldots, \infty$.

(ii) For each n, let $\mathbf{t}_{n,1}, \mathbf{t}_{n,2}, \ldots, \mathbf{t}_{n,m_n}$ denote a lattice of fixed points in $S_{\mathbf{X}}$. As $n \to \infty$, these points get dense in $S_{\mathbf{X}}$.

(iii) Let $z_n(\mathbf{t}) = \frac{(nb_n^p)^{1/2}[\hat{g}_n(\mathbf{t}) - \mathbb{E}_f \hat{g}_n(\mathbf{t})]}{\sigma(\mathbf{t})}$. Furthermore, let $z_{n,j} = z_n(\mathbf{t}_{n,j})$ for $j = 1, \ldots, m_n$. Then $z_{n,j}$ is a triangular array of random variables with mean zero. □

The grid $\{\mathbf{t}_{n,1}, \ldots, \mathbf{t}_{n,m_n}\}$ in $S_{\mathbf{X}} = [0, 1]^p$ is created by choosing $m_n^{1/p}$ points in each dimension. These points are chosen such that the distance between adjacent points in each dimension is $m_n^{-1/p}$. It is then easy to see that $z_{n,j}$ is independent of $z_{n,j+1}$,

[1] I am grateful to Professor Severini for giving me access to some of his unpublished notes. Section 4.6 is based upon these notes.

if $m_n b_n^p < 2^{-p}$. That is, if $m_n \to \infty$ at a slow enough rate. Therefore, to ensure row independence of the triangular array $z_{n,j}$, we make the following assumption. This assumption will allow us to use a CLT for triangular arrays, later on in this section.

ASSUMPTION 4.6.1. *For each n, $m_n b_n^p < \frac{1}{2^p}$. In particular, this implies $m_n = o(b_n^{-p})$.* □

Apart from the assumptions made earlier, also assume the following.

ASSUMPTION 4.6.2. *As $n \to \infty$,*

(i) $m_n^{1/2}(nb_n^p)^{1/2} \sup_{\mathbf{t}} |\mathbb{E}_{\hat{f}_n^*} \hat{g}_n(\mathbf{t}) - \mathbb{E}_{f^*} \hat{g}_n(\mathbf{t})| \xrightarrow{p} 0$,

(ii) $m_n^{1/2} \sup_{\mathbf{t}} |\hat{\sigma}_n^{-2}(\mathbf{t}) - \sigma^{-2}(\mathbf{t})| \xrightarrow{p} 0$,

(iii) $m_n^{1/2} \sup_{\mathbf{t}} |\mathbb{E}\, z_n^2(\mathbf{t}) - 1| \xrightarrow{p} 0$,

(iv) $m_n^{1/2} \sup_{\mathbf{t}} |\mathbb{E}\, z_n^3(\mathbf{t})| \xrightarrow{p} 0$,

(v) $m_n^{1/2} \sup_{\mathbf{t}} |\mathbb{E}\, z_n^4(\mathbf{t}) - 3| \xrightarrow{p} 0$,

(vi) *For some $\xi > 4$,* $\sup_n \sup_{\mathbf{t}} \mathbb{E}\,|z_n(\mathbf{t})|^{\xi} < \infty$. □

ASSUMPTION 4.6.3. *The sequence of local alternatives is given by,*

$$\mathrm{H}_{2n} : f(\mathbf{t}) = f^*(\mathbf{t}) + \frac{\delta(\mathbf{t})}{m_n^{1/4}\sqrt{nb_n^p}}.$$

Here, $\delta(\cdot) \in \mathrm{C}^2(S_{\mathbf{X}})$ such that $\sup_{\mathbf{t}} |\delta(\mathbf{t})| > 0$. □

To make explicit the dependence of m upon n, the test statistic in (4.3.1) is henceforth denoted by $\Lambda_{m_n,n}$. That is,

$$\Lambda_{m_n,n} = \sum_{j=1}^{m_n} \frac{(nb_n^p)\,\hat{T}_n^2(\mathbf{t}_{n,j})}{\hat{\sigma}^2(\mathbf{t}_{n,j})},$$

where $\hat{T}_n$ is defined as before. The asymptotic distribution of $\Lambda_{m_n,n}$ under H_0 and H_{2n} is then given by Theorem 4.6.1 and Theorem 4.6.2, respectively.

THEOREM 4.6.1. *Under* H_0,

$$\frac{\Lambda_{m_n,n} - m_n}{\sqrt{2m_n}} \xrightarrow{d} \mathrm{N}(0,1) \qquad \text{as } n \to \infty.$$

PROOF. After some tedious algebra, $\Lambda_{m_n,n} = \Lambda^{(1)}_{m_n,n} + o_p(m_n^{1/2})$ where,

$$\Lambda^{(1)}_{m_n,n} = \frac{nb_n^p \sum_{j=1}^{m_n} [\hat{g}_n(\mathbf{t}_{n,j}) - \mathbb{E}_{f^*}\, \hat{g}_n(\mathbf{t}_{n,j})]^2}{\hat{\sigma}_n^2(\mathbf{t}_{n,j})}.$$

Some more algebra yields

$$\Lambda^{(1)}_{m_n,n} = \frac{nb_n^p \sum_{j=1}^{m_n} [\hat{g}_n(\mathbf{t}_{n,j}) - \mathbb{E}_{f^*}\, \hat{g}_n(\mathbf{t}_{n,j})]^2}{\sigma^2(\mathbf{t}_{n,j})} + o_p(m^{1/2}).$$

Therefore, $\Lambda_{m_n,n} - m_n = \sum_{j=1}^{m_n} (z_{n,j}^2 - 1) + o_p(m_n^{1/2})$, and,

$$\frac{\Lambda_{m_n,n} - m_n}{\sqrt{2m_n}} = \frac{\sum_{j=1}^{m_n} (z_{n,j}^2 - 1)}{\sqrt{2m_n}} + o_p(1).$$

Hence, it suffices to show that

$$\frac{S_n}{\sqrt{2m_n}} = \frac{\sum_{j=1}^{m_n} (z_{n,j}^2 - 1)}{\sqrt{2m_n}} \xrightarrow{d} \mathrm{N}(0,1).$$

From Assumption 4.6.1, S_n is the sum of a row independent triangular array. Now a CLT for triangular arrays (Durrett 1991) yields, $\frac{S_n - \mathbb{E}\, S_n}{\sqrt{\operatorname{Var} S_n}} \xrightarrow{d} \mathrm{N}(0,1)$. After some computations it may also be seen that

$$\mathbb{E}\, S_n = o(m_n^{1/2})$$

$$\operatorname{Var} S_n = 2m_n + o(m^{1/2}).$$

Therefore, $\frac{S_n - \mathbb{E}\, S_n}{\sqrt{\operatorname{Var} S_n}} = \frac{S_n}{\sqrt{2m_n}} + o(1)$, which implies that

$$\begin{aligned} \frac{S_n}{\sqrt{2m_n}} &= \frac{S_n - \mathbb{E}\, S_n}{\sqrt{\operatorname{Var} S_n}} + o(1) \\ &\xrightarrow{d} \mathrm{N}(0,1) + o(1). \end{aligned}$$

Hence, by utilizing Slutsky's Theorem,

$$\frac{\Lambda_{m_n,n} - m_n}{\sqrt{2m_n}} = \frac{S_n}{\sqrt{2m_n}} + o_p(1)$$
$$\xrightarrow{d} \mathrm{N}(0,1).$$

□

THEOREM 4.6.2. *Under the sequence of local alternatives* H_{2n},

$$\frac{\Lambda_{m_n,n} - m_n}{\sqrt{2m_n}} - \Delta_n \xrightarrow{d} \mathrm{N}(0,1), \quad \textit{as } n \to \infty, \textit{ where,}$$

$$\Delta_n = \frac{1}{\sqrt{2m_n}} \sum_{j=1}^{m_n} \frac{\delta^2(\mathbf{t}_{n,j}) p^2(\mathbf{t}_{n,j})}{\sigma^2(\mathbf{t}_{n,j})}.$$

PROOF. Follows from the proof of Theorem 4.6.1 by noting that under H_{2n},

$$\Lambda_{m_n,n} = \Lambda^{(1)}_{m_n,n} + \frac{1}{m_n^{1/2}} \sum_{j=1}^{m_n} \frac{\delta^2(\mathbf{t}_{n,j}) p^2(\mathbf{t}_{n,j})}{\sigma^2(\mathbf{t}_{n,j})} + o_p(m_n^{1/2}).$$

□

Therefore, as $n \to \infty$ and $\{\mathbf{t}_{n,1}, \mathbf{t}_{n,2}, \ldots, \mathbf{t}_{n,m_n}\}$ become dense in $S_{\mathbf{X}}$, any test based on $\Lambda_{m_n,n}$ has positive power under H_{2n}.

4.7. Tables for Chapter 4

TABLE (4.7.1). Algorithm for Simulating $\mathbb{E}_{\hat{f}_n^*}\, \hat{g}_n(\mathbf{t})$

(i)	Generate sample: $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.
(ii)	For each $\mathbf{x}_j$, compute: $\mathrm{K}(\frac{\mathbf{t} - \mathbf{x}_j}{b_n})\hat{f}_n^*(\mathbf{x}_j)$.
(iii)	Take average: $\frac{1}{n}\sum_{j=1}^{n} \mathrm{K}(\frac{\mathbf{t} - \mathbf{x}_j}{b_n})\hat{f}_n^*(\mathbf{x}_j)$, and divide by b_n^p.

TABLE (4.7.2). Simulation Results ($m = 25$, Repetitions = 500)

Sample Size	# Rejections (H_0)	Size of	# Rejections (H_{1n})	Power of
(n)	(5% Level)	Test	(5% Level)	Test
50	106	0.21	500	1.00
100	94	0.19	500	1.00
250	60	0.12	500	1.00

APPENDIX A

TANGENT CONES

In this appendix we collect some results about tangent cones. These results are available in standard mathematical literature but seem to be scattered all over the place. We begin with a definition from Krabs (1979, Page 154).

DEFINITION A.1 (TANGENT VECTOR & TANGENT CONE). Let E be a normed vector space, A a non-empty subset of E, and x_0 any point of A. A vector $h \in E$ is called a tangent vector to A at x_0 if there is a sequence x_n of elements of A and a sequence λ_n of positive real numbers with $\lim_{n \to \infty} x_n = x_0$ and $\lim_{n \to \infty} \lambda_n(x_n - x_0) = h$. Furthermore, let $T(A, x_0)$ be the set of all tangent vectors to A at x_0. Then $T(A, x_0)$ is called the tangent cone to A at x_0. □

REMARK A.1. (i) Since $T(A, x_0)$ certainly contains the null vector of E, it is not empty.

(ii) In the above definition, x_0 is necessarily a point of closure of A. Moreover, in general $T(A, x_0)$ is not a convex set.

(iii) Notice that if $A \subset B$ and $x_0 \in A \cap B$, then $T(A, x_0) \subset T(B, x_0)$.

(iv) We can also show that if $x_0 \in A \cap \text{int(B)}$, then $T(A \cap B, x_0) = T(A, x_0)$. □

We now look at some properties of $T(A, x_0)$.

Lemma A.1. *$T(A, x_0)$ is a cone.*

Proof. Let $h \in T(A, x_0)$. Therefore, there exists a sequence of real numbers $\lambda_n > 0$ and a sequence of elements $x_n \in A$ with $x_n \to x_0$ such that $h = \lim_{n\to\infty} \lambda_n(x_n - x_0)$.

To show that $T(A, x_0)$ is a cone we have to show that $\alpha h \in T(A, x_0)$ for all $\alpha > 0$. Now, notice that,

$$\begin{aligned} \alpha h &= \alpha \lim_{n\to\infty} \lambda_n(x_n - x_0) \\ &= \lim_{n\to\infty} \alpha\lambda_n(x_n - x_0) \\ &= \lim_{n\to\infty} \mu_n(x_n - x_0), \end{aligned}$$

where $\mu_n = \alpha\lambda_n$. This shows that there exists a sequence of real numbers $\mu_n > 0$, and a sequence of elements $x_n \in A$ with $x_n \to x_0$ such that $\alpha h = \lim_{n\to\infty} \mu_n(x_n - x_0)$. i.e. αh is also a tangent vector at x_0, which implies that $\alpha h \in T(A, x_0)$. Therefore, $T(A, x_0)$ is a cone. □

Lemma A.2. *$T(A, x_0)$ is closed.*

Proof. See Krabs (1979, Page 154). □

The next lemma gives a sufficient condition under which $T(A, x_0)$ is a convex set.

Lemma A.3. *Let A be a non-empty convex subset of a vector space E. Then, $T(A, x_0)$ contains $A - x_0$ and is convex.*

Proof. We first show that $T(A, x_0)$ contains $A - x_0$ if A is convex. So let $h \in A$. Now define the sequence $h_n = x_0 + \frac{1}{n}(h - x_0)$, i.e. $h_n = \frac{1}{n}h + (1 - \frac{1}{n})x_0$. Clearly, $h_n \in A$ since A is convex. Also, $h_n \to x_0$ and $n(h_n - x_0) \to h - x_0$. Therefore, $h - x_0 \in T(A, x_0)$, and since h was an arbitrary element of A, this implies that $A - x_0 \subset T(A, x_0)$.

We now show that $T(A, x_0)$ is convex. Let $h_1, h_2 \in T(A, x_0)$. Then there exist sequences $x_n^1, x_n^2 \in A$ with $x_n^1 \to x_0, x_n^2 \to x_0$, and sequences of positive real numbers μ_n^1, μ_n^2 such that $h_1 = \lim_{n\to\infty} \mu_n^1(x_n^1 - x_0)$ and $h_2 = \lim_{n\to\infty} \mu_n^2(x_n^2 - x_0)$. Now, let $0 \leq \lambda \leq 1$ and define $h = \lambda h_1 + (1-\lambda)h_2$. Then, $h = \lim_{n\to\infty} \delta_n(z_n - x_0)$, where,

$$\begin{aligned} \delta_n &= \lambda\mu_n^1 + (1-\lambda)\mu_n^2, \quad \text{and,} \\ z_n &= \frac{\lambda\mu_n^1}{\lambda\mu_n^1 + (1-\lambda)\mu_n^2} x_n^1 + \frac{(1-\lambda)\mu_n^2}{\lambda\mu_n^1 + (1-\lambda)\mu_n^2} x_n^2. \end{aligned}$$

Now δ_n is a sequence of positive real numbers, and $z_n \in A$ since A is convex. So if we can show that $z_n \to x_0$ we would be done, since then h would be an element of $T(A, x_0)$. We show this as follows. Notice that,

$$\begin{aligned} \|z_n - x_0\| &= \left\| \frac{\lambda\mu_n^1}{\lambda\mu_n^1 + (1-\lambda)\mu_n^2}(x_n^1 - x_0) + \frac{(1-\lambda)\mu_n^2}{\lambda\mu_n^1 + (1-\lambda)\mu_n^2}(x_n^2 - x_0) \right\| \\ &\leq \frac{\lambda\mu_n^1}{\lambda\mu_n^1 + (1-\lambda)\mu_n^2}\|x_n^1 - x_0\| + \frac{(1-\lambda)\mu_n^2}{\lambda\mu_n^1 + (1-\lambda)\mu_n^2}\|(x_n^2 - x_0)\| \\ &= \frac{\lambda}{\lambda + (1-\lambda)\frac{\mu_n^2}{\mu_n^1}}\|x_n^1 - x_0\| + \frac{(1-\lambda)}{(1-\lambda) + \lambda\frac{\mu_n^1}{\mu_n^2}}\|(x_n^2 - x_0)\|. \end{aligned}$$

And since both the coefficients are bounded by 1, we have

$$\begin{aligned} \|z_n - x_0\| &\leq \|x_n^1 - x_0\| + \|x_n^2 - x_0\| \\ &\to 0, \end{aligned}$$

since $\|x_n^1 - x_0\| \to 0$, and $\|x_n^2 - x_0\| \to 0$. Hence we are done. □

Using the properties given above, we get the following characterization of a tangent cone.

THEOREM A.1. *Let A be a non-empty convex subset of a vector space E. Then, $T(A, x_0)$ is the smallest closed cone containing $A - x_0$.*

PROOF. By the previous theorems, $T(A, x_0)$ is a closed cone containing $A - x_0$. It only remains to show that it is the smallest closed cone containing $A - x_0$. So let $C(A - x_0)$ be any closed cone containing $A - x_0$, and suppose that $h \in T(A, x_0)$. Then $h = \lim_{n \to \infty} \lambda_n(x_n - x_0)$, where λ_n is a sequence of positive reals and x_n is a sequence of elements in A approaching x_0.

So define $h_n = \lambda_n(x_n - x_0)$. Clearly, $x_n - x_0 \in A - x_0$. But since $A - x_0 \subset C(A - x_0)$ we have that $x_n - x_0 \in C(A - x_0)$. Now, the fact that $C(\cdot)$ is a cone implies that $h_n = \lambda_n(x_n - x_0) \in C(A - x_0)$. i.e. h_n is a convergent sequence in $C(A - x_0)$. But since $C(A - x_0)$ is closed, the limit $h \in C(A - x_0)$. This implies that $T(A, x_0) \subset C(A - x_0)$. Note that $T(A, x_0)$ is also convex since A is a convex set. □

Using this theorem, we get the following important result about $T(A, x_0)$, when A is itself a cone. This is a result of Aubin and Frankowska (1990, Lemma 4.2.5, Page 143).

THEOREM A.2. *Let A be a non-empty convex cone in a vector space E, and $x_0 \in A$. Then, $T(A, x_0) = \overline{A - \mathbb{R}_{++}x_0}$.*

PROOF. $\underline{\Longrightarrow}$ Let $h \in T(A, x_0)$. Therefore, there exists a sequence of real numbers $\lambda_n > 0$ and a sequence of elements $x_n \in A$ with $x_n \to x_0$, such that $h = \lim_{n \to \infty} \lambda_n(x_n - x_0)$. But $\lambda_n x_n \in A$ since A is a cone, and clearly $\lambda_n x_0 \in \mathbb{R}_{++}x_0$. This implies that $\lambda_n x_n - \lambda_n x_0 \in A - \mathbb{R}_{++}x_0$, which shows that $h = \lim_{n \to \infty} \lambda_n(x_n - x_0) \in \overline{A - \mathbb{R}_{++}x_0}$.

$\underline{\Longleftarrow}$ Let x be any arbitrary element of A, and $\lambda > 0$ any element in $\mathbb{R}_{++}$. If we could show that $x - \lambda x_0 \in T(A, x_0)$, we would be done since this would imply that $A - \mathbb{R}_{++}x_0 \subset$

$T(A, x_0)$. Then taking the closure on both sides, and keeping in mind that $T(A, x_0)$ is closed, we would have $\overline{A - \mathbb{R}_{++}x_0} \subset T(A, x_0)$. So we show that $x - \lambda x_0 \in T(A, x_0)$.

Since $\lambda \in \mathbb{R}_{++}$, choose a $t > 0$ such that $\lambda t < 1$. Then since A is a convex cone, we have that $(1 - \lambda t)x_0 + tx \in A$, i.e. $x_0 + t(x - \lambda x_0) \in A$. This implies that $t(x - \lambda x_0) \in A - x_0$. But we know that $T(A, x_0)$ contains $A - x_0$. Hence, we have that $t(x - \lambda x_0) \in T(A, x_0)$. But since $t > 0$ and $T(A, x_0)$ is a cone, this implies that $x - \lambda x_0 \in T(A, x_0)$. □

COROLLARY A.1. *Let A be a closed linear subspace of a vector space E, and let $x_0 \in A$. Then $T(A, x_0) = A$.*

PROOF. Follows from the previous theorem. □

APPENDIX B

PROOFS OF RESULTS IN SECTION 2.7

REMARK B.1. The following proofs are from Severini (1987), with slight modifications. □

PROOF. [**Theorem 2.7.1**] Clearly $\boldsymbol{\beta}_t = (\beta_{10}, \beta_{20}, \ldots, \beta_{i0} + t, \beta_{(i+1)0}, \ldots, \beta_p)$ is an admissible curve in **B**. Then since $\eta_{\boldsymbol{\beta}}$ is a least favorable surface, $\eta_{\boldsymbol{\beta}_t}$ must be a least favorable curve for estimating $\boldsymbol{\beta}_t$. That is, it must minimize $\mathbb{E}\ [\frac{d}{dt}\ell(\boldsymbol{\beta}_t, \eta_{\boldsymbol{\beta}_t})|_{t=0}]^2$. Now,

$$\begin{aligned}
\frac{d}{dt}\ell(\boldsymbol{\beta}_t, \eta_{\boldsymbol{\beta}_t}) &= \sum_{i=1}^{p} \frac{\partial \ell(\boldsymbol{\beta}_t, \eta_{\boldsymbol{\beta}_t})}{\partial \beta_i} \frac{d\beta_i(t)}{dt} + \frac{\partial \ell(\boldsymbol{\beta}_t, \eta_{\boldsymbol{\beta}_t})}{\partial \eta} \sum_{i=1}^{p} \left(\frac{d}{d\beta_i(t)} \eta_{\boldsymbol{\beta}_t} \right) \frac{d\beta_i(t)}{dt} \\
&= \frac{\partial \ell(\boldsymbol{\beta}_t, \eta_{\boldsymbol{\beta}_t})}{\partial \beta_i} + \frac{\partial \ell(\boldsymbol{\beta}_t, \eta_{\boldsymbol{\beta}_t})}{\partial \eta} \left(\frac{d}{d\beta_i} \eta_{\boldsymbol{\beta}_t} \right), \quad \text{implying,} \\
\frac{d}{dt}\ell(\boldsymbol{\beta}_t, \eta_{\boldsymbol{\beta}_t})\Big|_{t=0} &= \frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \beta_i} + \frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \eta} \left(\frac{d}{d\beta_i} \eta_{\boldsymbol{\beta}_0} \right).
\end{aligned}$$

Therefore, minimizing $\mathbb{E}\,[\frac{d}{dt}\ell(\boldsymbol{\beta}_t, \eta_{\boldsymbol{\beta}_t})|_{t=0}]^2$ is equivalent to minimizing

$$\mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \beta_i} + \frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \eta} \left(\frac{d}{d\beta_i} \eta_{\boldsymbol{\beta}_0} \right) \right]^2 = \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \beta_i} + \frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \eta} \delta_i \right]^2,$$

where $\delta_i \in \overline{lin\, T(\mathcal{F}, f^*)}$ for $i = 1, \ldots, p$. Hence the minimizer δ_i^* satisfies

$$\mathbb{E}\left(\frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \beta_i} + \frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \eta} (\delta_i^*) \right) \frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \eta} (\delta^*) = 0$$

for all $\delta \in \overline{lin\, T(\mathcal{F}, f^*)}$, and $i = 1, \ldots, p$. □

Proof. **[Theorem 2.7.2]** $\Longrightarrow$ So suppose that the result is true. Then since it holds for all $(\delta_1,\dots,\delta_p) \in \overline{lin\,T(\mathcal{F},f^*)} \times \dots \times \overline{lin\,T(\mathcal{F},f^*)}$, it also holds for $(\delta_1,0,\dots,0) \in \overline{lin\,T(\mathcal{F},f^*)} \times \dots \times \overline{lin\,T(\mathcal{F},f^*)}$, because 0 is always an element of the tangent cone. Therefore,

$$\mathbb{E}\left(\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\beta_1} + \frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}\left(\frac{d}{d\beta_1}\eta_{\beta_0}\right)\right)\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\delta_1) = 0,$$

But this implies that δ_1 is the least favorable direction. The same holds for $\delta_2,\dots,\delta_p$, and we get that $\delta^* = (\delta_1,\dots,\delta_p)$ is the least favorable direction. Therefore, η_β is a least favorable curve since $\boldsymbol{\delta}^* = \frac{d}{d\beta}\eta_\beta|_{\beta=\beta_0}$ is the least favorable direction.

$\Longleftarrow$ Now suppose that η_β is a least favorable curve, and let the least favorable direction be $\boldsymbol{\delta} = (\frac{d}{d\beta_1}\eta_\beta|_{\beta=\beta_0},\dots,\frac{d}{d\beta_p}\eta_\beta|_{\beta=\beta_0})$. Then from Theorem 2.7.1, $\boldsymbol{\delta}$ satisfies

$$\mathbb{E}\left[\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\beta_i} + \frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta}(\delta_i^*)\right]\frac{\partial\ell(\beta_0,\eta_{\beta_0})}{\partial\eta_{\beta_0}}(\delta_i) = 0$$

for all $\delta_i \in \overline{lin\,T(\mathcal{F},f^*)}$, and $i = 1,\dots,p$. Summation over i then yields the required result. □

Proof. **[Theorem 2.7.3]** $\Longrightarrow$ Let I_{β_0} be given by

$$\mathbb{E}\left(\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right)\right)\left(\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right)\right)'.$$

Also assume that $\alpha'(I_{\beta_0} - I)\alpha \leq 0$ for all $\alpha \in \mathbb{R}^p$. We show that λ_β is a least favorable surface. Now since the given condition holds for all $\alpha \in \mathbb{R}^p$, choose $\alpha = \mathbf{e}_j$ the j^{th} unit vector in $\mathbb{R}^p$. Then $\mathbf{e}_j'(I_{\beta_0} - I)\mathbf{e}_j \leq 0$ becomes,

$$\mathbb{E}\left[\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta_j} + \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta_j}\lambda_{\beta_0}\right)\right]^2 \leq \mathbb{E}\left[\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta_j} + \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta_j}\eta_{\beta_0}\right)\right]^2$$

for $j = 1,\dots,p$. This[1] implies that $\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}(\frac{d}{d\beta_j}\lambda_{\beta_0})$ is the projection of $\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta_j}$ onto

[1] Since both λ_β and η_β are admissible curves in $\overline{lin\mathcal{F}}$ through f^*, the tangent vectors $\frac{d}{d\beta_j}\lambda_{\beta_0}$ and $\frac{d}{d\beta_j}\eta_{\beta_0}$, are both elements of $\overline{lin\,T(\mathcal{F},f^*)}$.

the space

$$\frac{\partial \ell(\boldsymbol{\beta}_0, \lambda_{\boldsymbol{\beta}_0})}{\partial \lambda} \left(\overline{lin\, T(\mathcal{F}, f^*)} \right).$$

Therefore, $\frac{d}{d\beta_j}\lambda_{\boldsymbol{\beta}_0}$ is the least favorable direction for $j = 1, \ldots, p$, implying that $\lambda_{\boldsymbol{\beta}}$ is a least favorable curve.

$\Longleftarrow$ Now suppose that $\eta_{\boldsymbol{\beta}}$ is a least favorable surface in $\overline{lin\mathcal{F}}$, and define

$$I_{\boldsymbol{\beta}_0} = \mathbb{E}\left(\frac{d\ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{d\boldsymbol{\beta}}\right)\left(\frac{d\ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{d\boldsymbol{\beta}}\right)'.$$

Then we show that $\boldsymbol{\alpha}'(\mathbb{I}_{\boldsymbol{\beta}_0} - I)\boldsymbol{\alpha} \leq 0$ for all $\boldsymbol{\alpha} \in \mathbb{R}^p$, where I is the Fisher information matrix corresponding to another p dimensional parameterization of η.

So let $\boldsymbol{\beta}_t = \boldsymbol{\beta}_0 + t\boldsymbol{\alpha}$. Clearly, $\boldsymbol{\beta}_t$ is an admissible curve in $\mathbf{B}$ through $\boldsymbol{\beta}_0$, with tangent vector $\boldsymbol{\alpha}$. Then since $\eta_{\boldsymbol{\beta}}$ is a least favorable surface, $\eta_{\boldsymbol{\beta}_t}$ must be a least favorable curve for $\boldsymbol{\beta}_t$, i.e. $\eta_{\boldsymbol{\beta}_t}$ minimizes $\mathbb{E}\left[\frac{d\ell(\boldsymbol{\beta}_t, \eta_{\boldsymbol{\beta}_t})}{dt}|_{t=0}\right]^2$. But by the chain rule,

$$\begin{aligned}
\frac{d\ell(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}_t})}{dt}\Big|_{t=0} &= \left[\frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \boldsymbol{\beta}} + \frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \eta}\left(\frac{d}{d\boldsymbol{\beta}}\eta_{\boldsymbol{\beta}_0}\right)\right]' \left[\frac{d\boldsymbol{\beta}_t}{dt}\Big|_{t=0}\right] \\
&= S'_{\boldsymbol{\beta}_0}\boldsymbol{\alpha}, \quad \text{which implies that,} \\
\mathbb{E}\left[\frac{d\ell(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}_t})}{dt}\Big|_{t=0}\right]^2 &= \mathbb{E}\left(S'_{\boldsymbol{\beta}_0}\boldsymbol{\alpha}\right)'\left(S'_{\boldsymbol{\beta}_0}\boldsymbol{\alpha}\right) \\
&= \boldsymbol{\alpha}'\mathbb{E}\, S_{\boldsymbol{\beta}_0} S'_{\boldsymbol{\beta}_0}\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}' I_{\boldsymbol{\beta}_0}\boldsymbol{\alpha},
\end{aligned}$$

and we get that $\eta_{\boldsymbol{\beta}_t}$ minimizes $\boldsymbol{\alpha}' I_{\boldsymbol{\beta}_0}\boldsymbol{\alpha}$. Now if $\lambda_{\boldsymbol{\beta}}$ is any other admissible curve in $\overline{lin\mathcal{F}}$, we have

$$\mathbb{E}\left[\frac{d\ell(\boldsymbol{\beta}_t, \lambda_{\boldsymbol{\beta}_t})}{dt}|_{t=0}\right]^2 = \boldsymbol{\alpha}' I \boldsymbol{\alpha},$$

where I is the information corresponding to $\lambda_{\boldsymbol{\beta}_t}$. But since $\eta_{\boldsymbol{\beta}_t}$ was a least favorable curve, it minimized $\boldsymbol{\alpha}' I_{\boldsymbol{\beta}_0}\boldsymbol{\alpha}$, i.e. $\boldsymbol{\alpha}' I_{\boldsymbol{\beta}_0}\boldsymbol{\alpha} \leq \boldsymbol{\alpha}' I \boldsymbol{\alpha}$, for all $\boldsymbol{\alpha} \in \mathbb{R}^p$. □

APPENDIX C

PROOF OF THEOREM 3.4.1

The proof of Theorem 3.4.1 requires the following definitions.

DEFINITION C.1 (LAN CONDITION). Let

$$L_n(\beta_n, \gamma_{\beta_n}) = \sum_{i=1}^{n} \ell(\beta, \gamma_{\beta_n}; \mathbf{x}_i, y_i, \mathbf{z}_i),$$

where γ_β is any curve in $\mathrm{C}^2(\mathbf{Z})$ such that $\gamma_\beta|_{\beta=\beta_0} = f^*$. Also let $\mathfrak{L}_n = L_n(\beta_n, \gamma_{\beta_n}) - L_n(\beta_0, f^*)$. Then for sufficiently large n,

$$\mathfrak{L}_n = n^{-1/2}\delta' \frac{d}{d\beta} L_n(\beta_0, \gamma_{\beta_0}) - \frac{1}{2}\delta' I_L \delta + o_p(1),$$

where $I_L = \mathbb{E}[\frac{d}{d\beta}\ell(\beta_0, \gamma_{\beta_0})][\frac{d}{d\beta}\ell(\beta_0, \gamma_{\beta_0})]'$. □

The following lemma gives sufficient conditions under which the LAN condition holds.

LEMMA C.1. *With γ_β as defined above and $i, j = 1, \ldots, p$, assume that for all $\beta \in \mathbb{R}^p$ the loglikelihood $\beta \mapsto \ell(\beta, \gamma_\beta; \mathbf{x}, y, \mathbf{z})$ satisfies,*

(i) $\mathbb{E}\,\frac{d}{d\beta_i}\ell(\beta, \gamma_\beta; \mathbf{x}, y, \mathbf{z}) = 0.$

(ii) $\mathbb{E}\,[\frac{d^2}{d\beta_j d\beta_i}\ell(\beta, \gamma_\beta; \mathbf{x}, y, \mathbf{z})] + \mathbb{E}\,[\frac{d}{d\beta_i}\ell(\beta, \gamma_\beta; \mathbf{x}, y, \mathbf{z})\frac{d}{d\beta_j}\ell(\beta, \gamma_\beta; \mathbf{x}, y, \mathbf{z})] = 0.$

(iii) *Let $\mathbb{Q}$ be the measure induced by* $(\mathbf{x}, y, \mathbf{z})$. *Then for* $i = 1, \ldots, p$, *the functions* $\frac{d}{d\beta_i}\ell(\cdot;\beta,\gamma_\beta)$ *are linearly* $\mathbb{Q}$ *- independent. That is, if*

$$\sum_{k=1}^{m} a_k \frac{d}{d\beta_j}\ell(\mathbf{x}_k, y_k, \mathbf{z}_k; \beta, \gamma_\beta) = 0$$

for $\mathbb{Q}$ *- a.a.* $(\mathbf{x}, y, \mathbf{z})$, *then* $a_k = 0$ *for all* k.

(iv) *There exists a neighborhood* N_0 *of* β_0, *such that for all* $(\mathbf{x}, y, \mathbf{z})$ *the map* $\beta \mapsto \frac{d^2}{d\beta_j d\beta_i}\ell(\beta, \gamma_\beta; \mathbf{x}, y, \mathbf{z})$ *is continuous on* N_0, *and,*

(v) $\mathbb{E} \sup_{\beta \in N_0} |\frac{d^2}{d\beta_j d\beta_i}\ell(\beta, \gamma_\beta; \mathbf{x}, y, \mathbf{z})| < \infty$.

Then the LAN condition holds.

Proof. See Pfanzagl (1994, Page 265). □

We now prove Theorem 3.4.1.

Proof. **[Theorem 3.4.1]** We obtain a proof by contradiction. So let Assumption 3.4.1 hold, and suppose that there exists a regular $n^{1/2}$ consistent estimator for β_0 that achieves the efficiency bounds when the parametric scores are projected onto $T(\mathcal{F}, f^*)$. Let $\hat{\beta}_n$ denote this estimator. Then as a consequence of the the convolution theorem (Pfanzagl 1994, Page 289), $\hat{\beta}_n$ is asymptotically linear. That is, there exists a curve $\lambda_\beta \in \mathcal{F}$ satisfying $\lambda_\beta|_{\beta=\beta_0} = f^*$, such that

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2} I_2^{-1} \sum_{i=1}^{n} \frac{d}{d\beta}\ell(\beta_0, \lambda_{\beta_0}; \mathbf{x}_i, y_i, \mathbf{z}_i) + o_p(1), \qquad \text{(C.1)}$$

where I_2 is defined in Assumption 3.4.1. Notice that since $\hat{\beta}_n$ achieves the lower bound I_2^{-1}, by Theorem 2.7.3 λ_β is a least favorable curve in $\mathcal{F}$. In particular, this implies that

$$I_2 = \mathbb{E}\left[\frac{d}{d\beta}\ell(\beta_0, \lambda_{\beta_0})\right]\left[\frac{d}{d\beta}\ell(\beta_0, \lambda_{\beta_0})\right]'.$$

But as the LAN condition holds for a larger class of functions, it implies that for all $(\delta, \gamma_\beta) \in \mathbb{R}^p \times \overline{lin\mathcal{F}}$ such that $\gamma_{\beta_0} = f^*$,

$$\mathcal{L}_n = n^{-1/2}\sum_{i=1}^{n} \boldsymbol{\delta}' \frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_0, \gamma_{\boldsymbol{\beta}_0}; \mathbf{x}_i, y_i, \mathbf{z}_i) - \frac{1}{2}\boldsymbol{\delta}' I_L \boldsymbol{\delta} + o_p(1), \tag{C.2}$$

where,

$$I_L = \mathbb{E}\left[\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_0, \gamma_{\boldsymbol{\beta}_0})\right]\left[\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_0, \gamma_{\boldsymbol{\beta}_0})\right]'.$$

From (C.1) and (C.2) it is clear that under $\boldsymbol{\beta}_0$,

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, I_2^{-1}) \quad \text{and,}$$

$$\mathcal{L}_n \xrightarrow{\mathrm{d}} \mathrm{N}(-\frac{1}{2}\boldsymbol{\delta}' I_L \boldsymbol{\delta}, \boldsymbol{\delta}' I_L \boldsymbol{\delta}).$$

Therefore, by using the Cramér-Wold device given in Proposition H.1

$$\begin{pmatrix} n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\ \mathcal{L}_n \end{pmatrix} \xrightarrow[\boldsymbol{\beta}_0]{\mathrm{d}} \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\boldsymbol{\delta}' I_L \boldsymbol{\delta} \end{bmatrix}, \begin{bmatrix} I_2^{-1} & \boldsymbol{\zeta} \\ \boldsymbol{\zeta}' & \boldsymbol{\delta}' I_L \boldsymbol{\delta} \end{bmatrix}\right),$$

with, $\boldsymbol{\zeta} = I_2^{-1}\mathbb{E}\left[\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_0, \lambda_{\boldsymbol{\beta}_0})\right]\left[\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_0, \gamma_{\boldsymbol{\beta}_0})\right]'\boldsymbol{\delta}$. Hence by LeCam's Third Lemma (Rieder 1994, Page 44),

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow[\boldsymbol{\beta}_n]{\mathrm{d}} \mathrm{N}\left(I_2^{-1}\mathbb{E}\left[\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_0, \lambda_{\boldsymbol{\beta}_0})\right]\left[\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_0, \gamma_{\boldsymbol{\beta}_0})\right]'\boldsymbol{\delta}, I_2^{-1}\right).$$

And since $n^{1/2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) = \boldsymbol{\delta}$, this implies that

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) = n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) - n^{1/2}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0) \xrightarrow[\boldsymbol{\beta}_n]{\mathrm{d}} \mathrm{N}\left(\boldsymbol{\mu}, I_2^{-1}\right),$$

where the bias ($\boldsymbol{\mu}$) of the asymptotic distribution is

$$\boldsymbol{\mu} = I_2^{-1}\mathbb{E}\left[\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_0, \lambda_{\boldsymbol{\beta}_0})\right]\left[\frac{d}{d\boldsymbol{\beta}}\ell(\boldsymbol{\beta}_0, \gamma_{\boldsymbol{\beta}_0})\right]'\boldsymbol{\delta} - \boldsymbol{\delta}.$$

Therefore, $\hat{\boldsymbol{\beta}}_n$ is a regular estimator iff its asymptotic distribution under $\boldsymbol{\beta}_n$ does not depend upon $\boldsymbol{\delta}$. That is, $\hat{\boldsymbol{\beta}}_n$ is regular iff $\boldsymbol{\mu} = \mathbf{0}$. But,

$$\mu = \mathbf{0} \Leftrightarrow I_2^{-1}\mathbb{E}\left[\frac{d}{d\beta}\ell(\beta_0,\lambda_{\beta_0})\right]\left[\frac{d}{d\beta}\ell(\beta_0,\gamma_{\beta_0})\right]' - \mathbb{I}_{p\times p} = \mathbf{0} \tag{C.3}$$

$$\Leftrightarrow \mathbb{E}\left[\frac{d}{d\beta}\ell(\beta_0,\lambda_{\beta_0})\right]\left[\frac{d}{d\beta}\ell(\beta_0,\gamma_{\beta_0})\right]' = I_2 \tag{C.4}$$

$$\Leftrightarrow \mathbb{E}\left[\frac{d}{d\beta}\ell(\beta_0,\lambda_{\beta_0})\right]\left[\frac{d}{d\beta}\ell(\beta_0,\gamma_{\beta_0})\right]' = \mathbb{E}\left[\frac{d}{d\beta}\ell(\beta_0,\lambda_{\beta_0})\right]\left[\frac{d}{d\beta}\ell(\beta_0,\lambda_{\beta_0})\right]' \tag{C.5}$$

$$\Leftrightarrow \mathbb{E}\left\{\frac{d}{d\beta}\ell(\beta_0,\lambda_{\beta_0})\left[\frac{d}{d\beta}\ell(\beta_0,\gamma_{\beta_0}) - \frac{d}{d\beta}\ell(\beta_0,\lambda_{\beta_0})\right]\right\} = \mathbf{0} \tag{C.6}$$

$$\Leftrightarrow \mathbb{E}\left\{\frac{d}{d\beta}\ell(\beta_0,\lambda_{\beta_0})\left[\frac{\partial\ell(\beta_0,\gamma_{\beta_0})}{\partial\gamma}\left(\frac{d}{d\beta}\gamma_{\beta_0}\right) - \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right)\right]\right\} = \mathbf{0} \tag{C.7}$$

$$\begin{aligned}\Leftrightarrow \mathbb{E}&\left[\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right)\right] \times \\ &\left[\frac{\partial\ell(\beta_0,\gamma_{\beta_0})}{\partial\gamma}\left(\frac{d}{d\beta}\gamma_{\beta_0}\right) - \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right)\right] = \mathbf{0}\end{aligned} \tag{C.8}$$

$$\begin{aligned}\Leftrightarrow \mathbb{E}&\left[\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right)\right]\frac{\partial\ell(\beta_0,\gamma_{\beta_0})}{\partial\gamma}\left(\frac{d}{d\beta}\gamma_{\beta_0}\right) - \\ \mathbb{E}&\left[\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right)\right]\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right) = \mathbf{0}\end{aligned} \tag{C.9}$$

$$\Leftrightarrow \mathbb{E}\left[\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right)\right]\frac{\partial\ell(\beta_0,\gamma_{\beta_0})}{\partial\gamma}\left(\frac{d}{d\beta}\gamma_{\beta_0}\right) = \mathbf{0} \tag{C.10}$$

$$\Leftrightarrow \mathbb{E}\left[\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right)\right]\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\gamma_{\beta_0}\right) = \mathbf{0}. \tag{C.11}$$

REMARK C.1. We now explain briefly how the above equations are obtained.

(i) (C.5) simply follows from the definition of I_2.

(ii) To see how (C.7) is obtained from the previous equation, notice that

$$\begin{aligned}\frac{d}{d\beta}\ell(\beta_0,\gamma_{\beta_0}) &= \frac{\partial\ell(\beta_0,\gamma_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\gamma_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\gamma_{\beta_0}\right) \text{ and,} \\ \frac{d}{d\beta}\ell(\beta_0,\lambda_{\beta_0}) &= \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right).\end{aligned}$$

However $\frac{\partial\ell(\beta_0,\gamma_{\beta_0})}{\partial\beta} = \frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta}$, since the scores with respect to the parameters of interest are equal.

(iii) Now $\frac{d}{d\beta}\lambda_{\beta_0}$ is the least favorable direction in $T(\mathcal{F}, f^*)$, and (C.10) follows from the the previous equation since

$$\mathbb{E}\left[\frac{\partial\ell(\beta_0, \lambda_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0, \lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right)\right]\frac{\partial\ell(\beta_0, \lambda_{\beta_0})}{\partial\lambda}\left(\frac{d}{d\beta}\lambda_{\beta_0}\right) = \mathbf{0}.$$

By Theorem H.3,this equation is just one of the conditions which are necessary and sufficient for projecting $\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\beta}$ onto $T(\mathcal{F}, f^*)$.

(iv) Finally, (C.11) follows from the previous equation because $\frac{\partial\ell(\beta_0,\lambda_{\beta_0})}{\partial\lambda}$ is the restriction of the Fréchet derivative $\frac{\partial\ell(\beta_0,\gamma_{\beta_0})}{\partial\gamma}$ to $T(\mathcal{F}, f^*)$.

Hence, Theorem H.2 and (C.11) imply that $\frac{d}{d\beta}\lambda_{\beta_0} \in \times_{i=1}^{p} T(\mathcal{F}, f^*)$ is the unique solution to the optimization problem

$$\inf_{\xi \in \times_{i=1}^{p} \overline{lin\, T(\mathcal{F}, f^*)}} \mathbb{E}\left[\frac{\partial\ell(\beta_0, \lambda_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0, \lambda_{\beta_0})}{\partial\lambda}(\xi)\right]\left[\frac{\partial\ell(\beta_0, \lambda_{\beta_0})}{\partial\beta} + \frac{\partial\ell(\beta_0, \lambda_{\beta_0})}{\partial\lambda}(\xi)\right]' .$$

This means that even when we search over a bigger space $\overline{lin\, T(\mathcal{F}, f^*)}$, the least favorable direction $\frac{d}{d\beta}\lambda_{\beta_0}$ is found to lie in the strictly smaller space $T(\mathcal{F}, f^*)$. But this violates Assumption 3.4.1, implying that $\mu \neq \mathbf{0}$. Hence under β_n, the bias in the asymptotic distribution of $n^{1/2}(\hat{\beta}_n - \beta_0)$ depends upon δ, which implies that that $\hat{\beta}_n$ is not regular. But we had started with the assumption that $\hat{\beta}_n$ was regular. Hence we get a contradiction. Notice that dependence upon δ is considered undesirable for the following reason. For some $0 < \alpha < 1/2$ choose $\delta = o(n^{\alpha})$ so that $\beta_n \rightarrow \beta_0$, while the bias μ explodes to $\pm\infty$. But this implies that the sequence $n^{1/2}(\hat{\beta}_n - \beta_0)$ is not even tight, much less regular. □

APPENDIX D

PROOFS OF RESULTS IN SECTION 3.5

PROOF. **[Theorem 3.5.1]** First notice that $\hat{\boldsymbol{\beta}}_n$ is measurable since $L_n(\boldsymbol{\beta}, \hat{\eta}_{\boldsymbol{\beta}}; \mathbf{x}, y, \mathbf{z})$ is continuous in $\boldsymbol{\beta}$, and is a measurable function of $(\mathbf{x}, y, \mathbf{z})$ for each $\boldsymbol{\beta}$. Now let $a(\boldsymbol{\beta}) = \mathbb{E}\,\ell(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}}; \mathbf{x}, y, \mathbf{z})$. Then by Assumption 3.5.1, $a(\boldsymbol{\beta}) < a(\boldsymbol{\beta}_0)$ if $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$. Now by the WLLN, $\frac{1}{n} L_n(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}}) \xrightarrow{p} a(\boldsymbol{\beta})$ for each $\boldsymbol{\beta} \in \mathbf{B}$. In particular, this implies that $\frac{1}{n} L_n(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}}) = O_p(1)$. Furthermore, by using the mean value theorem for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbf{B}$,

$$
\begin{aligned}
\frac{1}{n}|L_n(\boldsymbol{\beta}_1, \eta_{\boldsymbol{\beta}_1}) - L_n(\boldsymbol{\beta}_2, \eta_{\boldsymbol{\beta}_2})| &= \frac{1}{n}|\sum_{j=1}^{n} \ell(\boldsymbol{\beta}_1, \eta_{\boldsymbol{\beta}_1}; \mathbf{x}_j, y_j, \mathbf{z}_j) - \sum_{j=1}^{n} \ell(\boldsymbol{\beta}_2, \eta_{\boldsymbol{\beta}_2}; \mathbf{x}_j, y_j, \mathbf{z}_j)| \\
&\leq \frac{1}{n} \sum_{j=1}^{n} |\ell(\boldsymbol{\beta}_1, \eta_{\boldsymbol{\beta}_1}; \mathbf{x}_j, y_j, \mathbf{z}_j) - \ell(\boldsymbol{\beta}_2, \eta_{\boldsymbol{\beta}_2}; \mathbf{x}_j, y_j, \mathbf{z}_j)| \\
&\leq A_n \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|,
\end{aligned}
$$

with A_n defined as

$$
\begin{aligned}
&\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{p} \sup_{\boldsymbol{\beta}, \eta \in \mathbf{B} \times \mathcal{H}} \left| \frac{\partial \ell(\boldsymbol{\beta}, \eta; \mathbf{x}_j, y_j, \mathbf{z}_j)}{\partial \beta_i} \right| \\
&\qquad\qquad + p \sup_{\boldsymbol{\beta}, \mathbf{z} \in \mathbf{B} \times \mathbf{Z}} |\frac{d}{d\beta_i} \eta_{\boldsymbol{\beta}}(\mathbf{z})| \frac{1}{n} \sum_{j=1}^{n} \sup_{\boldsymbol{\beta}, \eta \in \mathbf{B} \times \mathcal{H}} \left| \frac{\partial \ell(\boldsymbol{\beta}, \eta; \mathbf{x}_j, y_j, \mathbf{z}_j)}{\partial \eta} \right|,
\end{aligned}
$$

and $A_n = O_p(1)$ by Assumption 3.5.2. Hence the modulus of continuity

$$w_{\frac{1}{n}L_n}(\delta) = \sup_{\|\beta_1-\beta_2\|<\delta} \frac{1}{n}|L_n(\beta_1, \eta_{\beta_1}) - L_n(\beta_2, \eta_{\beta_2})| \le A_n\delta.$$

Now choose any $\xi, \epsilon > 0$. Then since $\frac{1}{n}L_n(\beta, \eta_\beta) = O_p(1)$, if we can find a $\delta > 0$ such that $\Pr\{w_{\frac{1}{n}L_n}(\delta) > \epsilon\} < \xi$, $\frac{1}{n}L_n(\beta, \eta_\beta)$ will be tight in $\mathcal{C}(\mathbf{B})$. That such a δ exists, can be seen as follows.

First notice that since $A_n = O_p(1)$, there exists a M_ξ such that $\Pr\{|A_n| > M_\xi\} < \xi$. Now let $\delta = \epsilon/M_\xi$. Then,

$$\begin{aligned}
\Pr\{w_{\frac{1}{n}L_n}(\delta) > \epsilon\} &\le \Pr\{A_n\delta > \epsilon\} \\
&= \Pr\{A_n > \epsilon/\delta\} \\
&= \Pr\{A_n > M_\xi\} \\
&\le \Pr\{|A_n| > M_\xi\} \\
&< \xi.
\end{aligned}$$

Hence, $\frac{1}{n}L_n(\beta, \eta_\beta)$ is tight in C(**B**). This implies that for any subsequence $\{n'\} \subset \{n\}$, there exists a further subsequence $\{m\} \subset \{n'\}$ such that

$$\frac{1}{m}L_m(\beta, \eta_\beta) \xrightarrow{\mathrm{d}} a(\beta) \qquad \text{in } \mathrm{C}(B). \tag{D.1}$$

Now, for each $\beta \in \mathbf{B}$, and $\eta \in \mathcal{H}$

$$\begin{aligned}
\frac{1}{m}|L_m(\beta, \hat{\eta}_\beta) - L_m(\beta, \eta_\beta)| &= \frac{1}{m}|\sum_{j=1}^{m} \ell(\beta, \hat{\eta}_\beta; \mathbf{x}_j, y_j, \mathbf{z}_j) - \sum_{j=1}^{m} \ell(\beta, \eta_\beta; \mathbf{x}_j, y_j, \mathbf{z}_j)| \\
&\le \frac{1}{m}\sum_{j=1}^{m} \sup_{\beta,\eta} \left|\frac{\partial \ell(\beta, \eta; \mathbf{x}_j, y_j, \mathbf{z}_j)}{\partial \eta}\right| \sup_{\beta} |\hat{\eta}_\beta(\mathbf{z}_j) - \eta_\beta(\mathbf{z}_j)| \\
&\le \sup_{\beta,\mathbf{z}} |\hat{\eta}_\beta(\mathbf{z}) - \eta_\beta(\mathbf{z})|\, B_m,
\end{aligned}$$

with $B_m = \frac{1}{m}\sum_{j=1}^{m} \sup_{\boldsymbol{\beta},\eta} |\frac{\partial \ell(\boldsymbol{\beta},\eta;\mathbf{x}_j,y_j,\mathbf{z}_j)}{\partial \eta}|$. Then since $B_m = O_p(1)$ by Assumption 3.5.2 and $\sup_{\boldsymbol{\beta},\mathbf{z}} |\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z}) - \eta_{\boldsymbol{\beta}}(\mathbf{z})| = o_p(1)$ by Assumption 3.5.3, we have

$$\sup_{\boldsymbol{\beta}} \frac{1}{m} |L_m(\boldsymbol{\beta}, \hat{\eta}_{\boldsymbol{\beta}}) - L_m(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}})| \xrightarrow{p} 0. \tag{D.2}$$

Also notice that from (D.1)

$$\sup_{\boldsymbol{\beta}} |\frac{1}{m} L_m(\boldsymbol{\beta}, \hat{\eta}_{\boldsymbol{\beta}}) - a(\boldsymbol{\beta})| \xrightarrow{p} 0. \tag{D.3}$$

Therefore, by (D.2) and (D.3)

$$\begin{aligned} \sup_{\boldsymbol{\beta}} |\frac{1}{m} L_m(\boldsymbol{\beta}, \hat{\eta}_{\boldsymbol{\beta}}) - a(\boldsymbol{\beta})| &\le \sup_{\boldsymbol{\beta}} |\frac{1}{m} L_m(\boldsymbol{\beta}, \hat{\eta}_{\boldsymbol{\beta}}) - \frac{1}{m} L_m(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}})| \\ &\quad + \sup_{\boldsymbol{\beta}} |\frac{1}{m} L_m(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}}) - a(\boldsymbol{\beta})| \\ &\xrightarrow{p} 0. \end{aligned} \tag{D.4}$$

But this immediately implies that

$$\sup_{\boldsymbol{\beta}} \frac{1}{m} L_m(\boldsymbol{\beta}, \hat{\eta}_{\boldsymbol{\beta}}) \xrightarrow{p} \sup_{\boldsymbol{\beta}} a(\boldsymbol{\beta}) = a(\boldsymbol{\beta}_0). \tag{D.5}$$

Furthermore,

$$\begin{aligned} |a(\hat{\boldsymbol{\beta}}_m) - a(\boldsymbol{\beta}_0)| &\le |\frac{1}{m} L_m(\hat{\boldsymbol{\beta}}_m, \hat{\eta}_{\hat{\boldsymbol{\beta}}_m}) - a(\hat{\boldsymbol{\beta}}_m)| \\ &\quad + |\frac{1}{m} L_m(\hat{\boldsymbol{\beta}}_m, \hat{\eta}_{\hat{\boldsymbol{\beta}}_m}) - a(\boldsymbol{\beta}_0)|. \end{aligned} \tag{D.6}$$

Now since (D.4) holds for all $\boldsymbol{\beta} \in B$, $|\frac{1}{m} L_m(\hat{\boldsymbol{\beta}}_m, \hat{\eta}_{\hat{\boldsymbol{\beta}}_m}) - a(\hat{\boldsymbol{\beta}}_m)| \xrightarrow{p} 0$. Moreover, since $L_m(\hat{\boldsymbol{\beta}}_m, \hat{\eta}_{\hat{\boldsymbol{\beta}}_m}) = \sup_{\boldsymbol{\beta}} L_m(\boldsymbol{\beta}, \hat{\eta}_{\boldsymbol{\beta}})$, (D.5) implies that $\frac{1}{m} L_m(\hat{\boldsymbol{\beta}}_m, \hat{\eta}_{\hat{\boldsymbol{\beta}}_m}) \xrightarrow{p} a(\boldsymbol{\beta}_0)$. Following these observations (D.6) is reduced to $|a(\hat{\boldsymbol{\beta}}_m) - a(\boldsymbol{\beta}_0)| \xrightarrow{p} 0$, i.e.

$$a(\hat{\boldsymbol{\beta}}_m) \xrightarrow{p} a(\boldsymbol{\beta}_0). \tag{D.7}$$

Now the identification condition implies that for any $\epsilon > 0$ and any $\beta \neq \beta_0$, there exists a neighborhood N_β of β such that

$$\inf_{\tilde{\beta}} |a(\tilde{\beta}) - a(\beta_0)| > \epsilon. \tag{D.8}$$

But this means that

$$\Pr\{\hat{\beta}_m \in N_\beta\} \leq \Pr\{|a(\hat{\beta}_m) - a(\beta_0)| > \epsilon\} \to 0 \tag{D.9}$$

by (D.7). So let N_0 denote any neighborhood of β_0. Notice that the collection of neighborhoods $\{N_\beta : \beta \in B, \beta \neq \beta_0\}$ is an open cover of $B \backslash N_0$, with N_β satisfying (D.8). But since $B \backslash N_0$ is compact, there exists a finite subcover $\{N_{\beta_1}, \dots, N_{\beta_k}\}$. Therefore, from (D.7) and (D.8)

$$\begin{aligned}\Pr\{\hat{\beta}_m \notin N_0\} &= \Pr\{\hat{\beta}_m \in B \backslash N_0\} \\ &\leq \Pr\{\hat{\beta}_m \in \cup_{i=1}^{k} N_{\beta_i}\} \\ &\leq \sum_{i=1}^{k} \Pr\{\hat{\beta}_m \in N_{\beta_i}\} \\ &\to 0.\end{aligned}$$

And this implies that $\hat{\beta}_m \xrightarrow{p} \beta_0$, as $m \to \infty$. But since β_0 does not depend in any way upon the subsequence $\{m\} \subset \{n'\}$, this convergence holds with m replaced by n. That is, $\hat{\beta}_n \xrightarrow{p} \beta_0$, as $n \to \infty$. □

For proving asymptotic normality of $\hat{\beta}_n$, we need the following uniform CLT, and Lemma D.1 and Lemma D.2 given below.

THEOREM D.1 (JAIN AND MARCUS). *Let* C(S) *be the space of real-valued continuous functions on a compact metric space* (S, d). *Also let* X_n *be a sequence of* C(S) *- valued random variables on* $(\Omega, \mathfrak{F}, P)$ *satisfying*

(i) $\mathbb{E} X_n(s) = 0$, *for all* $s \in S$, *and,*

(ii) $\sup_{s\in S} \mathbb{E} X_n^2(s) = 1$.

Suppose there exist a nonnegative random variable M, $\mathbb{E} M^2 = 1$, *and a metric* ρ *on* S, *which is continuous with respect to* d, *such that given* $s, t \in S$, $\omega \in \Omega$,

$$|X_n(s,\omega) - X_n(t,\omega)| \le M(\omega)\rho(s,t).$$

If $\int H_\rho^{1/2}(S,\epsilon)\,d\epsilon < \infty$, *then the sequence* X_n *obeys the central limit theorem. That is,* $n^{-1/2}\sum_{i=1}^n X_i$ *converges weakly to a Gaussian measure on* $\mathrm{C}(S)$.

PROOF. See Jain and Marcus (1975). □

NOTATION D.0.1. Let $\mathfrak{G}_0$ be a compact subset of the closed unit ball in $\overline{lin\,T(\mathcal{F},f^*)}$, w.r.t. the sup norm, centered at the zero function. That is,

$$\mathfrak{G}_0 = \{f \in \overline{lin\,T(\mathcal{F},f^*)} : \|f\| \le 1, \|\frac{\partial f(u,v)}{\partial u}\| \le 1, \|\frac{\partial f(u,v)}{\partial v}\| \le 1\},$$

where $\|\cdot\|$ denotes the sup norm. Furthermore, let $H(\mathfrak{G}_0,\epsilon)$ denote the metric entropy of $\mathfrak{G}_0$ under the sup metric when $\mathfrak{G}_0$ is covered by a finite ϵ net. □

REMARK D.1. (i) $\mathfrak{G}_0$ is compact in $\mathrm{C}(\mathbf{Z})$, and can therefore itself be regarded as a compact metric space with the sup metric.

(ii) Let $K \subset \mathbb{R}^k$ be compact. Then by a result of Kolmogorov and Tihomirov (1961), for every $\mathcal{G} \subset \mathrm{C}^r(K)$

$$H(\mathcal{G},\epsilon) \le \frac{1}{\epsilon^{k/r}}.$$

In our case $k = 2$, $r = 2$, and $\mathcal{G} = \mathfrak{G}_0$. Therefore, $H(\mathfrak{G}_0,\epsilon) \le \frac{1}{\epsilon}$. □

LEMMA D.1. *Let* η_β *be a least favorable curve. Then for* $i = 1,\ldots,p$,

(i) $n^{-1/2}[\frac{d}{d\beta_i}\frac{\partial L_n(\beta_0,\eta_{\beta_0})}{\partial\eta}](\hat{\eta}_{\beta_0} - \eta_{\beta_0}) = o_p(1)$, *and,*

(ii) $n^{-1/2}\frac{\partial L_n(\beta_0,\eta_{\beta_0})}{\partial\eta}(\frac{d}{d\beta_i}\hat{\eta}_{\beta_0} - \frac{d}{d\beta_i}\eta_{\beta_0}) = o_p(1)$.

Proof. We will show (i). The proof of (ii) is similar. Now since the Fréchet derivative

$$\frac{\partial \ell(\boldsymbol{\beta}, \eta_{\boldsymbol{\beta}}; \mathbf{x}_j, y_j, \mathbf{z}_j)}{\partial \eta} = y_j - \mathbf{x}_j\boldsymbol{\beta} - \eta_{\boldsymbol{\beta}}(\mathbf{z}_j),$$

we have

$$\begin{aligned}
[\frac{d}{d\beta_i}\frac{\partial L_n(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0})}{\partial \eta}](\hat{\eta}_{\boldsymbol{\beta}_0} - \eta_{\boldsymbol{\beta}_0}) &= [\sum_{j=1}^{n} \frac{d}{d\beta_i}\frac{\partial \ell(\boldsymbol{\beta}_0, \eta_{\boldsymbol{\beta}_0}; \mathbf{x}_j, y_j, \mathbf{z}_j)}{\partial \eta}](\hat{\eta}_{\boldsymbol{\beta}_0}(\mathbf{z}_j) - \eta_{\boldsymbol{\beta}_0}(\mathbf{z}_j)) \\
&= -\sum_{j=1}^{n}[x_{ij} + \frac{d}{d\beta_i}\eta_{\boldsymbol{\beta}_0}(\mathbf{z}_j)](\hat{\eta}_{\boldsymbol{\beta}_0}(\mathbf{z}_j) - \eta_{\boldsymbol{\beta}_0}(\mathbf{z}_j)) \\
&= -\sum_{j=1}^{n} A(x_{ij}, \mathbf{z}_j)\xi_0(\mathbf{z}_j),
\end{aligned}$$

where, $A(x_{ij}, \mathbf{z}_j) = x_{ij} + \frac{d}{d\beta_i}\eta_{\boldsymbol{\beta}_0}(\mathbf{z}_j)$, and $\xi_0(\mathbf{z}_j) = \hat{\eta}_{\boldsymbol{\beta}_0}(\mathbf{z}_j) - \eta_{\boldsymbol{\beta}_0}(\mathbf{z}_j) \in \overline{lin\, T(\mathcal{F}, f^{*})}$.

Remark D.2. The notation x_{ij} here refers to the ith. element of the vector $\mathbf{x}_j$. □

Notice that since x and $\mathbf{z}$ come from distributions with compact support,

$$\sup_{x,\mathbf{z}} |A(x, \mathbf{z})| = M < \infty.$$

This also implies that $\mathbb{E}\,|A(x, \mathbf{z})|^2 < \infty$.

For any $\xi \in \mathfrak{G}_0$, let us now look at the map $\xi \mapsto A(x, \mathbf{z})\xi(\mathbf{z})$. First, notice that since

$$\begin{aligned}
||A(x, \mathbf{z})\xi(\mathbf{z})|| &= \sup_{x,\mathbf{z}} |A(x, \mathbf{z})\xi(\mathbf{z})| \\
&= \sup_{x,\mathbf{z}} |A(x, \mathbf{z})||\xi(\mathbf{z})| \\
&\leq ||\xi|| \sup_{x,\mathbf{z}} |A(x, \mathbf{z})| \\
&\leq M||\xi||,
\end{aligned}$$

$\xi \mapsto A(x, \mathbf{z})\xi(\mathbf{z})$ is a continuous mapping on $\mathfrak{G}_0$. That is, this map is an element of C($\mathfrak{G}_0$), the space of all continuous functions on $\mathfrak{G}_0$.

Secondly, since each ξ in $\mathfrak{G}_0$ is also an element of $\overline{lin\,T(\mathcal{F}, f^*)}$,

$$\mathbb{E}\,A(x,\mathbf{z})\xi(\mathbf{z}) = 0$$

by the least favorable curve property of η_β. With these two points in mind, we can now treat $\{A(x_{ij}, \mathbf{z}_j)\xi(\mathbf{z}_j)\}_{j=1}^n$ as a sequence of C($\mathfrak{G}_0$) - valued random variables with zero mean and finite variance. Note that the random variable $A(x,\mathbf{z})\xi(\mathbf{z})$ also has the property that for all $\xi_1, \xi_2 \in \mathfrak{G}_0$,

$$\begin{aligned}|A(x,\mathbf{z})\xi_1(\mathbf{z}) - A(x,\mathbf{z})\xi_2(\mathbf{z})| &= |A(x,\mathbf{z})||\xi_1(\mathbf{z}) - \xi_2(\mathbf{z})| \\ &\leq |A(x,\mathbf{z})|\|\xi_1 - \xi_2\|,\end{aligned}$$

with $\mathbb{E}\,|A(x,\mathbf{z})|^2 < \infty$. This fact coupled with the observation that

$$\int_0^1 H^{1/2}(\mathfrak{G}_0, \epsilon)\,d\epsilon < \infty,$$

allows us to utilize the uniform CLT of Jain and Marcus (1975). Hence from Theorem D.1, for all $\xi \in \mathfrak{G}_0$

$$\begin{aligned}-n^{-1/2}\sum_{j=1}^n A(x_{ij}, \mathbf{z}_j)\xi(\mathbf{z}_j) &= -n^{-1/2}\sum_{j=1}^n [x_{ij} + \frac{d}{d\beta_i}\eta_{\beta_0}(\mathbf{z}_j)]\xi(\mathbf{z}_j) \qquad \text{(D.10)} \\ &= O_p(1). \qquad \text{(D.11)}\end{aligned}$$

Now by Assumption 3.5.3, as $n \to \infty$,

$$\Pr\{n^{\alpha_1}(\hat{\eta}_{\beta_0} - \eta_{\beta_0}) \in \mathfrak{G}_0\} \to 1.$$

So w.l.o.g. assume that $n^{\alpha_1}(\hat{\eta}_{\beta_0} - \eta_{\beta_0}) \in \mathfrak{G}_0$, for the probability that this event does not happen can be made arbitrarily small. Then since (D.10) and (D.11) hold for all $\xi \in \mathfrak{G}_0$, we have

$$-n^{-1/2}\sum_{j=1}^n [x_{ij} + \frac{d}{d\beta_i}\eta_{\beta_0}(\mathbf{z}_j)]n^{\alpha_1}(\hat{\eta}_{\beta_0}(\mathbf{z}_j) - \eta_{\beta_0}(\mathbf{z}_j)) = O_p(1),$$

which implies that

$$n^{-1/2}\sum_{j=1}^{n}[x_{ij} + \frac{d}{d\beta_i}\eta_{\beta_0}(\mathbf{z}_j)](\hat{\eta}_{\beta_0}(\mathbf{z}_j) - \eta_{\beta_0}(\mathbf{z}_j)) = o_p(1).$$

□

Lemma D.2. *Let η_β be a least favorable curve. Then for $i, j = 1, \ldots, p$,*

(i) $L_n(\beta, \hat{\eta}_\beta) - L_n(\beta, \eta_\beta) = r_n^{(1)}(\beta)$, *with*

$$\sup_{\beta}\left|\frac{1}{n}\frac{d^2 r_n^{(1)}(\beta)}{d\beta_i d\beta_j}\right| = o_p(1),$$

(ii) $L_n(\beta, \hat{\eta}_\beta) = L_n(\beta, \eta_\beta) + \frac{\partial L_n(\beta,\eta_\beta)}{\partial\eta}(\hat{\eta}_\beta - \eta_\beta) + r_n^{(2)}(\beta)$, *with*

$$n^{-1/2}\frac{d}{d\beta_i}r_n^{(2)}(\beta_0) = o_p(1).$$

Proof. **[Lemma D.2(i)]** By a Taylor expansion,

$$\ell(\beta, \hat{\eta}_\beta; \mathbf{x}, y, \mathbf{z}) = \ell(\beta, \eta_\beta; \mathbf{x}, y, \mathbf{z}) + r_a(\beta; \mathbf{x}, y, \mathbf{z}),$$

where,

$$r_a(\beta; \mathbf{x}, y, \mathbf{z}) = \int_{t=0}^{1}\frac{\partial\ell(\beta, t\hat{\eta}_\beta + (1-t)\eta_\beta)}{\partial\eta}dt \cdot (\hat{\eta}_\beta(\mathbf{z}) - \eta_\beta(\mathbf{z})).$$

Now let $r_a(\beta; \mathbf{x}, y, \mathbf{z}) = Q^{(1)}(\beta; \mathbf{x}, y, \mathbf{z})(\hat{\eta}_\beta(\mathbf{z}) - \eta_\beta(\mathbf{z}))$, and note that since, $L_n(\beta, \hat{\eta}_\beta) - L_n(\beta, \eta_\beta) = \sum_{k=1}^{n}\ell(\beta, \hat{\eta}_\beta; \mathbf{x}_k, y_k, \mathbf{z}_k) - \ell(\beta, \eta_\beta; \mathbf{x}_k, y_k, \mathbf{z}_k)$, we have

$$\begin{aligned}
r_n^{(1)}(\boldsymbol{\beta}) &= \sum_{k=1}^{n} Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)(\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z}_k)-\eta_{\boldsymbol{\beta}}(\mathbf{z}_k)), \quad \text{and}\\
\frac{d}{d\beta_i} r_n^{(1)}(\boldsymbol{\beta}) &= \sum_{k=1}^{n}\left[\frac{d}{d\beta_i}Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\right](\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z}_k)-\eta_{\boldsymbol{\beta}}(\mathbf{z}_k))\\
&\quad + \sum_{k=1}^{n} Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\left[\frac{d}{d\beta_i}\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z}_k)-\frac{d}{d\beta_i}\eta_{\boldsymbol{\beta}}(\mathbf{z}_k)\right], \text{implying}\\
\frac{1}{n}\frac{d^2}{d\beta_j d\beta_i} r_n^{(1)}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{k=1}^{n}\left[\frac{d^2}{d\beta_j d\beta_i}Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\right](\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z}_k)-\eta_{\boldsymbol{\beta}}(\mathbf{z}_k))\\
&\quad + \frac{1}{n}\sum_{k=1}^{n}\left[\frac{d}{d\beta_i}Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\right]\left[\frac{d}{d\beta_j}\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z}_k)-\frac{d}{d\beta_j}\eta_{\boldsymbol{\beta}}(\mathbf{z}_k)\right]\\
&\quad + \frac{1}{n}\sum_{k=1}^{n}\left[\frac{d}{d\beta_j}Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\right]\left[\frac{d}{d\beta_i}\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z}_k)-\frac{d}{d\beta_i}\eta_{\boldsymbol{\beta}}(\mathbf{z}_k)\right]\\
&\quad + \frac{1}{n}\sum_{k=1}^{n} Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\left[\frac{d^2}{d\beta_j d\beta_i}\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z}_k)-\frac{d^2}{d\beta_j d\beta_i}\eta_{\boldsymbol{\beta}}(\mathbf{z}_k)\right].
\end{aligned}$$

Therefore, using Assumption 3.5.2 and Assumption 3.5.3

$$\begin{aligned}
\sup_{\boldsymbol{\beta}}\left|\frac{1}{n}\frac{d^2}{d\beta_j d\beta_i} r_n^{(1)}(\boldsymbol{\beta})\right| \le\ & \underbrace{\sup_{\boldsymbol{\beta},\mathbf{z}}\left|\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z})-\eta_{\boldsymbol{\beta}}(\mathbf{z})\right|}_{o_p(1)}\underbrace{\frac{1}{n}\sum_{k=1}^{n}\left|\frac{d^2}{d\beta_j d\beta_i}Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\right|}_{O_p(1)}\\
& + \underbrace{\sup_{\boldsymbol{\beta},\mathbf{z}}\left|\frac{d}{d\beta_j}\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z})-\frac{d}{d\beta_j}\eta_{\boldsymbol{\beta}}(\mathbf{z})\right|}_{o_p(1)}\underbrace{\frac{1}{n}\sum_{k=1}^{n}\left|\frac{d}{d\beta_i}Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\right|}_{O_p(1)}\\
& + \underbrace{\sup_{\boldsymbol{\beta},\mathbf{z}}\left|\frac{d}{d\beta_i}\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z})-\frac{d}{d\beta_i}\eta_{\boldsymbol{\beta}}(\mathbf{z})\right|}_{o_p(1)}\underbrace{\frac{1}{n}\sum_{k=1}^{n}\left|\frac{d}{d\beta_j}Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\right|}_{O_p(1)}\\
& + \underbrace{\sup_{\boldsymbol{\beta},\mathbf{z}}\left|\frac{d^2}{d\beta_j d\beta_i}\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z})-\frac{d^2}{d\beta_j d\beta_i}\eta_{\boldsymbol{\beta}}(\mathbf{z})\right|}_{o_p(1)}\underbrace{\frac{1}{n}\sum_{k=1}^{n}\left|Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\right|}_{O_p(1)}\\
=\ & o_p(1).
\end{aligned}$$

□

Remark D.3. It is instructive to see how the terms involving $Q^{(1)}$ are $O_p(1)$. So let us show that $\frac{1}{n}\sum_{k=1}^{n}\left|\frac{d^2}{d\beta_j d\beta_i}Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\right| = O_p(1)$. A similar reasoning can be used to verify that the other terms are bounded in probability. Thus,

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n}\left|\frac{d^2}{d\beta_j d\beta_i}Q^{(1)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)\right| &= \frac{1}{n}\sum_{k=1}^{n}\left|\frac{d^2}{d\beta_j d\beta_i}\int_{t=0}^{1}\frac{\partial\ell(\boldsymbol{\beta},t\hat{\eta}_{\boldsymbol{\beta}}+(1-t)\eta_{\boldsymbol{\beta}})}{\partial\eta}\,dt\right|\\ &= \frac{1}{n}\sum_{k=1}^{n}\left|\int_{t=0}^{1}\frac{\partial^3\ell(\boldsymbol{\beta},t\hat{\eta}_{\boldsymbol{\beta}}+(1-t)\eta_{\boldsymbol{\beta}})}{\partial\beta_j\partial\beta_i\partial\eta}\,dt\right|\\ &\leq \frac{1}{n}\sum_{k=1}^{n}\int_{t=0}^{1}\left|\frac{\partial^3\ell(\boldsymbol{\beta},t\hat{\eta}_{\boldsymbol{\beta}}+(1-t)\eta_{\boldsymbol{\beta}})}{\partial\beta_j\partial\beta_i\partial\eta}\right|\,dt\\ &\leq \frac{1}{n}\sum_{k=1}^{n}\sup_{\boldsymbol{\beta}\in\mathbf{B}}\sup_{\eta\in\mathcal{H}}\left|\frac{\partial^3\ell(\boldsymbol{\beta},\eta)}{\partial\beta_j\partial\beta_i\partial\eta}\right|\\ &= O_p(1),\end{aligned}$$

since

$$\mathbb{E}\left[\sup_{\boldsymbol{\beta}\in\mathbf{B}}\sup_{\eta\in\mathcal{H}}\left|\frac{\partial^3\ell(\boldsymbol{\beta},\eta)}{\partial\beta_j\partial\beta_i\partial\eta}\right|\right]^2 \leq \mathbb{E}\left\{\sup_{\boldsymbol{\beta}\in\mathbf{B}}\sup_{\eta\in\mathcal{H}}\left|\frac{\partial^3\ell(\boldsymbol{\beta},\eta)}{\partial\beta_j\partial\beta_i\partial\eta}\right|^2\right\} < \infty$$

by Assumption 3.5.2. □

Proof. **[Lemma D.2(ii)]** By a Taylor expansion,

$$\ell(\boldsymbol{\beta},\hat{\eta}_{\boldsymbol{\beta}};\mathbf{x},y,\mathbf{z}) = \ell(\boldsymbol{\beta},\eta_{\boldsymbol{\beta}};\mathbf{x},y,\mathbf{z}) + \frac{\ell(\boldsymbol{\beta},\eta_{\boldsymbol{\beta}};\mathbf{x},y,\mathbf{z})}{\partial\eta}(\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z})-\eta_{\boldsymbol{\beta}}(\mathbf{z})) + r_b(\boldsymbol{\beta};\mathbf{x},y,\mathbf{z}),$$

where,

$$\begin{aligned}r_b(\boldsymbol{\beta};\mathbf{x},y,\mathbf{z}) &= \frac{1}{2}\int_{t=0}^{1}(1-t)\frac{\partial^2\ell(\boldsymbol{\beta},t\hat{\eta}_{\boldsymbol{\beta}}+(1-t)\eta_{\boldsymbol{\beta}})}{\partial^2\eta}\,dt\cdot[\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z})-\eta_{\boldsymbol{\beta}}(\mathbf{z})]^2\\ &= Q^{(2)}(\boldsymbol{\beta};\mathbf{x},y,\mathbf{z})[\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z})-\eta_{\boldsymbol{\beta}}(\mathbf{z})]^2.\end{aligned}$$

But since $L_n(\boldsymbol{\beta},\eta_{\boldsymbol{\beta}}) = \sum_{k=1}^{n}\ell(\boldsymbol{\beta},\eta_{\boldsymbol{\beta}};\mathbf{x}_k,y_k,\mathbf{z}_k)$, we have

$$r_n^{(2)}(\boldsymbol{\beta}) = \sum_{k=1}^{n}Q^{(2)}(\boldsymbol{\beta};\mathbf{x}_k,y_k,\mathbf{z}_k)[\hat{\eta}_{\boldsymbol{\beta}}(\mathbf{z}_k)-\eta_{\boldsymbol{\beta}}(\mathbf{z}_k)]^2,$$

which implies that

$$\frac{d}{d\beta_i} r_n^{(2)}(\beta) = \sum_{k=1}^{n} \left[\frac{d}{d\beta_i} Q^{(2)}(\beta; \mathbf{x}_k, y_k, \mathbf{z}_k) \right] (\hat{\eta}_\beta(\mathbf{z}_k) - \eta_\beta(\mathbf{z}_k))^2$$
$$+ 2 \sum_{k=1}^{n} Q^{(2)}(\beta; \mathbf{x}_k, y_k, \mathbf{z}_k)(\hat{\eta}_\beta(\mathbf{z}_k) - \eta_\beta(\mathbf{z}_k)) \left[\frac{d}{d\beta_i} (\hat{\eta}_\beta(\mathbf{z}_k) - \frac{d}{d\beta_i} \eta_\beta(\mathbf{z}_k)) \right].$$

Thus,

$$\left| \frac{d}{d\beta_i} r_n^{(2)}(\beta) \right| \leq \sup_{\mathbf{z}} |\hat{\eta}_\beta(\mathbf{z}) - \eta_\beta(\mathbf{z})|^2 \sum_{k=1}^{n} \left| \frac{d}{d\beta_i} Q^{(2)}(\beta; \mathbf{x}_k, y_k, \mathbf{z}_k) \right|$$
$$+ 2 \sup_{\mathbf{z}} |\hat{\eta}_\beta(\mathbf{z}) - \eta_\beta(\mathbf{z})| \sup_{\mathbf{z}} \left| \frac{d}{d\beta_i} \hat{\eta}_\beta(\mathbf{z}) - \frac{d}{d\beta_i} \eta_\beta(\mathbf{z}) \right|$$
$$\times \sum_{k=1}^{n} |Q^{(2)}(\beta; \mathbf{x}_k, y_k, \mathbf{z}_k)|.$$

Therefore, using Assumption 3.5.2 and Assumption 3.5.3

$$n^{-1/2} \left| \frac{d}{d\beta_i} r_n^{(2)}(\beta_0) \right| \leq \underbrace{\sup_{\mathbf{z}} [n^{1/4} |\hat{\eta}_{\beta_0}(\mathbf{z}) - \eta_{\beta_0}(\mathbf{z})|]^2}_{o_p(1)} \underbrace{\frac{1}{n} \sum_{k=1}^{n} \left| \frac{d}{d\beta_i} Q^{(2)}(\beta_0; \mathbf{x}_k, y_k, \mathbf{z}_k) \right|}_{O_p(1)}$$
$$+ \underbrace{2 \sup_{\mathbf{z}} n^{1/4} |\hat{\eta}_{\beta_0}(\mathbf{z}) - \eta_{\beta_0}(\mathbf{z})|}_{o_p(1)} \underbrace{\sup_{\mathbf{z}} n^{1/4} \left| \frac{d}{d\beta_i} \hat{\eta}_{\beta_0}(\mathbf{z}) - \frac{d}{d\beta_i} \eta_{\beta_0}(\mathbf{z}) \right|}_{o_p(1)}$$
$$\times \underbrace{\sum_{k=1}^{n} |Q^{(2)}(\beta_0; \mathbf{x}_k, y_k, \mathbf{z}_k)|}_{O_p(1)}$$
$$= o_p(1).$$

Similarly, the terms involving $Q^{(2)}$ can be shown to be $O_p(1)$ by using the reasoning in Remark D.3. □

We are now ready to prove Theorem 3.5.2.

Proof. **[Theorem 3.5.2]** From a Taylor expansion w.r.t. β,

$$\begin{aligned}0 &= \frac{dL_n(\beta, \hat{\eta}_\beta)}{d\beta}\bigg|_{\beta=\hat{\beta}_n} \\ &= \frac{dL_n(\beta_0, \hat{\eta}_{\beta_0})}{d\beta} + \frac{d^2L_n(\beta, \hat{\eta}_\beta)}{d\beta d\beta'}\bigg|_{\beta=\hat{\beta}^*_n}(\hat{\beta}_n - \beta_0),\end{aligned}$$

for some $\hat{\beta}^*_n$ between $\hat{\beta}_n$ and β_0. Notice that since $\hat{\beta}_n$ is consistent, $\hat{\beta}^*_n \xrightarrow{p} \beta_0$. Thus,

$$n^{1/2}(\hat{\beta}_n - \beta_0) = \left[-\frac{1}{n}\frac{d^2L_n(\beta, \hat{\eta}_\beta)}{d\beta d\beta'}\bigg|_{\beta=\hat{\beta}^*_n}\right]^{-1} n^{-1/2}\frac{dL_n(\beta_0, \hat{\eta}_{\beta_0})}{d\beta}.$$

Now using a Taylor expansion w.r.t. η,

$$L_n(\beta, \hat{\eta}_\beta) - L_n(\beta, \eta_\beta) = r_n^{(1)}(\beta)$$

where $r_n^{(1)}(\beta)$ as defined in Lemma D.2(i). Therefore,

$$\frac{1}{n}\frac{d^2L_n(\beta, \hat{\eta}_\beta)}{d\beta_i d\beta_j} - \frac{1}{n}\frac{d^2L_n(\beta, \eta_\beta)}{d\beta_i d\beta_j} = \frac{1}{n}\frac{d^2r_n^{(1)}(\beta)}{d\beta_i d\beta_j},$$

and from Lemma D.2(i) we get that

$$\sup_\beta\left|\frac{1}{n}\frac{d^2L_n(\beta, \hat{\eta}_\beta)}{d\beta d\beta'} - \frac{1}{n}\frac{d^2L_n(\beta, \eta_\beta)}{d\beta d\beta'}\right| = o_p(1). \tag{D.12}$$

Now, yet another Taylor expansion w.r.t. η yields

$$n^{-1/2}\frac{d}{d\beta}[L_n(\beta, \hat{\eta}_\beta) - L_n(\beta, \eta_\beta)] = n^{-1/2}\frac{d}{d\beta}\left[\frac{\partial L_n(\beta, \eta_\beta)}{\partial\eta}(\hat{\eta}_\beta - \eta_\beta) + r_n^{(2)}(\beta)\right],$$

with $r_n^{(2)}(\beta)$ defined in Lemma D.2(ii). Hence, using Lemma D.1 and Lemma D.2(ii),

$$\begin{aligned}n^{-1/2}\frac{dL_n(\beta, \hat{\eta}_{\beta_0})}{d\beta} - n^{-1/2}\frac{dL_n(\beta, \eta_{\beta_0})}{d\beta} &= n^{-1/2}\left[\frac{d}{d\beta}\frac{\partial L_n(\beta_0, \eta_{\beta_0})}{\partial\eta}\right](\hat{\eta}_{\beta_0} - \eta_{\beta_0}) \\ &\quad + n^{-1/2}\frac{\partial L_n(\beta_0, \eta_{\beta_0})}{\partial\eta}\left(\frac{d}{d\beta}\hat{\eta}_{\beta_0} - \frac{d}{d\beta}\eta_{\beta_0}\right) \\ &\quad + \frac{d}{d\beta}r_n^{(2)}(\beta_0) \\ &= o_p(1),\end{aligned}$$

which implies that

$$n^{-1/2}\frac{dL_n(\beta_0,\hat{\eta}_{\beta_0})}{d\beta} = n^{-1/2}\frac{dL_n(\beta_0,\eta_{\beta_0})}{d\beta} + o_p(1). \tag{D.13}$$

Therefore, (D.12) and (D.13) imply

$$n^{1/2}(\hat{\beta}_n - \beta_0) = \left[-\frac{1}{n}\frac{d^2L_n(\beta_0,\eta_{\beta_0})}{d\beta d\beta'}\right]^{-1} n^{-1/2}\frac{dL_n(\beta_0,\eta_{\beta_0})}{d\beta} + o_p(1).$$

But since η_β is a least favorable curve, $-\frac{1}{n}\frac{d^2L_n(\beta_0,\eta_{\beta_0})}{d\beta d\beta'} \xrightarrow{p} I_{\beta_0}$, and by Slutsky's Lemma

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2}I_{\beta_0}^{-1}\frac{dL_n(\beta_0,\eta_{\beta_0})}{d\beta} + o_p(1). \tag{D.14}$$

Now by the CLT,

$$n^{-1/2}\frac{dL_n(\beta_0,\eta_{\beta_0})}{d\beta} \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, I_{\beta_0}^{-1}),$$

and thus $n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, I_{\beta_0}^{-1})$. □

PROOF. **[Theorem 3.5.3]** Using the asymptotic linearity of $\hat{\beta}_n$ from (D.14),

$$n^{1/2}(\hat{\beta}_n - \beta_0) = n^{-1/2}\sum_{i=1}^{n} I_{\beta_0}^{-1}\frac{d}{d\beta}\ell_n(\beta_0,\eta_{\beta_0};\mathbf{x}_i,y_i,\mathbf{z}_i) + o_p(1), \quad \text{and}$$
$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow[\beta_0]{d} \mathrm{N}(\mathbf{0}, I_{\beta_0}^{-1}).$$

Similarly, the LAN condition implies that for any $\delta \in \mathbb{R}^p$

$$\mathcal{L}_n = n^{-1/2}\sum_{i=1}^{n}\delta'\frac{d}{d\beta}\ell_n(\beta_0,\eta_{\beta_0};\mathbf{x}_i,y_i,\mathbf{z}_i) - \frac{1}{2}\delta' I_{\beta_0}\delta + o_p(1), \quad \text{and}$$
$$\mathcal{L}_n \xrightarrow[\beta_0]{d} \mathrm{N}(-\frac{1}{2}\delta' I_{\beta_0}\delta, \delta' I_{\beta_0}\delta).$$

Hence by the Cramér - Wold device,

$$\begin{pmatrix} n^{1/2}(\hat{\beta}_n - \beta_0) \\ \mathcal{L}_n \end{pmatrix} \xrightarrow[\beta_0]{\mathrm{d}} \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ -\frac{1}{2}\delta' I_{\beta_0}\delta \end{bmatrix}, \begin{bmatrix} I_{\beta_0}^{-1} & \delta \\ \delta' & \delta' I_{\beta_0}\delta \end{bmatrix}\right).$$

Now let $\boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + n^{-1/2}\boldsymbol{\delta}$. Then using LeCam's Third Lemma,

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow[\boldsymbol{\beta}_n]{d} \mathrm{N}(\boldsymbol{\delta}, I_{\boldsymbol{\beta}_0}^{-1}).$$

But this implies that

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) \xrightarrow[\boldsymbol{\beta}_n]{d} \mathrm{N}(\mathbf{0}, I_{\boldsymbol{\beta}_0}^{-1}),$$

and since the limiting distribution does not depend upon $\boldsymbol{\delta}$, $\hat{\boldsymbol{\beta}}_n$ is regular. □

APPENDIX E

PROOFS OF RESULTS IN SECTION 3.6

PROOF. **[Theorem 3.6.1]** Let $\lambda > 0$. Since $\delta^*(\lambda u, \lambda v) = \frac{(\lambda v)^r \mathbb{E}[x_i z_2^r | \frac{\lambda u}{\lambda v}]}{\mathbb{E}[z_2^{2r} | \frac{\lambda u}{\lambda v}]} = \lambda^r \delta^*(u, v)$, $\delta^*(\cdot)$ is homogeneous of degree r. To verify the orthogonality conditions, let $g(\cdot)$ be a homogeneous function of degree r. Then since $g(z_1, z_2) = z_2^r g(\frac{z_1}{z_2}, 1)$,

$$\begin{aligned}
\mathbb{E}[(x_i - \delta^*(z_1, z_2))g(z_1, z_2)] &= \mathbb{E}[\{x_i - \frac{z_2^r \mathbb{E}(x_i z_2^r | \frac{z_1}{z_2})}{\mathbb{E}(z_2^{2r} | \frac{z_1}{z_2})}\} z_2^r g(\frac{z_1}{z_2}, 1)] \\
&= \mathbb{E}[x_i g(z_1, z_2) - z_2^{2r} \frac{\mathbb{E}(z_2^r x_i g(\frac{z_1}{z_2}, 1) | \frac{z_1}{z_2})}{\mathbb{E}(z_2^{2r} | \frac{z_1}{z_2})}] \\
&= \mathbb{E}[x_i g(z_1, z_2) - z_2^{2r} \frac{\mathbb{E}(x_i g(z_1, z_2) | \frac{z_1}{z_2})}{\mathbb{E}(z_2^{2r} | \frac{z_1}{z_2})}] \\
&= \mathbb{E}[x_i g(z_1, z_2)] - \mathbb{E}[z_2^{2r} \frac{\mathbb{E}(x_i g(z_1, z_2) | \frac{z_1}{z_2})}{\mathbb{E}(z_2^{2r} | \frac{z_1}{z_2})}] \\
&= \mathbb{E}[x_i g(z_1, z_2)] - \mathbb{E}[\mathbb{E}\{z_2^{2r} \frac{\mathbb{E}(x_i g(z_1, z_2) | \frac{z_1}{z_2})}{\mathbb{E}(z_2^{2r} | \frac{z_1}{z_2})} \bigg| \frac{z_1}{z_2}\}] \\
&= \mathbb{E}[x_i g(z_1, z_2)] - \mathbb{E}[\frac{\mathbb{E}(x_i g(z_1, z_2) | \frac{z_1}{z_2})}{\mathbb{E}(z_2^{2r} | \frac{z_1}{z_2})} \mathbb{E}\left\{z_2^{2r} | \frac{z_1}{z_2}\right\}] \\
&= \mathbb{E}[x_i g(z_1, z_2)] - \mathbb{E}[\mathbb{E}\left\{x_i g(z_1, z_2) | \frac{z_1}{z_2}\right\}] \\
&= \mathbb{E}[x_i g(z_1, z_2)] - \mathbb{E}[x_i g(z_1, z_2)] = 0.
\end{aligned}$$

Thus, the necessary and sufficient conditions for $\delta^*(z_1, z_2)$ to be the required projection are satisfied. □

Proof. **[Proposition 3.6.1]** To see that $\eta_\beta(u,v)$ is a least favorable curve, first notice that for $i = 1, \dots, p$

$$\frac{d}{d\beta_i}\eta_{\beta_0}(u,v) = -\frac{v^r\,\mathbb{E}\left[x_{ij}z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]},$$

which by Theorem 3.6.1 is the least favorable direction. Hence, the curve η_{β_0} certainly has the least favorable direction as the tangent vector. Now η_β is clearly a homogeneous function of degree r, i.e. $\eta_\beta \in \mathcal{F}$. So to verify that η_β is a least favorable curve, all we have to do is to show that $\eta_{\beta_0} = f^*$. To see this, notice that at $\boldsymbol{\beta}_0$,

$$\begin{aligned}
\frac{v^r\,\mathbb{E}\left[y_j z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} &= \frac{v^r\,\mathbb{E}\left[(\sum_{i=1}^p x_{ij}\beta_{i0} + f^*(z_{1j}, z_{2j}))z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} \\
&= \sum_{i=1}^p \beta_{i0} \frac{v^r\mathbb{E}\left[x_{ij}z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} \\
&\quad + \frac{v^r\,\mathbb{E}\left[f^*(z_{1j}, z_{2j})z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} \\
&= \sum_{i=1}^p \beta_{i0} \frac{v^r\mathbb{E}\left[x_{ij}z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} \\
&\quad + \frac{v^r\,\mathbb{E}\left[f^*(\frac{z_{1j}}{z_{2j}}, 1)z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} \\
&= \sum_{i=1}^p \beta_{i0} \frac{v^r\mathbb{E}\left[x_{ij}z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} + v^r f^*(\frac{u}{v}, 1) \\
&= \sum_{i=1}^p \beta_{i0} \frac{v^r\mathbb{E}\left[x_{ij}z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} + f^*(u,v).
\end{aligned}$$

Hence,

$$\begin{aligned} f^*(u,v) &= \frac{v^r\,\mathbb{E}\left[y_j z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} - \sum_{i=1}^{p} \beta_{i0} \frac{v^r\mathbb{E}\left[x_{ij} z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} \\ &= \eta_{\beta_0}(u,v). \end{aligned}$$

Therefore, η_β as defined in (3.6.1) is indeed a least favorable curve. To see that $\hat{\eta}_\beta$ consistently estimates η_β, notice that from standard results on kernel estimation

$$\frac{v^r \sum_{j=1}^{n} y_j z_{2j}^r \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])}{\sum_{j=1}^{n} z_{2j}^{2r} \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])} \xrightarrow{p} \frac{v^r\,\mathbb{E}\left[y_j z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}, \quad \text{and} \tag{E.1}$$

$$\frac{v^r \sum_{j=1}^{n} x_{ij} z_{2j}^r \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])}{\sum_{j=1}^{n} z_{2j}^{2r} \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])} \xrightarrow{p} \frac{v^r\mathbb{E}\left[x_{ij} z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}. \tag{E.2}$$

Hence from (E.1) and (E.2)

$$\begin{gathered} \frac{v^r \sum_{j=1}^{n} y_j z_{2j}^r \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])}{\sum_{j=1}^{n} z_{2j}^{2r} \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])} - \sum_{i=1}^{p} \beta_i \frac{v^r \sum_{j=1}^{n} x_{ij} z_{2j}^r \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])}{\sum_{j=1}^{n} z_{2j}^{2r} \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])} \\ \xrightarrow{p} \\ \frac{v^r\,\mathbb{E}\left[y_j z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]} - \sum_{i=1}^{p} \beta_i \frac{v^r\mathbb{E}\left[x_{ij} z_{2j}^r \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}{\mathbb{E}\left[z_{2j}^{2r} \middle| \frac{z_{1j}}{z_{2j}} = \frac{u}{v}\right]}, \end{gathered}$$

which implies that $\hat{\eta}_\beta \xrightarrow{p} \eta_\beta$. Therefore, $\hat{\eta}_\beta$ is a consistent estimator of η_β. □

APPENDIX F

PROOF OF THEOREM 3.6.2

To prove Theorem 3.6.2 we will closely follow Ichimura (1993). The proof will be obtained as a series of lemmas and propositions. Because of its length, we will skip most of the tedious algebra. Furthermore, we will only prove (1). The proof of (2), (1′), and (2′) is similar. In what follows, the limits of integration are always $(-\infty, \infty)$ unless otherwise specified.

Notation F.0.1. Let,

$$\begin{aligned}
\hat{A}_n(u,v) &= \frac{v^r}{na_n}\sum_{j=1}^{n} y_j z_{2j}^r \mathrm{K}\left(\frac{1}{a_n}\left[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}\right]\right) \\
\hat{B}_n(u,v) &= \frac{1}{na_n}\sum_{j=1}^{n} z_{2j}^{2r} \mathrm{K}\left(\frac{1}{a_n}\left[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}\right]\right) \\
t &= u/v \\
t_j &= z_{1j}/z_{2j} \\
g_{nj}(u,v) &= v^r z_{2j}^r \mathrm{K}\left(\frac{1}{a_n}\left[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}\right]\right).
\end{aligned}$$

We will also make use of the following inequality, given in Ichimura (1993), to majorize various terms.

Theorem F.1 (Bernstein's Inequality). *Let $Y_{1n}, \ldots, Y_{nn}$ be independent random* *variables with zero means and bounded ranges, that is, $|Y_{in}| \leq c_n$. Write σ^2_{in} for the variance of Y_{in}. Suppose $V_n \geq \sigma^2_{1n} + \ldots + \sigma^2_{nn}$. Then for each $\xi_n > 0$,*

$$\Pr\{|Y_{1n} + \ldots + Y_{nn}| > \xi_n\} \leq \exp\left\{\frac{-\xi_n^2}{2(V_n + \frac{1}{3}c_n\xi_n)}\right\}. \quad \square$$

Now let $\hat{A}_n(u,v) \xrightarrow{p} A(u,v)$. Clearly, $A(u,v) = v^r \mathbb{E}[y_j z^r_{2j}|t_j = t]p(t)$. Also,

$$|\hat{A}_n(u,v) - A(u,v)| \leq |\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)| + |\mathbb{E}\,\hat{A}_n(u,v) - A(u,v)|.$$

We first determine the rate at which $\hat{A}_n(u,v)$ converges to its probability limit. This will be done with the help of the following lemmas.

Lemma F.1. $\sup_{u,v} |\mathbb{E}\,\hat{A}_n(u,v) - A(u,v)| = O(a_n^2)$.

Proof. Since

$$\begin{aligned}
\mathbb{E}\,\hat{A}_n(u,v) &= \frac{v^r}{na_n}\sum_{j=1}^{n} \mathbb{E}\, y_j z^r_{2j} \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}]) \\
&= \frac{v^r}{a_n}\int \mathrm{K}(\frac{1}{a_n}[t - t_j])\mathbb{E}[y_j z^r_{2j}|t_j]p(t_j)\,dt_j \\
&= v^r \int \mathrm{K}(s)\mathbb{E}[y_j z^r_{2j}|t - a_n s]p(t - a_n s)\,ds.
\end{aligned}$$

For some t^* between $t - a_n s$ and t,

$$\begin{aligned}
|\mathbb{E}\,\hat{A}_n(u,v) - A(u,v)| &= |v^r \int \mathrm{K}(s)[\psi(t - a_n s) - \psi(t)]\,ds| \\
&= |v^r|a_n^2/2| \int s^2 \mathrm{K}(s)\psi''(t^*)\,ds| \\
&\leq a_n^2 M
\end{aligned}$$

where,

$$M = \sup_t |\psi''(t)| \sup_v |v^r| \int s^2 \mathrm{K}(s)\,ds.$$

Hence, $\sup_{u,v} |\mathbb{E}\,\hat{A}_n(u,v) - A(u,v)| = O(a_n^2)$. $\square$

We now look at the term $|\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)|$. Since the dependent variable y may be unbounded, a truncation procedure is used to deal with this term. So let $\mathbb{I}_{nj} = \mathbb{I}_{\{|y_j| \leq M_n\}}$ and $\mathbb{I}^c_{nj} = 1 - \mathbb{I}_{nj}$, where M_n is a sequence of positive numbers chosen such that $M_n \to \infty$ as $n \to \infty$. Then,

$$\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v) \equiv (\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v))\mathbb{I} + (\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v))\mathbb{I}^c,$$

where,

$$(\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v))\mathbb{I} \equiv \frac{v^r}{na_n}\sum_{j=1}^{n}[y_j\mathbb{I}_{nj}\mathrm{K}(\frac{1}{a_n}[t-t_j]) - \mathbb{E}\,y_j\mathbb{I}_{nj}\mathrm{K}(\frac{1}{a_n}[t-t_j])]$$
$$(\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v))\mathbb{I}^c \equiv \frac{v^r}{na_n}\sum_{j=1}^{n}[y_j\mathbb{I}^c_{nj}\mathrm{K}(\frac{1}{a_n}[t-t_j]) - \mathbb{E}\,y_j\mathbb{I}^c_{nj}\mathrm{K}(\frac{1}{a_n}[t-t_j])].$$

Therefore,

$$\Pr\{\sup_{u,v}|\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)| > 2\epsilon\} \leq \Pr\{\sup_{u,v}|\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)|\mathbb{I}^c > \epsilon\}$$
$$+ \Pr\{\sup_{u,v}|\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)|\mathbb{I} > \epsilon\}.$$

LEMMA F.2. *Let* $c_0 = 2\mathbb{E}\,|y_j|^q sup_{u,v}|g_{nj}(u,v)|$. *Then,*

$$\Pr\{\sup_{u,v}|\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)|\mathbb{I}^c > \epsilon\} \leq \frac{c_0}{a_n\epsilon M_n^{q-1}}.$$

PROOF. By an application of Hölder and Chebychev inequalities,

$$\Pr\{\sup_{u,v}|\hat{A}_n - \mathbb{E}\,\hat{A}_n|\mathbb{I}^c > \epsilon\} = \Pr\{\sup_{u,v}|\sum_{j=1}^{n}[y_j\mathbb{I}^c_{nj}g_{nj}(u,v) - \mathbb{E}\,y_j\mathbb{I}^c_{nj}g_{nj}(u,v)]| > na_n\epsilon\}$$
$$\leq \Pr\{\sum_{j=1}^{n}\sup_{u,v}|[y_j\mathbb{I}^c_{nj}g_{nj}(u,v) - \mathbb{E}\,y_j\mathbb{I}^c_{nj}g_{nj}(u,v)]| > na_n\epsilon\}$$
$$\leq c\frac{\mathbb{E}\,|y_j|^q}{a_n\epsilon M_n^{q-1}},$$

where $c = 2\sup_{u,v}|g_{nj}(u,v)|$. □

Now since $Z_1 \times Z_2$ is compact by assumption, w.l.o.g. assume that it lies inside the unit cube in $\mathbb{R}^2$. Also let δ and ν denote some positive numbers in this proof that are chosen appropriately. What values of δ, ν to choose will be determined according to the requirements of the subsequent lemmas. For $i = 1, 2$, partition Z_i into N_i cubes having each side of length δa_n^{ν}. Therefore, $Z_1 \times Z_2$ is partitioned into $N = N_1 \times N_2$ cubes with $N = \delta^{-2} a_n^{-2\nu}$. That is, $Z_1 \times Z_2 \subset \cup_{i=1}^{N} \mathrm{S}_i$, where S_i is an open cube in $\mathbb{R}^2$ with each side of length $\delta^2 a_n^{2\nu}$, and $N = \delta^{-2} a_n^{-2\nu}$. Furthermore, for each S_i let $(u_i, v_i) \in \mathrm{S}_i$ with $i = 1, \dots, N$. Then,

$$\begin{aligned}
\Pr\{\sup_{u,v} |\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)|\mathbb{I} > \epsilon\} &\leq \Pr\{\cup_{i=1}^{N} \sup_{\mathrm{S}_i} |\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)|\mathbb{I} > \epsilon\} \\
&\leq \sum_{i=1}^{N} \Pr\{\sup_{\mathrm{S}_i} |\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)|\mathbb{I} > \epsilon\} \\
&= T1 + T2 + T3, \quad \text{where,} \qquad \text{(F.1)}
\end{aligned}$$

$$\begin{aligned}
T1 &\equiv \sum_{i=1}^{N} \Pr\{|\frac{1}{a_n} \sum_{j=1}^{n} [y_j \mathbb{I}_{nj} g_{nj}(u,v) - \mathbb{E}\, y_j \mathbb{I}_{nj} g_{nj}(u,v)]| > \frac{n\epsilon}{2}\} \\
T2 &\equiv \sum_{i=1}^{N} \Pr\{\sup_{\mathrm{S}_i} |\frac{1}{a_n} \sum_{j=1}^{n} [y_j \mathbb{I}_{nj} g_{nj}(u,v) - y_j \mathbb{I}_{nj} g_{nj}(u_i,v_i)]| > \frac{n\epsilon}{4}\} \\
T3 &\equiv \sum_{i=1}^{N} \Pr\{\sup_{\mathrm{S}_i} |\frac{1}{a_n} \sum_{j=1}^{n} [\mathbb{E}\, y_j \mathbb{I}_{nj} g_{nj}(u,v) - \mathbb{E}\, y_j \mathbb{I}_{nj} g_{nj}(u_i,v_i)]| > \frac{n\epsilon}{4}\}.
\end{aligned}$$

LEMMA F.3. *Let T1 be defined as above. Then*

$$T1 \leq \delta^{-2} \exp\{-2\nu \log a_n - \frac{c_3 n a_n \epsilon^2}{1 + M_n \epsilon}\},$$

where, $c_3^{-1} = 16 \sup_v v^{2r} \sup_t \mathbb{E}\,[(y_j \mathbb{I}_{nj} z_{2j}^r)^2|t] \sup_t p(t)$.

PROOF. As in Ichimura (1993), apply Bernstein's Inequality with

$$\xi_n = \frac{n a_n \epsilon}{2}, \quad c_n = c_1 M_n, \quad V_n = c_2 n a_n$$

where,

$$c_1 = \sup_{u,v} |g_{nj}(u,v)|$$
$$c_2 = 2[\sup_{u,v} v^{2r}]\,[\sup_t \mathbb{E}\,[(y_j \mathbb{I}_{nj} z_{2j}^r)^2|t]]\,[\sup_t p(t)].$$

□

Lemma F.4. *With $T2$ as defined before,* $T2 \le \delta^{-2}\exp\{-2\nu \log a_n - \frac{c_3 n a_n \epsilon^2}{1+M_n\epsilon}\} + o(1)$.

Proof. First notice that

$$\Pr\{\sup_{S_i} |\frac{1}{a_n}\sum_{j=1}^{n}[y_j\mathbb{I}_{nj}g_{nj}(u,v) - y_j\mathbb{I}_{nj}g_{nj}(u_i,v_i)]| > \frac{n\epsilon}{4}\} = T2_A + T2_B,$$

where,

$$T2_A \equiv \Pr\{|\sum_{j=1}^{n}\frac{1}{a_n}[\sup_{S_i}|y_j\mathbb{I}_{nj}g_{nj}(u,v) - y_j\mathbb{I}_{nj}g_{nj}(u_i,v_i)|$$
$$- \mathbb{E}\sup_{S_i}|y_j\mathbb{I}_{nj}g_{nj}(u,v) - y_j\mathbb{I}_{nj}g_{nj}(u_i,v_i)|]| > \frac{n\epsilon}{8}\}$$
$$T2_B \equiv \Pr\{\frac{1}{a_n}\mathbb{E}\sup_{S_i}|y_j\mathbb{I}_{nj}g_{nj}(u,v) - y_j\mathbb{I}_{nj}g_{nj}(u_i,v_i)| > \frac{\epsilon}{8}\}.$$

Once again, apply Bernstein's inequality with

$$\xi_n = \frac{n a_n \epsilon}{8}, \quad c_n = c_6 M_n a_n^{\nu-1}, \quad V_n = c_7 n a_n^{2(\nu-1)}$$

where $c_6 = 2\delta c_4$, $c_7 = \delta^2 c_4^2$, and

$$c_4 = \sup_{u,v\in S_i} |v^{r-1} z_{2j}^r \mathrm{K}'(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])| + \sup_{u,v\in S_i} |r v^{r-1} z_{2j}^r \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])|$$
$$+ \sup_{u,v\in S_i} |\frac{u}{v^{2-r}} z_{2j}^r \mathrm{K}'(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])|$$

to get,

$$T2_A \le \exp\{-\frac{n a_n \epsilon^2 c_8}{a_n^{2\nu-3} + M_n a_n^{\nu-1}\epsilon}\},$$

with $c_8^{-1} = 128c_7$.

To handle $T2_B$, apply Chebychev and Hölder inequalities to get

$$T2_B \le \frac{8(\mathbb{E}\,|y_j\mathbb{I}_{nj}|^2)^{1/2}(\mathbb{E}\,[\sup_{\mathrm{S}_i}|g_{nj}(u,v) - g_{nj}(u_i,v_i)|]^2)^{1/2}}{\epsilon a_n}.$$

Now $q \ge 2$ implies that $(\mathbb{E}\,|y_j\mathbb{I}_{nj}|^2)^{1/2} < \infty$. Moreover, uniform continuity of $g_{nj}(u,v)$ on $\overline{\mathrm{S}}_i$ allows us to conclude that $T2_B \le \frac{c_{10}a_n^{\nu-1}}{N\epsilon}$, where $c_{10} = \delta[\mathbb{E}\,|y_j\mathbb{I}_{nj}|^2]^{1/2}$. Therefore,

$$\begin{aligned} T2 &\le \sum_{i=1}^{N} T2_A + \sum_{i=1}^{N} T2_B \\ &\le N\exp\{-\frac{na_n\epsilon^2c_8}{a_n^{2\nu-3} + M_na_n^{\nu-1}\epsilon}\} + \frac{c_{10}a_n^{\nu-1}}{\epsilon}. \end{aligned}$$

But notice that for $\nu > 1$, $a_n^{\nu-1} \to 0$. Hence by choosing $\nu > 1$, $\frac{c_{10}a_n^{\nu-1}}{\epsilon} \to 0$, and

$$T2 \le N\exp\{-\frac{na_n\epsilon^2c_8}{a_n^{2\nu-3} + M_na_n^{\nu-1}\epsilon}\} + o(1).$$

Furthermore, it may also be shown that if $\nu > 1.5$ and n is sufficiently large,

$$N\exp\{-\frac{na_n\epsilon^2c_8}{a_n^{2\nu-3} + M_na_n^{\nu-1}\epsilon}\} \le \delta^{-2}\exp\{-2\nu\log a_n - \frac{c_3na_n\epsilon^2}{1 + M_n\epsilon}\}.$$

Hence, $T2 \le \delta^{-2}\exp\{-2\nu\log a_n - \frac{c_3na_n\epsilon^2}{1+M_n\epsilon}\} + o(1)$. □

LEMMA F.5. *With $T3$ as defined before, $T3 = o(1)$.*

PROOF. Notice that by the previous result,

$$\begin{aligned} T3 &= \sum_{i=1}^{N} \Pr\{\sup_{\mathrm{S}_i}|\frac{1}{a_n}\sum_{j=1}^{n}[\mathbb{E}\,y_j\mathbb{I}_{nj}g_{nj}(u,v) - \mathbb{E}\,y_j\mathbb{I}_{nj}g_{nj}(u_i,v_i)]| > \frac{n\epsilon}{4}\} \\ &\le \sum_{i=1}^{N} \Pr\{\mathbb{E}\,\sup_{\mathrm{S}_i}|y_j\mathbb{I}_{nj}g_{nj}(u,v) - y_j\mathbb{I}_{nj}g_{nj}(u_i,v_i)| > \frac{a_n\epsilon}{4}\} \\ &\le \sum_{i=1}^{N} T2_B \\ &\le o(1). \end{aligned}$$

□

Proposition F.1. $|\hat{A}_n(u,v) - A(u,v)| = O_p(n^{-\lambda})$, *if*

$$\frac{\lambda}{2} \le \alpha < \frac{(1-2\lambda)(q-1)}{q}.$$

Proof. Since we have obtained the terms that majorize $T1$,$T2$ and $T3$, by (F.1) we have

$$\begin{aligned}\Pr\{\sup_{\mathrm{S},} |\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)|\mathbb{I} > \epsilon\} &= T1 + T2 + T3 \\ &\le 2\delta^{-2}\exp\{-2\nu \log a_n - \frac{c_3 n a_n \epsilon^2}{1 + M_n\epsilon}\} + o(1).\end{aligned}$$

This inequality, combined with the previously obtained result

$$\Pr\{\sup_{u,v} |\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)|\mathbb{I}^c > \epsilon\} \le \frac{c_0}{a_n \epsilon M_n^{q-1}},$$

leads to the conclusion that

$$\begin{aligned}\Pr\{\sup_{u,v} |\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)| > 2\epsilon\} &\le \frac{c_0}{a_n \epsilon M_n^{q-1}} \\ &+ 2\delta^{-2}\exp\{-2\nu \log a_n - \frac{c_3 n a_n \epsilon^2}{1 + M_n\epsilon}\} + o(1).\end{aligned}$$

Now in the above inequality, replace ϵ by $n^{-\lambda}\epsilon_0$, a_n by $n^{-\alpha}$, and choose

$$M_n = \frac{n a_n}{(-\log a_n)} \left[\frac{(-\log a_n)}{n a_n^{q/(q-1)}}\right]^{1/2}.$$

Then after some tedious algebra we can show that

$$\Pr\{\sup_{u,v} |\hat{A}_n(u,v) - \mathbb{E}\,\hat{A}_n(u,v)| > 2n^{-\lambda}\epsilon_0\} \to 0,$$

if,

(i) $\alpha < \min\{1, \frac{q-1-2\lambda}{q}, 1 - 2\lambda, \frac{(1-2\lambda)(q-1)}{q}\}$, and

(ii) ν is chosen such that $\nu > \max\{1.5, 1 + \lambda/\alpha\}$.

Therefore, combining this result with the fact that

$$\sup_{u,v} |\mathbb{E}\hat{A}_n(u,v) - A(u,v)| = O(a_n^2) = O(n^{-2\alpha}),$$

we get that

$$\sup_{u,v} |\hat{A}_n(u,v) - A(u,v)| = O(\max\{n^{-\lambda}, n^{-2\alpha}\}).$$

Hence $\sup_{u,v} |\hat{A}_n(u,v) - A(u,v)| = O_p(n^{-\lambda})$, if

$$\frac{\lambda}{2} \leq \alpha < \min\{1, \frac{q-1-2\lambda}{q}, 1-2\lambda, \frac{(1-2\lambda)(q-1)}{q}\}.$$

But since,

$$1 > 1-2\lambda > \frac{(1-2\lambda)(1-q)}{q},$$

and

$$\frac{q-1-2\lambda}{q} > \frac{(1-2\lambda)(q-1)}{q}$$

for $q > 2$, the condition on α simplifies to

$$\frac{\lambda}{2} \leq \alpha < \frac{(1-2\lambda)(q-1)}{q}.$$

□

Remark F.1. This proposition shows that ν should be chosen such that

$$\nu > \max\{1.5, 1 + \lambda/\alpha\}$$

while constructing the cubes S_i. Clearly, this value of ν satisfies the requirements of Lemma F.4. □

Proposition F.2. *Let* $\hat{B}_n(u,v) \xrightarrow{p} B(u,v)$. *Then,*

$$\sup_{u,v} |\hat{B}_n(u,v) - B(u,v)| = O_p(n^{-\lambda})$$

if,

$$\frac{\lambda}{2} \leq \alpha < \frac{(1-2\lambda)(q-1)}{q}.$$

Proof. Since $\hat{B}_n(u,v)$ is obtained from $\hat{A}_n(u,v)$ by putting $v^r y_j = 1$, the result of the previous proposition applies. □

We are now in a position to finally prove Result (1) of Theorem 3.6.2.

Proof. **[Theorem 3.6.2]** First notice that

$$\hat{\eta}_\beta(u,v) = \frac{\hat{A}_n(u,v)}{\hat{B}_n(u,v)} - \sum_{i=1}^{p} \beta_i \frac{\hat{C}_{ni}(u,v)}{\hat{B}_n(u,v)}$$

where,

$$\hat{C}_{ni}(u,v) = \frac{\sum_{j=1}^n x_{ij} z_{2j}^r \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])}{\sum_{j=1}^n z_{2j}^r \mathrm{K}(\frac{1}{a_n}[\frac{u}{v} - \frac{z_{1j}}{z_{2j}}])}$$

$$\eta_\beta(u,v) = \frac{A(u,v)}{B(u,v)} - \sum_{i=1}^{p} \beta_i \frac{C_i(u,v)}{B(u,v)},$$

and $\hat{C}_{ni}(u,v) \xrightarrow{P} C_i(u,v)$. Now,

$$\hat{\eta}_\beta(u,v) - \eta_\beta(u,v) = \left[\frac{\hat{A}_n(u,v)}{\hat{B}_n(u,v)} - \frac{A(u,v)}{B(u,v)}\right] - \sum_{j=1}^{n} \beta_i \left[\frac{\hat{C}_{ni}(u,v)}{\hat{B}_n(u,v)} - \frac{C_i(u,v)}{B(u,v)}\right].$$

Now, since $\inf_{u,v} |B(u,v)| > 0$, $0 < \sup_{u,v} |A(u,v)| < \infty$, and $0 < \sup_{u,v} |C_i(u,v)| < \infty$ by assumption, Proposition F.1 and Proposition F.2 readily imply that

$$\sup_{u,v} \left| \frac{\hat{A}_n(u,v)}{\hat{B}_n(u,v)} - \frac{A(u,v)}{B(u,v)} \right| = o_p(n^{-\lambda}),$$

$$\sup_{u,v} \left| \frac{\hat{C}_{ni}(u,v)}{\hat{B}_n(u,v)} - \frac{C_i(u,v)}{B(u,v)} \right| = o_p(n^{-\lambda}).$$

And since each β_i is bounded,

$$\begin{aligned}\sup_{u,v,\beta}|\hat{\eta}_{\beta}(u,v)-\eta_{\beta}(u,v)| &\le \sup_{u,v}\left|\frac{\hat{A}_n(u,v)}{\hat{B}_n(u,v)}-\frac{A(u,v)}{B(u,v)}\right| \\ &\quad+\sum_{i=1}^{p}\sup_{u,v}\left|\frac{\hat{C}_{ni}(u,v)}{\hat{B}_n(u,v)}-\frac{C_i(u,v)}{B(u,v)}\right|\sup_{\beta}|\beta_i| \\ &\le \sup_{u,v}\left|\frac{\hat{A}_n(u,v)}{\hat{B}_n(u,v)}-\frac{A(u,v)}{B(u,v)}\right| \\ &\quad+M\sum_{i=1}^{p}\sup_{u,v}\left|\frac{\hat{C}_{ni}(u,v)}{\hat{B}_n(u,v)}-\frac{C_i(u,v)}{B(u,v)}\right|\sup_{\beta}|\beta_i| \\ &= o_p(n^{-\lambda}).\end{aligned}$$

Hence, $\sup_{u,v,\beta}|\hat{\eta}_{\beta}(u,v)-\eta_{\beta}(u,v)| = o_p(n^{-\lambda})$, which certainly implies that

$$\sup_{u,v}|\hat{\eta}_{\beta_0}(u,v)-\eta_{\beta_0}(u,v)| = o_p(n^{-\lambda}).$$

□

APPENDIX G

PROOFS OF RESULTS IN SECTION 3.7

REMARK G.1. In this section, all convergence is w.r.t. the C^2 norm. □

PROOF. **[Theorem 3.7.1]** Using Theorem A.2 we have to show that $\overline{\mathcal{F} - \mathbb{R}_+ f^*} = T(\mathcal{F}, f^*)$. We proceed case by case.

Case I: f^* is strictly concave on **Z**:

If f^* is strictly concave, $f^* \in \text{int}(\mathcal{F})$ which implies that $T(\mathcal{F}, f^*) = \mathcal{H}$.

Case II: f^* is affine on **Z**:

$\Longrightarrow$ Let $f \in \overline{\mathcal{F} - \mathbb{R}_+ f^*}$. Then there exists a sequence $(\lambda_n, f_n) \in (0, \infty) \times \mathcal{F}$, such that $f_n - \lambda_n f^* \rightarrow f$. But since f^* is linear, $f_n - \lambda_n f^*$ is a convergent sequence of concave functions. Hence the limit f is also a concave function, i.e. $f \in \mathcal{F}$. This shows that $\overline{\mathcal{F} - \mathbb{R}_{++} f^*} \subseteq \mathcal{F}$.

$\Longleftarrow$ To show the reverse inequality, let $\delta \in \mathcal{F}$. Now for $t > 0$, define $\eta(u) = f^*(u) + t\delta(u)$. Then, for all $u \in \mathbf{Z}$ and $\alpha \in \mathbb{R}^2$,

$$\begin{aligned} \alpha'[\nabla^2(\eta(u) - t\delta(u))]\alpha] &= \alpha'[\nabla^2 f^*(u)]\alpha \\ &= 0, \end{aligned}$$

since f^* is affine on **Z**. This implies that $\eta - t\delta \in \mathcal{F}$, which further implies that $\eta \in \mathcal{F}$.

Therefore, $\delta \in \mathcal{F} - \frac{1}{t}f^* \subseteq \overline{\mathcal{F} - \mathbb{R}_+ f^*}$.

Case III: f^* is concave (but not strictly concave) on $\mathbf{Z}$:

Since f^* is concave but not strictly concave, there exists a nonempty set $\mathbf{Z}_0 \subset \mathbf{Z}$ such that, $\det[\nabla^2 f^*(u)] = 0$ for $u \in \mathbf{Z}_0$, while for $u \in \mathbf{Z} - \mathbf{Z}_0$ the Hessian matrix $\nabla^2 f^*(u)$ is negative definite.

We first show that $\overline{\mathcal{F} - \mathbb{R}_{++} f^*} \subseteq \mathcal{W}$. So let $f \in \overline{\mathcal{F} - \mathbb{R}_{++} f^*}$. Then there exists a sequence $(\lambda_n, f_n) \in (0, \infty) \times \mathcal{F}$, such that $f_n - \lambda_n f^* \to f$. Let $g_n(u) = f_n(u) - \lambda_n f^*(u)$. Now notice that on $\mathbf{Z}_0$, $\det[\nabla^2 g_n - \nabla^2 f_n] = \lambda_n^2 \det[\nabla^2 f^*] = 0$. i.e. the determinant of the Hessian of $g_n - f_n$ vanishes on $\mathbf{Z}_0$. But this implies that on $\mathbf{Z}_0$, $g_n - f_n \in \mathcal{W}$, i.e. $g_n \in W + f_n$. However, $f_n \in \mathcal{F} \subset \mathcal{W}$ implying that $g_n \in \mathcal{W}$ since $\mathcal{W}$ is a convex cone. But this says that g_n is a convergent sequence in the closed cone $\mathcal{W}$. Therefore, its limit f also lies in $\mathcal{W}$. That is, $\overline{\mathcal{F} - \mathbb{R}_{++} f^*} \subseteq \mathcal{W}$.

Now for the reverse inequality. Let δ be any function in $\mathcal{W}$, and for $t > 0$ define $\eta_t(u) = f^*(u) + t\delta(u)$. Then if we can show that $\eta_t(u) \in \mathcal{F}$ for all $u \in \mathbf{Z}$, we are done because then $\delta = \frac{\eta_t - f^*}{t} \in \mathcal{F} - \frac{1}{t}f^* \subseteq \overline{\mathcal{F} - \mathbb{R}_+ f^*}$, implying that $\mathcal{W} \subseteq \overline{\mathcal{F} - \mathbb{R}_+ f^*}$.

So we show that $\eta_t \in \mathcal{F}$. Now as before, notice that on $\mathbf{Z}_0$, $\det[\nabla^2 \eta_t - t\nabla^2 \delta] = \det[\nabla^2 f^*] = 0$. i.e. the determinant of the Hessian of $\eta_t - t\delta$, vanishes on $\mathbf{Z}_0$. Therefore, $\eta_t \in \mathcal{W} + t\delta$. However, since δ was chosen to be in $\mathcal{W}$ and $\mathcal{W}$ is a convex cone, we have that $\eta \in \mathcal{W}$. Thus the Hessian of η is negative semi-definite on $\mathbf{Z}_0$.

Now on $\mathbf{Z} - \mathbf{Z}_0$ the Hessian of f^* is negative definite, while no such statement can be made about the Hessian of δ. This implies that for all $\alpha \in \mathbb{R}^2$, $\alpha'[\nabla^2 \eta]\alpha = \alpha'[\nabla^2 f^*]\alpha + t\alpha'[\nabla^2 \delta]\alpha < 0$, for sufficiently small t. i.e. on $\mathbf{Z} - \mathbf{Z}_0$, η is strictly concave. Thus we have shown that for all $\alpha \in \mathbb{R}^2$,

$$\alpha' \nabla^2 \eta(u) \alpha = \begin{cases} \leq 0 & \text{if } u \in \mathbf{Z}_0, \\ < 0 & \text{if } u \in \mathbf{Z} - \mathbf{Z}_0. \end{cases}$$

That is, η is concave on $\mathbf{Z}$. And since $\eta \in \mathcal{H}$ by construction, we have that $\eta \in \mathcal{F}$. □

To prove Theorem 3.7.2, we need the following definition and the subsequent lemmas.

Definition G.1. Let $\mathcal{D}_{\mathcal{W}} = \{f \in \mathcal{H} : f = f_1 - f_2, \text{for some} f_1, f_2 \in \mathcal{W}\}$. That is, $\mathcal{D}_{\mathcal{W}}$ is the set of all functions in $\mathcal{H}$, which can be expressed as the difference of two functions in $\mathcal{W}$.

Remark G.2. It is easy to see from its definition, that $\mathcal{D}_{\mathcal{W}} \subset \mathcal{H}$ is a linear space containing $\mathcal{W}$. □

Lemma G.1. *$\mathcal{D}_{\mathcal{W}}$ is a Banach space containing $\mathcal{W}$.*

Proof. As the above remark shows, $\mathcal{D}_{\mathcal{W}}$ is a linear space containing $\mathcal{W}$. So it only remains to show that $\mathcal{D}_{\mathcal{W}}$ is complete. So let f_n be a sequence in $\mathcal{D}_{\mathcal{W}}$ that converges in C^2 norm to f. We show that $f \in \mathcal{D}_{\mathcal{W}}$. Since the convergence is in the C^2 norm, $f \in \mathcal{H}$. So if we can show that f can be written as the difference of two functions in $\mathcal{W}$, we are done. This is done as follows. For $\mathbf{z} \in \mathbf{Z}$, define $g_0(\mathbf{z}) = -(z_1^2 + z_2^2)$. Then since the Hessian of g_0 is negative definite on $\mathbf{Z}$, $g_0 \in \mathcal{W}$. Now for all $\mathbf{z} \in \mathbf{Z}$, consider the function $h(\mathbf{z}) = f(\mathbf{z}) + \frac{1}{\epsilon} g_0(\mathbf{z})$, where $\epsilon > 0$. Therefore, for all $\boldsymbol{\alpha} \in \mathbb{R}^2$,

$$\begin{aligned}\boldsymbol{\alpha}'[\nabla^2 h(\mathbf{z})]\boldsymbol{\alpha} &= \boldsymbol{\alpha}'[\nabla^2 f(\mathbf{z})]\boldsymbol{\alpha} + \frac{1}{\epsilon}\boldsymbol{\alpha}'[\nabla^2 g_0(\mathbf{z})]\boldsymbol{\alpha} \\ &< 0,\end{aligned}$$

for sufficiently small ϵ, since g_0 has a negative definite Hessian[1] and $\mathbf{Z}$ is compact. This implies that h is strictly concave for sufficiently small ϵ, i.e. $h \in \mathcal{W}$, and $f = h - \frac{1}{\epsilon} g_0$. Moreover, $\frac{1}{\epsilon} g_0 \in \mathcal{W}$, since $\epsilon > 0$ and $\mathcal{W}$ is a cone. Therefore, f can be written as the difference of two functions in $\mathcal{W}$. Hence, $f \in \mathcal{D}_{\mathcal{W}}$. □

[1] And therefore, g_0 is strictly concave.

LEMMA G.2. *Let $\mathfrak{C}$ be the collection of all Banach spaces containing $\mathcal{W}$, and let $\overline{lin\,\mathcal{W}} = \cap_{\mathcal{X}\in\mathfrak{C}}\mathcal{X}$.* [2] *Then, $\mathcal{D}_\mathcal{W} = \overline{lin\,\mathcal{W}}$.*

PROOF. Notice that since $\mathcal{H} \in \mathfrak{C}$, $\mathfrak{C}$ is not empty. We first show that $\mathcal{D}_\mathcal{W} \subset \overline{lin\,\mathcal{W}}$. So let $\mathcal{X}$ be any Banach space containing $\mathcal{W}$. Then due to the linearity of $\mathcal{X}$, $-\mathcal{W} \subset \mathcal{X}$. Now let $f \in \mathcal{D}_\mathcal{W}$. Therefore, $f = f_1 - f_2$ for some $f_1, f_2 \in \mathcal{W}$. But this implies that $f_1 \in \mathcal{W} \subset \mathcal{X}$ and $-f_2 \in -\mathcal{W} \subset \mathcal{X}$. Therefore, again by linearity of $\mathcal{X}$, $f_1 + (-f_2) \in \mathcal{X}$. That is, $\mathcal{D}_\mathcal{W} \subset \mathcal{X}$, and since $\mathcal{X}$ was an arbitrary element of $\mathfrak{C}$, $\mathcal{D}_\mathcal{W} \subset \cap_{\mathcal{X}\in\mathfrak{C}}\mathcal{X} = \overline{lin\,\mathcal{W}}$.

The other direction is even easier to show. By Lemma G.1, $\mathcal{D}_\mathcal{W}$ is a Banach space that contains $\mathcal{W}$. Also, by definition $\overline{lin\,\mathcal{W}}$ is the smallest Banach space containing $\mathcal{W}$. Therefore, these two statements together imply that $\overline{lin\,\mathcal{W}} \subset \mathcal{D}_\mathcal{W}$. □

LEMMA G.3. *$\mathcal{D}_\mathcal{W} = \mathcal{H}$.*

PROOF. Clearly $\mathcal{D}_\mathcal{W} \subset \mathcal{H}$. But $\mathcal{H} \subset \mathcal{D}_\mathcal{W}$ from Lemma G.1. □

We are now ready to prove Theorem 3.7.2.

PROOF. [**Theorem 3.7.2**] When f^* is strictly concave, $T(\mathcal{F}, f^*) = \mathcal{H}$, and there is nothing to prove. When f^* is concave, but not strictly concave, the proof is straightforward from Lemma G.2 and Lemma G.3. We now look at the case when f^* is affine on $\mathbf{Z}$. Notice that when $\mathbf{Z}_0 = \mathbf{Z}$, we obtain $\mathcal{W} = \mathcal{F}$. Hence, letting $\mathbf{Z}_0 = \mathbf{Z}$ and replacing $\mathcal{W}$ by $\mathcal{F}$ in Lemma G.1, Lemma G.2 and Lemma G.3, we obtain the required result. □

[2] That is, $\overline{lin\,\mathcal{W}}$ is the smallest closed linear space containing $\mathcal{W}$.

APPENDIX H

SOME MISCELLANEOUS RESULTS

PROOF. **[Theorem 2.6.1]** $\Longrightarrow$ We first show that $linT(\mathcal{F},f^*) \subset \overline{lin\mathcal{F}}$. This will imply that $\overline{lin\,T(\mathcal{F},f^*)} \subset \overline{lin\mathcal{F}}$. So let $f \in linT(\mathcal{F},f^*)$. This means that for some m, there exist $\{\alpha_1,\dots,\alpha_m\} \in \mathbb{R}^m$, and $\{f_1,\dots,f_m\} \in \times_{i=1}^m T(\mathcal{F},f^*)$ such that $f = \sum_{i=1}^m \alpha_i f_i$. Now since each $f_i \in T(\mathcal{F},f^*) = \overline{\mathcal{F} - \mathbb{R}_{++}f^*}$, there exist $(g_n,\lambda_n) \in \mathcal{F} \times \mathbb{R}_{++}$, such that $f_i = \lim_{n\to\infty}(g_n - \lambda_n f^*)$. But since $g_n \in \mathcal{F} \subset lin\mathcal{F}$ and $\lambda_n f^* \in \mathcal{F} \subset lin\mathcal{F}$, we get that $g_n - \lambda_n f^*$ is a convergent sequence in $lin\mathcal{F}$. Hence its limit is an element of $\overline{lin\mathcal{F}}$, i.e., $f_i \in \overline{lin\mathcal{F}}$. Therefore, each f_i is an element of $\overline{lin\mathcal{F}}$. Thus, $f = \sum_{i=1}^m \alpha_i f_i \in \overline{lin\mathcal{F}}$ which implies that $linT(\mathcal{F},f^*) \subset \overline{lin\mathcal{F}}$.

$\Longleftarrow$ We now show that $lin\mathcal{F} \subset linT(\mathcal{F},f^*)$. This will imply that $\overline{lin\mathcal{F}} \subset \overline{lin\,T(\mathcal{F},f^*)}$. So let $f \in lin\mathcal{F}$. This implies that there exist $\{\alpha_1,\dots,\alpha_m\} \in \mathbb{R}^m$, and $\{f_1,\dots,f_m\} \in \times_{i=1}^m \mathcal{F}$ such that $f = \sum_{i=1}^m \alpha_i f_i$. But since $f^* \in T(\mathcal{F},f^*)$, this means that each $f_i \in \mathcal{F} \subset T(\mathcal{F},f^*)$. Therefore, $f = \sum_{i=1}^m \alpha_i f_i \in linT(\mathcal{F},f^*)$. Hence, $lin\mathcal{F} \subset linT(\mathcal{F},f^*)$. □

THEOREM H.1 (A USEFUL RESULT). *Let f be a real valued* C^2 *function on* $\mathbb{R}^k$. *Let* $\mathbf{z} \in \mathbb{R}^k$, *and suppose that f is convex (resp. concave) at* $\mathbf{z}$. *Then,* $\det[\nabla^2 f(\mathbf{z})] = 0$ *iff there exists at least one non-zero* $\boldsymbol{\alpha} \in \mathbb{R}^k$, *such that* $\boldsymbol{\alpha}'[\nabla^2 f(\mathbf{z})]\boldsymbol{\alpha} = 0$.

PROOF. The proof is well known, but instructive. We provide it for the sake of completeness.

$\Longrightarrow$ Let $\det[\nabla^2 f(\mathbf{z})] = 0$. Then the system of linear equations $[\nabla^2 f(\mathbf{z})]\mathbf{x} = \mathbf{0}$, has a non-zero solution $\mathbf{x}_0 \in \mathbb{R}^k$. That is, $[\nabla^2 f(\mathbf{z})]\mathbf{x}_0 = \mathbf{0}$, which implies that $\mathbf{x}_0'[\nabla^2 f(\mathbf{z})]\mathbf{x}_0 = 0$.

$\Longleftarrow$ Now suppose that there exists a non-zero $\boldsymbol{\alpha}_0 \in \mathbb{R}^k$, such that $\boldsymbol{\alpha}_0'[\nabla^2 f(\mathbf{z})]\boldsymbol{\alpha}_0 = 0$. Let $Q(\boldsymbol{\alpha}) = \boldsymbol{\alpha}'[\nabla^2 f(\mathbf{z})]\boldsymbol{\alpha}$. Then since f was convex (resp. concave) at $\mathbf{z}$, we have that $Q(\boldsymbol{\alpha})$ is ≥ 0 (resp. ≤ 0). In either case, $\boldsymbol{\alpha}_0$ minimizes (resp. maximizes) $Q(\boldsymbol{\alpha})$. The first order conditions then imply that $\frac{dQ(\boldsymbol{\alpha})}{d\boldsymbol{\alpha}}|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0} = \mathbf{0}$. That is, $[\nabla^2 f(\mathbf{z})]\boldsymbol{\alpha}_0 = \mathbf{0}$. But this means that the system of linear equations $[\nabla^2 f(\mathbf{z})]\boldsymbol{\alpha} = \mathbf{0}$ has a non-zero solution $\boldsymbol{\alpha}_0$. That is, $\det[\nabla^2 f(\mathbf{z})] = 0$. □

THEOREM H.2 (CLASSICAL PROJECTION THEOREM). *Let H be a Hilbert space and M a closed subspace of H. Corresponding to any vector $x \in H$, there is a unique vector $m_0 \in M$ such that $||x - m_0|| \leq ||x - m||$ for all $m \in M$. Furthermore, a necessary and sufficient condition that $m_0 \in M$ be the unique minimizing vector is that $x - m_0$ be orthogonal to M.*

PROOF. See Luenberger (1969). □

THEOREM H.3 (PROJECTION ON CONVEX CONES). *Let H be a Hilbert space and M a closed convex cone in H. Corresponding to any vector $x \in H$, there is a unique vector $m_0 \in M$ such that $||x - m_0|| \leq ||x - m||$ for all $m \in M$. Furthermore, a necessary and sufficient condition that $m_0 \in M$ be the unique minimizing vector is that $\langle x - m_0, m_0\rangle = 0$, and that $\langle x - m_0, m\rangle \leq 0$ for all $m \in M$.*

PROOF. See Barlow, Bartholomew, Bremner, and Brunk (1972). □

PROPOSITION H.1 (CRAMÉR - WOLD DEVICE). *Let $\mathbf{X}_n$ denote a sequence of random variables in $\mathbb{R}^p$. Then $\mathbf{X}_n \xrightarrow{d} \mathbf{X} \Longleftrightarrow \boldsymbol{\delta}'\mathbf{X}_n \xrightarrow{d} \boldsymbol{\delta}'\mathbf{X}$, for every $\boldsymbol{\delta} \in \mathbb{R}^p$.*

APPENDIX I

PROOF OF LEMMA 4.3.1

Let $\mathbf{t}$ be a fixed point in $S_{\mathbf{X}}$. Then since $\hat{g}_n(\mathbf{t}) = (nb_n^p)^{-1}\sum_{j=1}^n y_j \mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})$,

$$(nb_n^p)^{1/2}[\hat{g}_n(\mathbf{t}) - \mathbb{E}_f\, \hat{g}_n(\mathbf{t})] = n^{-1/2}\sum_{j=1}^n w_{n,j}(\mathbf{t}), \quad \text{where,}$$

$$w_{n,j}(\mathbf{t}) = b_n^{-p/2}\{y_j\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n}) - \mathbb{E}\, y_j\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})\}.$$

Notice that $\mathbb{E}\, w_{n,j}(\mathbf{t}) = 0$, and that $\{w_{n,1}(\mathbf{t}),\dots,w_{n,n}(\mathbf{t})\}$ are i.i.d. terms. It is then easy to see that $\mathrm{Var}\, w_{n,j}(\mathbf{t}) = \sigma^2(\mathbf{t}) + o(1)$. Furthermore, after some more algebra we can also verify that $w_{n,j}(\mathbf{t})$ satisfies the sufficient condition in Lyapunov's CLT. That is, for some $\alpha > 0$,

$$\begin{aligned}\sum_{j=1}^n \mathbb{E}\,|\frac{w_{n,j}(\mathbf{t})}{\sqrt{n}}|^{2+\alpha} &= O(\left\{\frac{1}{nb_n^p}\right\}^{\alpha/2}) \\ &= o(1),\end{aligned}$$

since $nb_n^p \to \infty$. Therefore, we have proved that for each fixed $\mathbf{t} \in S_{\mathbf{X}}$,

$$(nb_n^p)^{1/2}[\hat{g}_n(\mathbf{t}) - \mathbb{E}_f\, \hat{g}_n(\mathbf{t})] \xrightarrow{d} \mathrm{N}(0, \sigma^2(\mathbf{t})).$$

This clearly shows that for each $\mathbf{t}_j$, where $j = 1, \ldots, m$,

$$(nb_n^p)^{1/2}\left[\hat{g}_n(\mathbf{t}_j) - \mathbb{E}_f\,\hat{g}_n(\mathbf{t}_j)\right] \xrightarrow{d} \mathrm{N}(0, \sigma^2(\mathbf{t}_j)).$$

However, in order to obtain the final result we still have to show that the covariance between the various terms is zero. We verify this fact for only two terms. The extension to m terms follows similarly.

So let $\mathbf{s}$ and $\mathbf{t}$ be any two fixed points in $S_{\mathbf{X}}$. Then using the result just obtained. the Cramér-Wold device yields that

$$\begin{pmatrix} n^{-1/2}\sum_{j=1}^n w_{n,j}(\mathbf{s}) \\ n^{-1/2}\sum_{j=1}^n w_{n,j}(\mathbf{t}) \end{pmatrix} \xrightarrow{d} \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2(\mathbf{s}) & \rho(\mathbf{s},\mathbf{t}) \\ \rho(\mathbf{s},\mathbf{t}) & \sigma^2(\mathbf{t}) \end{bmatrix}\right),$$

where, $\rho(\mathbf{s},\mathbf{t}) = \mathrm{cov}(w_{n,j}(\mathbf{s}), w_{n,j}(\mathbf{t}))$. Now to show $\rho(\mathbf{s},\mathbf{t}) = o(1)$, notice that

$$\begin{aligned} \rho(\mathbf{s},\mathbf{t}) &= \mathrm{cov}(w_{n,j}(\mathbf{s}), w_{n,j}(\mathbf{t})) \\ &= \frac{1}{b_n^p}\mathrm{cov}(y_j\mathrm{K}(\frac{\mathbf{s}-\mathbf{x}_j}{b_n}), y_j\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})) \\ &= \frac{1}{b_n^p}\mathbb{E}\{\mathrm{K}(\frac{\mathbf{s}-\mathbf{x}_j}{b_n})\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})\gamma(\mathbf{x}_j)\} \\ &\quad - \frac{1}{b_n^p}\mathbb{E}\{\mathrm{K}(\frac{\mathbf{s}-\mathbf{x}_j}{b_n})f(\mathbf{x}_j)\}\,\mathbb{E}\{\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})f(\mathbf{x}_j)\}, \end{aligned}$$

where, $\gamma(\mathbf{t}) = \mathbb{E}(y^2|\mathbf{t})$. Now let $\mathbf{s} - \mathbf{x}_j = b_n\mathbf{u}$. Then by Lebesgue's Dominated Convergence Theorem,

$$\begin{aligned} &\frac{1}{b_n^p}\mathbb{E}\{\mathrm{K}(\frac{\mathbf{s}-\mathbf{x}_j}{b_n})\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})\gamma(\mathbf{x}_j)\} = \\ &\int_{[-1,1]^p}\mathrm{K}(\mathbf{u})\mathrm{K}(\frac{\mathbf{t}-\mathbf{s}}{b_n}+\mathbf{u})\gamma(\mathbf{s}-b_n\mathbf{u})p(\mathbf{s}-b_n\mathbf{u})\,d\mathbf{u} \\ &= o(1), \end{aligned}$$

since $\mathrm{K}(\frac{\mathbf{t}-\mathbf{s}}{b_n} + \mathbf{u}) \longrightarrow 0$ as $b_n \longrightarrow 0$. Similarly, we can show that

$$\begin{aligned}
&\frac{1}{b_n^p}\mathbb{E}\{\mathrm{K}(\frac{\mathbf{s}-\mathbf{x}_j}{b_n})f(\mathbf{x}_j)\}\mathbb{E}\{\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})f(\mathbf{x}_j)\} = \\
&b_n^p\int_{[-1,1]^p}\mathrm{K}(\mathbf{u})f(\mathbf{s}-b_n\mathbf{u})p(\mathbf{s}-b_n\mathbf{u})\,d\mathbf{u}\cdot\int_{[-1,1]^p}\mathrm{K}(\mathbf{u})f(\mathbf{t}-b_n\mathbf{u})p(\mathbf{t}-b_n\mathbf{u})\,d\mathbf{u} \\
&= o(1).
\end{aligned}$$

Substituting these results in the expression for $\rho(\mathbf{s},\mathbf{t})$ yields that $\rho(\mathbf{s},\mathbf{t}) \to 0$. Hence, we are done. □

APPENDIX J

PROOF OF LEMMA 4.3.2

First note that since $\mathbb{E}_f\,(y_j|\mathbf{x}_j) = f(\mathbf{x}_j)$, we have that $\mathbb{E}_{\hat{f}_n^*}\,(y_j|\mathbf{x}_j) = \hat{f}_n^*(\mathbf{x}_j)$. Therefore, using the law of iterated mathematical expectations and the fact that the observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are i.i.d.,

$$\begin{aligned}
\mathbb{E}_{\hat{f}_n^*}\,\hat{g}_n(\mathbf{t}) &= \frac{1}{nb_n^p}\,\mathbb{E}_{\hat{f}_n^*}\,\sum_{j=1}^{n} y_j \mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n}) \\
&= \frac{1}{b_n^p}\,\mathbb{E}\,\{\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})\mathbb{E}_{\hat{f}_n^*}\,(y_j|\mathbf{x}_j)\} \\
&= \frac{1}{b_n^p}\,\mathbb{E}\,\{\mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})\hat{f}_n^*(\mathbf{x}_j)\} \\
&= \frac{1}{b_n^p}\int_{[-1,1]^p} \mathrm{K}(\frac{\mathbf{t}-\mathbf{x}_j}{b_n})\hat{f}_n^*(\mathbf{x}_j)p(\mathbf{x}_j)\,d\mathbf{x}_j \\
&= \int_{[-1,1]^p} \mathrm{K}(\mathbf{u})\hat{f}_n^*(\mathbf{t}-b_n\mathbf{u})p(\mathbf{t}-b_n\mathbf{u})\,d\mathbf{u}.
\end{aligned}$$

Similarly,

$$\mathbb{E}_f\,\hat{g}_n(\mathbf{t}) = \int_{[-1,1]^p} \mathrm{K}(\mathbf{u})f(\mathbf{t}-b_n\mathbf{u})p(\mathbf{t}-b_n\mathbf{u})\,d\mathbf{u}, \quad \text{and the bias term}$$

$$B_n(\mathbf{t}) = (nb_n^p)^{1/2}\int_{[-1,1]^p} \mathrm{K}(\mathbf{u})p(\mathbf{t}-b_n\mathbf{u})\{\hat{f}_n^*(\mathbf{t}-b_n\mathbf{u}) - f(\mathbf{t}-b_n\mathbf{u})\}\,d\mathbf{u}.$$

Now let $B_n(\mathbf{t}|\mathrm{H}_0)$ denote the bias under H_0, while $B_n(\mathbf{t}|\mathrm{H}_{1n})$ denotes the bias under

H_{1n}. Then,

$$\begin{aligned}
B_n(\mathbf{t}|\mathrm{H}_{1n}) &= (nb_n^p)^{1/2}\int_{[-1,1]^p}\mathrm{K}(\mathbf{u})p(\mathbf{t}-b_n\mathbf{u})\,\{\,\hat{f}_n^*(\mathbf{t}-b_n\mathbf{u}) - f^*(\mathbf{t}-b_n\mathbf{u}) \\
&\qquad - \frac{\delta(\mathbf{t}-b_n\mathbf{u})}{\sqrt{nb_n^p}}\,\}\,d\mathbf{u} \qquad\qquad (\mathrm{J}.1)\\
&= (nb_n^p)^{1/2}\int_{[-1,1]^p}\mathrm{K}(\mathbf{u})p(\mathbf{t}-b_n\mathbf{u})\{\hat{f}_n^*(\mathbf{t}-b_n\mathbf{u}) - f^*(\mathbf{t}-b_n\mathbf{u})\}\,d\mathbf{u} \\
&\qquad - \int_{[-1,1]^p}\mathrm{K}(\mathbf{u})p(\mathbf{t}-b_n\mathbf{u})\delta(\mathbf{t}-b_n\mathbf{u})\,d\mathbf{u} \\
&= B_n(\mathbf{t}|\mathrm{H}_0) - \int_{[-1,1]^p}\mathrm{K}(\mathbf{u})p(\mathbf{t}-b_n\mathbf{u})\delta(\mathbf{t}-b_n\mathbf{u})\,d\mathbf{u} \\
&= B_n(\mathbf{t}|\mathrm{H}_0) - p(\mathbf{t})\delta(\mathbf{t}) + o(1).
\end{aligned}$$

Therefore, to prove Lemma 4.3.2 we simply have to show that $B_n(\mathbf{t}|\mathrm{H}_0) = o(1)$. However, before showing this we consider some local alternatives that cannot be distinguished from the null hypothesis.

REMARK J.1. As pointed out by Severini and Staniswalis (1991), (J.1) clearly shows that local alternatives that converge to f^* at rates faster than $(nb_n^p)^{1/2}$ will not be detectable. For in this case we would simply have that $B_n(\mathbf{t}|\mathrm{H}_{1n}) = B_n(\mathbf{t}|\mathrm{H}_0) + o(1)$. Moreover, the same would also hold for all points $\mathbf{t}$ at which $\delta(\mathbf{t}) = 0$, and therefore such local alternatives are also not detectable. One way to get rid of such local alternatives is to make m a function of n, such that $m(n) \to \infty$, as $n \to \infty$. This means that the function δ, which satisfies $\delta(\mathbf{t}_j) = 0$ for $j = 1, \ldots, m(n)$, exhibits a highly oscillatory behavior as $n \to \infty$, and is therefore unsuitable as an alternative. □

Now back to our original problem, i.e. showing $B_n(\mathbf{t}|\mathrm{H}_0) = o(1)$. To see this, notice that since the kernel vanishes outside $[-1,1]^p$,

$$\begin{aligned}|B_n(\mathbf{t}|\mathrm{H}_0)| &= (nb_n^p)^{1/2}\left|\int_{[-1,1]^p}\mathrm{K}(\mathbf{u})p(\mathbf{t}-b_n\mathbf{u})\{\hat{f}_n^*(\mathbf{t}-b_n\mathbf{u})-f^*(\mathbf{t}-b_n\mathbf{u})\}\,d\mathbf{u}\right|\\ &\leq (nb_n^p)^{1/2}\sup_{\mathbf{u}\in[-1,1]^p}|\hat{f}_n^*(\mathbf{u})-f^*(\mathbf{u})|\sup_{\mathbf{u}\in[-1,1]^p}p(\mathbf{u}).\end{aligned}$$

But from Remark 4.2.1(iii) we have

$$\sup_{\mathbf{u}\in[0,1]^p}|\hat{f}_n^*(\mathbf{u})-f^*(\mathbf{u})| = O_p(\{\log n/n\}^{\frac{2}{3+p}}),$$

and this implies that $B_n(\mathbf{t}|\mathrm{H}_0) = O_p(\sqrt{nb_n^p}\{\frac{\log n}{n}\}^{\frac{2}{3+p}})$. Now, from standard results on kernel regression we know that the asymptotically optimal choice of bandwidth is given by $b_n = O(n^{-\frac{1}{4+p}})$. With this choice of b_n, it is easily seen that for $\xi = \frac{3+p}{4+p} \in (0,1)$,

$$\begin{aligned}B_n(\mathbf{t}|\mathrm{H}_0) &= O_p(\sqrt{nb_n^p}\{\log n/n\}^{\frac{2}{3+p}})\\ &= O_p(\left\{\frac{\log n}{n^{1-\xi}}\right\}^{\frac{2}{3+p}})\\ &= o_p(1).\end{aligned}$$

Hence, $B_n(\mathbf{t}|\mathrm{H}_0) = o_p(1)$ uniformly in $\mathbf{t}$. □

BIBLIOGRAPHY

AUBIN, J. P., AND H. FRANKOWSKA (1990): *Set Valued Analysis*. Birkhauser.

BAHADUR, R. (1964): "On Fisher's Bound for Asymptotic Variances," *Annals of Mathematical Statistics*, 35, 1545–1552.

BARLOW, R., D. BARTHOLOMEW, J. BREMNER, AND H. BRUNK (1972): *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression*. John Wiley and Sons.

BICKEL, P., C. KLASSEN, Y. RITOV, AND J. WELLNER (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Press.

CHAMBERLAIN, G. (1986): "Asymptotic efficiency in semiparametric models with censoring," *Journal of Econometrics*, 32, 189–218.

COSSLETT, S. (1987): "Efficiency bounds for distribution free estimators of binary choice and censored regression models," *Econometrica*, 55, 559–585.

DURRETT, R. (1991): *Probability:Theory and Examples*. Wadsworth.

ICHIMURA, H. (1993): "Semiparametric Least Squares (SLS) and weighted SLS estimation of single index models," *Journal of Econometrics*, 58(1), 71–120.

JAIN, N., AND M. MARCUS (1975): "Central Limit Theorem for C(S)-valued random variables," *Journal of Functional Analysis*, 19, 216–231.

KOLMOGOROV, A., AND V. TIHOMIROV (1961): "ϵ - entropy and ϵ - capacity of sets in functional spaces," *American Mathematical Society Translation*, 17, 277–364.

KRABS, W. (1979): *Optimization and Approximation*. John Wiley and Sons.

LeCam, L. (1953): "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates," *Univ. California Publ. Statist.*, 1, 277–330.

Luenberger, D. G. (1969): *Optimization by Vector Space Methods*. John Wiley and Sons.

Matzkin, R. (1994): "Restrictions of Economic Theory in Nonparametric Methods," in *Handbook of Econometrics, vol. IV*, ed. by R. Engle, and D. McFadden, pp. 2524–2558. Elsevier Science B.V.

Newey, W. K. (1988): "Efficient estimation of Tobit models under symmetry," in *Nonparametric and semiparametric methods in econometrics and statistics. Proceedings of the fifth international symposium in Economic Theory and Econometrics*, ed. by W. A. Barnett, J. Powell, and G. Tauchen, pp. 291–336. Cambridge University Press.

——— (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1991): "The asymptotic variance of semiparametric estimators," *Research Memorandum No.* 346, Econometrics Research Program, Princeton University.

Pfanzagl, J. (1994): *Parametric Statistical Theory*. de Gruyter.

Rieder, H. (1994): *Robust Asymptotic Statistics*. Springer-Verlag.

Robinson, P. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56(4), 931–954.

Severini, T. A. (1987): "Efficient Estimation in Semiparametric Models," Ph.D. thesis, Department of Statistics, University of Chicago.

Severini, T. A., and J. G. Staniswalis (1991): "Diagnostics for Assessing Regression Models," *Journal of the American Statistical Association*, 86(415), 684–692.

Severini, T. A., and W. H. Wong (1992): "Profile likelihood and conditionally parametric models," *The Annals of Statistics*, 20(4), 1768–1802.

Stein, C. (1956): "Efficient Nonparametric Testing and Estimation," in *Proceedings of the third Berkeley Symposium on mathematical statistics and probability*, vol. 1, pp. 187–195. University of California Press, Berkeley, CA.

Stone, C. J. (1982): "Optimal Global Rates Of Convergence For Nonparametric Regression," *The Annals of Statistics*, 10(4), 1040–1053.

van der Vaart, A. (1989): "On the Asymptotic Information Bound," *The Annals of Statistics*, 17(4), 1487–1500.