Spline Based Estimation for Sufficient Dimension Reduction

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE
in
MATHEMATICS

Author :	Abed Razawy
Student ID :	1541420
Supervisor :	Thomas Nagler
2nd corrector :	-

Leiden, The Netherlands, February 15, 2022

Spline Based Estimation for Sufficient Dimension Reduction

Abed Razawy

February 15, 2022

Abstract

In regression analysis we attempt to quantify the relationship between a response variable and a set of predictors. When the number of predictors is large, it is computationally expensive to estimate the function that encapsulates their relationship. In sufficient dimension reduction, we attempt to reduce the high dimensional space of the predictors to a lower dimensional space of linear combinations of the same predictors. In this thesis we consider the sufficient dimension reduction paradigm applied to the conditional mean of a regression model. In particular, we propose a B-spline method to estimate the dimension reduction subspace. The linearity of this estimation technique allows us to compute the estimation criterion efficiently, and faster than existing competitors. Under some mild conditions we prove that our proposed central mean subspace estimator achieves $\sqrt{n}$-consistency and asymptotic normality, and show that our estimation of the structural dimension and central mean subspace are consistent. Our approach to proving these results is centered around empirical process theory. In simulations we demonstrate that our methodology is easy to implement and performs well for various models.

Contents

Chapter 1

Introduction

Many statistical problems are motivated by finding a relationship between a response Y and a set of predictors $X = (X_1, ..., X_p)^T \in \mathbb{R}^p$. When the number of predictors p is large, in many situations there is redundant or irrelevant information present among the predictors. Thus a method for finding a subset of covariates that have a relevant relationship with the response Y is desirable. There are two common approaches to this problem. The first is variable selection, where the assumption is made that only a few covariates are truly related to the response, and all others have no explanatory effect. The second approach is to use dimension reduction. In this setting one still assumes that many predictors could have explanatory power, however these are expressed through a few linear combinations. The high-dimensional covariate space is then replaced by the low-dimensional space of linear combinations. The problem of correctly identifying the most informative *smallest* space of linear combinations of covariates is known as the *sufficient dimension reduction* paradigm. Typically there is no pre-specified model required and minimal assumptions are made, which makes this approach very appealing when one knows there are only a few relevant linear combinations. The name is derived from two concepts in statistics, sufficiency and dimension reduction.

When the relationship between Y and X is known, this significantly reduces the complexity of the problem. For example, if the mean of Y conditional on X is a linear function in X, by which we mean that $E(Y|X) = \beta^T X$, this amounts to linear regression. If the function through which the relationship between Y and X is expressed is unknown, but could be any function g, this problem becomes significantly more difficult. The goal is to determine a sufficiently large function space to estimate g, but not give up any efficiency in the estimation of the dimension reduction subspace. To make the least amount of assumptions, we could choose a model with no finite or fixed number of parameters. Such a model is called a *nonparametric* model. For our purposes it is sufficient to define a nonparametric model as a model which is not *parametric*. A parametric model is a model which is (smoothly) indexed by a Euclidean parameter. The estimation of the parametric part in the dimension reduction, combined with nonparametric estimation of the function which relates Y and $\beta^T X$ is an example of *semiparametric* estimation.

1.1 Thesis Motivation

In Huang and Chiang (2017) a semi-parametric estimation criterion for the sufficient dimension reduction model is presented. A *Kernel regression*-type estimator is used to estimate the nonparametric function. Then a cross-validation criterion is proposed, simultaneously estimating the basis and structural dimension of the smallest dimension reduction, and the bandwidth of a kernel estimator. Asymptotic results are proven, among which $\sqrt{n}$-consistency and asymptotic normality. However this approximation is computationally expensive when n is not small, in part due to the required n model estimations, once for each omitted case, in the cross-validation criterion. The estimation method we propose is a linear model, which allows us to estimate the model only once on the complete data set (Seber and Lee (2012)). The idea for our estimator is rooted in the Stone-Weierstrass theorem, which demonstrates that any continuous function on the closed interval $[a, b]$ can be arbitrarily closely approximated by polynomials. Although appealing because of their simple form, polynomials perform poorly for function approximation. A favorable alternative to estimate a function are *splines*, piecewise polynomials. We demonstrate that we can use splines to estimate the unknown function, and obtain asymptotically similar results to Huang and Chiang. Results from *empirical process theory* are applied to prove $\sqrt{n}$-consistency and asymptotic normality.

1.2 Overview

The first few chapters of this thesis are a gentle introduction to techniques used in sufficient dimension reduction and techniques we use in our proofs. Chapter 2 introduces kernel and spline methods for regression. In Chapter 3 we review some of the sufficient dimension reduction literature. We are particularly interested in fundamental results, such as the existence and uniqueness of dimension reductions. We summarize Huang and Chiang (2017), and review the motivation to use their estimation criterion. We then propose our estimator, and explain the difference in the computation of the cross-validation criterion. In Chapter 4 we familiarize the reader with definitions and results from empirical process theory. This chapter is based on van der Vaart and Wellner (1996). The goal is to give an exposition of major components of the modern theory of empirical processes. In Chapter 5 we demonstrate how the results from empirical process theory can be applied in a parametric version of the dimension reduction problem. We assume that the function which relates Y and $\beta^T X$ is known, and it is seen how this simplifies our problem significantly. In Chapter 6 we propose a novel approach for estimating the sufficient dimension reduction spaces. We use empirical process theory to prove consistency and a rate of convergence. The main challenge in the latter is achieving sharp bounds on certain maximal inequalities. We introduce the semi-parametric estimation paradigm formally, and discuss how one can go about estimation in this setting. By combining results we demonstrate with the theory from Chapter 4 and the semiparametric framework, we show $\sqrt{n}$-consistency and asymptotic normality. This first few sections in this chapter are based on the first four chapters of the wonderfully written book *Semiparametric Theory and Missing Data* written by Tsiatis (2006). To conclude our discussion of sufficient dimension reduction, we perform simulations to demonstrate how our approach can be implemented, and how it performs under various models in Chapter 7.

Chapter 2

Nonparametric Regression

In this chapter we briefly review two nonparametric estimation methods. Nonparametric techniques allow the complexity of the fitted model to depend on the sample: The larger the sample size, the greater the complexity of the fitted model. In this setting it is acknowledged that fitted models are approximations, and therefore inherently misspecified. This misspecification implies estimation bias, however one can increase the complexity of the fitted model in order to decrease this bias. As a result the estimation variance is increased. This relation is known as the bias-variance trade-off, and minimizing the mean squared error (MSE), a measure of fit, which consists of both of these terms is a central problem in nonparametric estimation. In the first section we give a heuristic derivation of kernel regression. The rest of the chapter is used to review a particular kind of spline regression, B-splines, which we use in Chapter 6.

2.1 Kernel Regression

The idea of kernel density estimation resembles that of histograms, but with some added ”smoothness”. Let X be a random variable drawn from a probability density f. Then, for any $x \in \mathbb{R}$ and $h > 0$ sufficiently small,

$$P(X \in [x-h, x+h]) = \int_{x-h}^{x+h} f(u)du \approx 2hf(x).$$

This yields,

$$f(x) \approx \frac{1}{2h} P(X \in [x-h, x+h]).$$

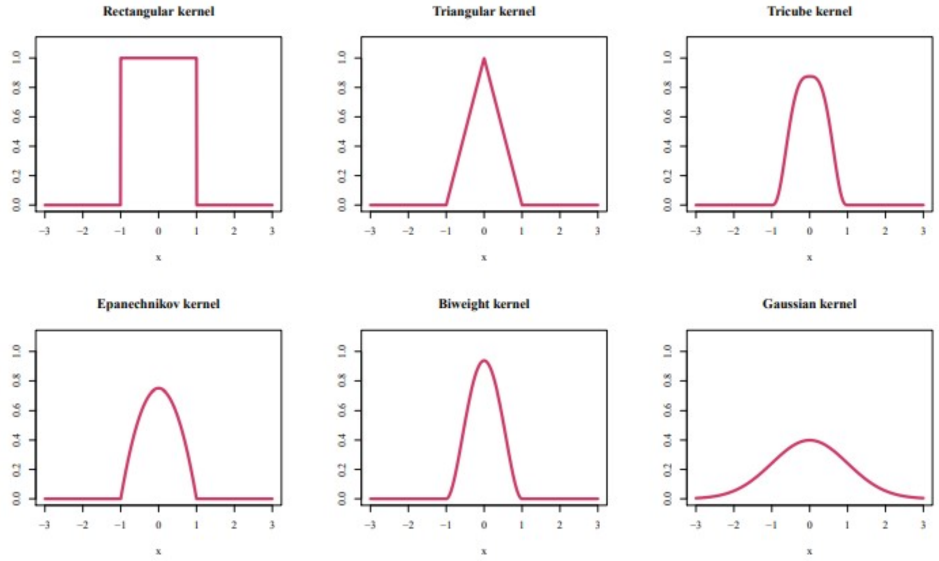

Figure 2.1: *Some commonly used kernels.*

This motivates our decision to estimate the probability density of X at x by the relative frequency of observations in a small interval around this point. If we observe independent and identically distributed (i.i.d.) copies, $X_1, ..., X_n$ of X, we can estimate f by,

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} w_h(x - X_i),$$

where w is a so-called *window function*, given by,

$$w_h(x) = \begin{cases} 1/2h, & \text{if } x \in [x-h, x+h] \\ 0, & \text{if } x \notin [x-h, x+h]. \end{cases}$$

Note that our estimate for $f(x)$ will only depend on observations that are sufficiently ”close” to x. More generally, we can replace the weight $w(x)$ with a *kernel*. A kernel is a function $K : \mathbb{R} \to \mathbb{R}$ that satisfies,

$$\int K(v)dv = 1; \qquad \int vK(v)dv = 0.$$

Some commonly used kernel functions are depicted in Figure 2.1. This gives us an estimate of the shape of f, known as the *kernel density estimator*

$$\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Suppose we observe i.i.d. random copies (X_i, Y_i); $i = 1, ..., n$, of (X, Y) with,

$$Y = f(X) + \epsilon, \tag{2.1}$$

and $E(\epsilon|X) = 0$. The conditional expectation of Y given X can be written as,

$$E(Y|X) = f(X).$$

Model (2.1) is known in the literature as nonparametric regression with random design.

The *Nadaraya-Watson estimator with kernel K* constructs an estimate of the regression function f as,

$$\widehat{f}_{n,h}(x) = \frac{\sum_{i=1}^n Y_i K(\frac{x-X_i}{h})}{\sum_{i=1}^n K(\frac{x-X_i}{h})},$$

where $h > 0$ is a smoothing parameter called the *bandwidth* of the kernel. When the denominator equals 0, we estimate f by 0. Note that this is a weighted average of Y_i values, and the denominator is a weighting term. It turns out that most choices for kernels K give similar results, and choosing the optimal bandwidth value h is of much more importance for the properties of the estimator. The value of the bandwidth determines how much bias and variance you introduce in the estimation, the balancing of these two quantities is the aforementioned bias-variance tradeoff. In practice we often choose h by cross-validation.

We often denote the dependence on h as $K_h(\cdot - X_i) = \frac{1}{h}K(\frac{\cdot - X_i}{h})$. At each observed point we place a kernel which has mass 1 and is centered around the observed point. An example of a kernel is any probability density with mean zero. A kernel K_q is of *order q* if additionally the $(q-1)$-th moments are zero, i.e.

$$\int v^k K(v) dv = 0 \text{ for } k = 1, ..., q-1,$$

and $\int v^q K(v) dv < \infty$. Higher order kernels are kernels of order $q > 2$, and have negative parts. As such, they are not probability distributions. Higher order kernels can be obtained by multiplying a second order kernel by a polynomial of order $(\frac{q}{2} - 1)$ in x^2. For example a kernel of order 4 is defined by,

$$K_4(x) = \frac{1}{2\sqrt{2\pi}}(3-x)^2 \exp(-\frac{x^2}{2}).$$

If we have multivariate predictors $X \in \mathbb{R}^p$, we can get multivariate kernels by multiplying univariate kernels,

$$K_{q,h_p}(u) = \prod_{k=1}^{p} K_q(u_k/h_k)/h_k, \qquad u = (u_1, ..., u_p) \in \mathbb{R}^p$$

where $h_p = (h_1, ..., h_p)^T$ is a positive valued bandwidth vector, and K_q is a q-th order (univariate) kernel function. We refer the interested reader to Wasserman (2005) and Hastie et al. (2001) for a more detailed review of kernel regression.

2.2 Splines

Analytic functions can be approximated locally reasonably well by polynomial functions, for example by a Taylor polynomial of sufficiently high order. A main drawback of using a polynomial to estimate a function is the lack of robustness: a slight change of one data point may cause large changes in the regression parameters and polynomial approximations. A preferable method are piecewise polynomials, which only react locally to changes in the data. These are known as *splines*. We use a spline function known as *B-splines*. B-splines have some very appealing properties, for example, any B-spline can be written as a linear combination of basis functions. These basis functions have compact support, and are given by explicit formulas. Splines are often used in applied mathematics, numerical analysis, geometrical modelling, and in particular in applications requiring the interpolation or smoothing of data. In the remainder of this section we discuss the basic properties of univariate splines. In the rest of this chapter we illustrate how B-splines are constructed and used in regression-type problems. Then we review how these methods can be easily generalized to multivariate B-splines using tensor products.

Suppose we want to approximate a univariate function $f : [L, U] \to \mathbb{R}$. A spline $S(x)$ maps values from the interval $[L, U]$ to the set of real numbers,

$$S : [L, U] \to \mathbb{R}.$$

In particular we partition $[L, U]$ into subintervals $[t_i, t_{i+1}]$ such that S is a polynomial on each subinterval. The points t_i indicate where S changes from one polynomial to another, and are called *knots*. A *knot sequence* $\boldsymbol{t} = \{t_0, t_1, ..., t_K\}$ is defined as an non-decreasing sequence of $K + 1$ knots, i.e.

$$L = t_0 \leq t_1 \leq ... \leq t_K = U,$$

and we define polynomials p_i such that,

$$S(x) \equiv p_i(x) \qquad \text{on } [t_i, t_{i+1}), \qquad \text{for } i = 0, ..., K-1.$$

The degree ρ of the polynomials which define $S(x)$ is known as the *degree of the spline*. The *order of the spline* is the degree plus one, i.e. $\rho + 1$.

2.3 Function Approximation with B-splines

B-splines were first described in Schoenberg's (1988) paper *Contributions to the problem of approximation of equidistant data by analytic functions*. It turns out that any spline of a given degree ρ, can be written as a B-spline of that same degree. Thus the B-splines form a basis for the space of spline functions.[*] The smoothness of the estimate is usually enforced by requiring the derivatives (up to the $\rho - 1$-th derivative) of the adjacent polynomials to be the same at any knot. The function $f(x)$ can be approximated using a B-spline,

$$\widehat{f}(x) = S(x) = \sum_{i=0}^{q_n-1} \psi_{i,\rho}(x) a_i, \tag{2.2}$$

where $\psi_{i,\rho}(x)$, $i = 0, ...q_n - 1$ are B-spline basis functions of degree ρ on a knot sequence t, and a_i are the corresponding coefficients. The basis functions are also piecewise polynomials that are smoothly connected at the knots, and are only non-zero on their respective domains. The amount of smoothing is determined by the degree of the spline and the number of the knots. The similarity of the role of the degree and number of knots for the B-splines, and the degree and bandwidth in the kernel regression model should be apparent. In this thesis we denote the spline approximation in Equation (2.2) as the vector product,

$$\psi(x)^T a = \sum_{i=0}^{q_n-1} \psi_{i,p}(x) a_i = (\psi_{0,\rho}(x), ..., \psi_{q_n-1,\rho}(x)) \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{q_n-1} \end{pmatrix}. \tag{2.3}$$

Whenever we write basis functions, it is implicit that there exist a sequence of corresponding knots. The B-spline basis functions can be constructed recursively by the Cox-de-Boor formula as follows,

$$\psi_{i,0}(x) = \begin{cases} 1, & \text{if } x \in [t_i, t_{i+1}) \\ 0, & \text{otherwise,} \end{cases} \tag{2.4}$$

[*] Coincidentally, the "B" stands for Basis.

$$\psi_{i,\rho}(x) = \frac{x - t_i}{t_{i+\rho} - t_i}\psi_{i,\rho-1}(x) + \frac{t_{i+\rho+1} - x}{t_{i+\rho+1} - t_{i+1}}\psi_{i+1,\rho-1}(x). \tag{2.5}$$

Figure 2.2 shows the B-spline basis functions up to degree 3 (order 4) with 10 uniformly placed knots in the interval $[0, 1]$. B-splines basis functions constructed by the Cox-de-Boor recursive formula have the following properties:

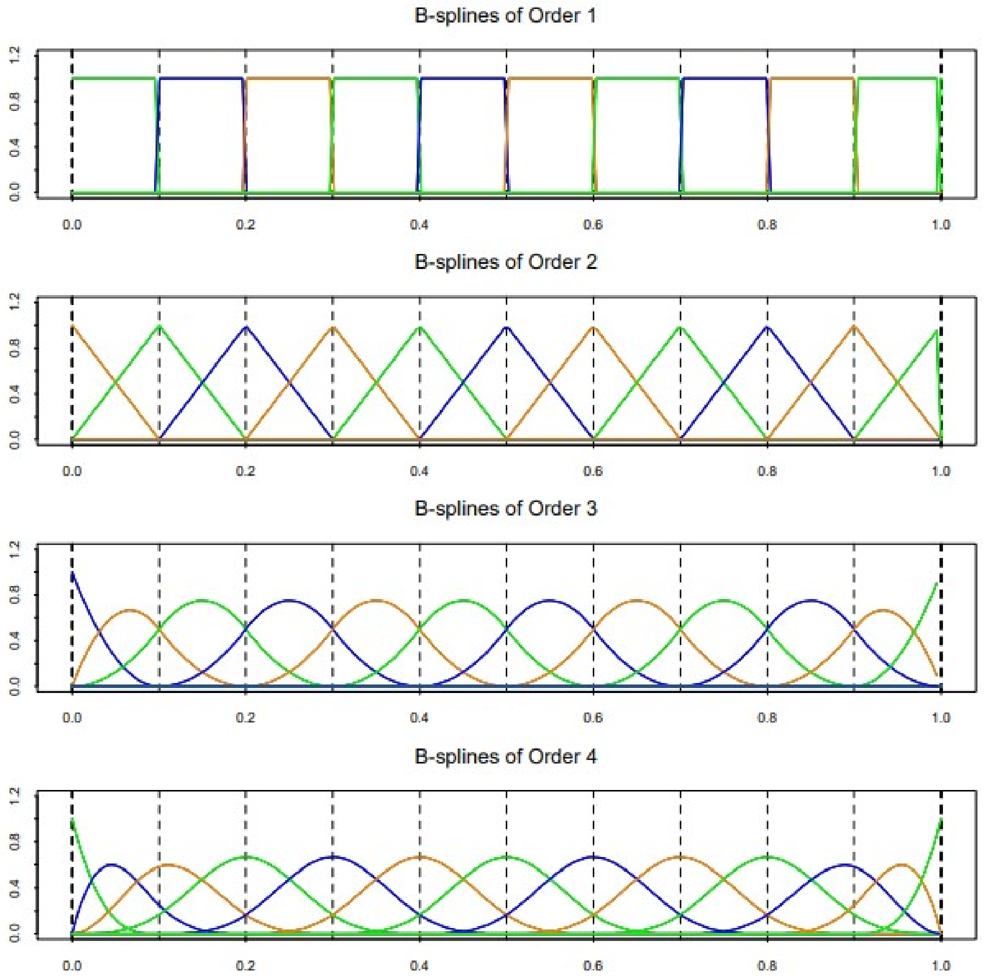

Figure 2.2: *The sequence of B-splines up to order four. The ten knots are uniformly spaced from 0 to 1. (Hastie et al. (2001))*

1. Basis functions are non-negative, i.e., $\psi_{i,\rho}(x) \geq 0$, for all x;

2. At any point, B-splines of degree ρ sum to 1, i.e., $\sum_i \psi_{i,\rho}(x) = 1$ for all x in the interior domain;
3. For all points outside of their domain, the basis functions equal zero, i.e. $\psi_{i,\rho}(x) = 0$ for all $x \notin [t_i, t_{i+\rho+1})$;
4. $\psi_{i,\rho}(x)\Big|_{x\in[t_{i+j},t_{i+j+1})}$ is a polynomial of degree ρ, for all j;
5. The derivative of a B-spline of degree ρ is another B-spline of degree $\rho - 1$ and
$$\psi'_{i,\rho}(x) = \rho\left\{\frac{\psi_{i,\rho-1}(x)}{t_{i+\rho} - t_i} - \frac{\psi_{i+1,\rho-1}(x)}{t_{i+\rho+1} - t_{i+1}}\right\};$$
6. When the knots are distinct, a B-spline of degree ρ, its derivatives are continuous up to the $\rho - 1$-th derivative.

If we have repeating knots, the convention that $0/0 = 0$ is used. Note that B-splines are only non-zero on a compact subset, and are positive within their support. This can be exploited to ensure a faster computation time (for more details, see p. 186 in Hastie et al. (2001)). The *partition of unity property of B-splines* (property 2) states that on any point within the natural domain of a B-spline curve of degree ρ, which is given by $[t_\rho, t_{q_n-\rho-1}]$, the sum of B-splines on any point within this domain equals 1,

$$\sum_{i=0}^{q_n-1} \psi_{i,\rho}(x) = 1, \qquad t_\rho \leq x \leq t_{q_n-\rho-1}.$$

From property 1 and 2 it follows that for splines of any degree ρ, at any point x within its domain,

$$\|\psi(x)\|_1 = \|(\psi_{0,\rho}(x), \psi_{1,\rho}(x), \dots, \psi_{q_n-1,\rho}(x))\|_1 = 1. \tag{2.6}$$

Let $(X_1, Y_1), ..., (X_n, Y_n)$ be drawn according to model (2.1). We slightly abuse notation by denoting $X = (X_1, ..., X_n)^T$ and $Y = (Y_1, ..., Y_n)^T$ in the following. The B-spline of degree ρ estimating f takes the following form,

$$\widehat{f}(X) = \begin{pmatrix} \widehat{f}(X_1) \\ \vdots \\ \widehat{f}(X_n) \end{pmatrix} = \psi(X)^T a = \begin{pmatrix} \psi_{0,\rho}(X_1) & \dots & \psi_{q_n-1,\rho}(X_1) \\ \vdots & \ddots & \vdots \\ \psi_{0,\rho}(X_n) & \dots & \psi_{q_n-1,\rho}(X_n) \end{pmatrix} a.$$

We can estimate the coefficients a by minimizing the MSE,

$$S(a) = \|Y - \widehat{f}(X)\|_2^2 = \|Y - \psi(X)^T a\|_2^2.$$

We can easily derive the minimizer of this expression, $\widehat{a}$, by setting the gradient with respect to a to zero. For this, note that,

$$\begin{aligned} S(a) &= (Y - \psi(X)^T a))^T (Y - \psi(X)^T a)) \\ &\quad = Y^T Y - 2Y^T \psi(X)^T a + (\psi(X)^T a)^T (\psi(X)^T a), \\ \nabla_a S(a) &= -2\psi(X)^T Y + 2\psi(X)\psi(X)^T a, \\ \nabla_a^2 S(a) &= \psi(X)^T \psi(X). \end{aligned}$$

These equations imply that whenever $\psi(X)$ has full column rank, the Hessian is positive semi-definite. Consequently, the minimizer of $S(a)$ is equal to

$$\widehat{a} = (\psi(X)\psi(X)^T)^{-1}\psi(X)Y. \tag{2.7}$$

2.4 Multivariate generalization of B-splines

We can generalize the B-spline theory we developed so far to the d-dimensional case by using *tensor products*. We illustrate how we can compute the B-spline basis functions for the simplest generalization, i.e., $d = 2$. Let $X = (X_1, X_2) \in \mathbb{R}^2$, and $Y \in \mathbb{R}$. We can compute the B-spline vectors in X_1 and X_2 using Equations (2.4) & (2.5). We define,

$$\psi_1(X_1) = (\psi_{0,\rho}(X_1), \psi_{1,\rho}(X_1), \ldots, \psi_{q-1,\rho}(X_1))^T$$
$$\psi_2(X_2) = (\psi_{0,\rho}(X_2), \psi_{1,\rho}(X_2), \ldots, \psi_{q-1,\rho}(X_2))^T.$$

The *spline surface* that lies over the rectangle spanned by the knots can be constructed as,

$$\widehat{f}(X) = \widehat{f}(X_1, X_2) = \sum_{i_1=0}^{q-1} \sum_{i_2=0}^{q-1} \psi_{i_1,\rho}(X_1)\psi_{i_2,\rho}(X_2)a_{i_1,i_2},$$

where a_{i_1,i_2} is an element in the $q \times q$ coefficient matrix. We can write the B-spline basis matrix $\psi(X)$ as the tensor product of $\psi_1(X_1)$ and $\psi_2(X_2)$,

$$\psi(X) = \begin{pmatrix} \psi_{0,\rho}(X_1)\psi_{0,\rho}(X_2) & \psi_{0,\rho}(X_1)\psi_{1,\rho}(X_2) & \ldots & \psi_{0,\rho}(X_1)\psi_{q-1,\rho}(X_2) \\ \psi_{1,\rho}(X_1)\psi_{0,\rho}(X_2) & \psi_{1,\rho}(X_1)\psi_{1,\rho}(X_2) & \ldots & \psi_{1,\rho}(X_1)\psi_{q-1,\rho}(X_2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{q-1,\rho}(X_1)\psi_{0,\rho}(X_2) & \psi_{q-1,\rho}(X_1)\psi_{1,\rho}(X_2) & \ldots & \psi_{q-1,\rho}(X_1)\psi_{q-1,\rho}(X_2) \end{pmatrix}.$$

Let vec$(\cdot)$ denote the vectorization operator that stacks the columns of a matrix. The resulting spline surface then equals,

$$\widehat{f}(X) = \text{vec}(\psi(X))^T \text{vec}(a).$$

For simplicity we use the shorthand notation $\widehat{f}(X) = \psi(X)^T a$, and from the context (i.e. the dimensionality of the basis tensor) it should be clear that we consider vectorized tensor products. More generally, if $X = (X_1, ..., X_d)^T \in \mathbb{R}^d$, one can construct the d-dimensional hyper-surface from B-splines using the tensor product,

$$\widehat{f}(X) = \widehat{f}(X_1, ..., X_d) = \sum_{i_1=0}^{q-1} \sum_{i_2=0}^{q-1} \ldots \sum_{i_d=0}^{q-1} \psi_{i_1,\rho}(X_1)\psi_{i_2,\rho}(X_2) \cdot \ldots \cdot \psi_{i_d,\rho}(X_d)a_{i_1 i_2 \ldots i_d}.$$

For simplicity we write the B-spline vectors for $i = 1, ..., d$,

$$\psi_i(X_i) = (\psi_{0,\rho}(X_i), \psi_{1,\rho}(X_i), \ldots, \psi_{q-1,\rho}(X_i))^T.$$

The B-spline basis tensor $\psi(X)$ can then be written as the tensor product of these B-spline vectors,

$$\psi(X) = \psi_1(X_1) \otimes \ldots \otimes \psi_d(X_d).$$

The coefficients can be easily found in a similar manner as the one-dimensional case, and the minimizer of the MSE is the same as in Equation (2.7), using the vectorized tensor $\psi(X)$. If f has p bounded derivatives, Section 5.3 in Huang (2003) gives the following asymptotic bias,

$$\sup_x |E(\widehat{f}(x)|X_1, ..., X_n) - f(x)| = O_P(q_n^{-p}). \tag{2.8}$$

We refer the interested reader to de Boor (1976) and Hastie et al. (2001) for a more elaborate overview of B-splines.

Chapter 3

A Review of Sufficient Dimension Reduction Literature

In this chapter we give a formal introduction to the sufficient dimension reduction (SDR) paradigm. We state some useful properties and give the particular setting we consider in the following chapters. Some of the SDR literature is reviewed, and the article that motivated this thesis is discussed in detail. Lastly, we present our approach to the estimation of the central mean subspace.

3.1 Introduction

A *statistic* is a function $T = T(X_1, ..., X_n)$ of a random sample $X_1, ..., X_n$ from a density function p_θ parametrized by θ. We call this statistic *sufficient* for θ if the random sample $X_1, ..., X_n$ does not contain any additional information about the parameter θ, other than the statistic. Intuitively this means that any information about θ we can extract from our random sample, is already contained in our statistic T. The sufficient dimension reduction paradigm combines this idea with that of *dimension reduction*. Let $Y \in \mathbb{R}$ be the response variable, and let $X = (X_1, ..., X_p)$ be the p-dimensional covariates. A dimension reduction is a mapping from $\mathbb{R}^p$ to $\mathbb{R}^d$, with $d < p$. In SDR we consider linear combinations of our covariates. Consequently, we can write the dimension reduction as $\beta^T X$, with $\beta \in \mathbb{R}^{p \times d}$. In regression analysis our goal is to study the relationship between Y and the covariates X. One class of dimension reduction problems in regression analysis concern the distribution of Y conditional on X, i.e.

the distribution of $Y|X$. Here, we assume there exists a matrix β such that

$$P(Y \leq y|X) = P(Y \leq y|\beta^T X), \text{ almost surely.} \tag{3.1}$$

Note that if such a matrix β exists, then as far as the relation between Y and X is concerned, the covariates X can be replaced by the d-dimensional linear combinations $\beta^T X$. Typically d is much smaller than p. Thus the regression problem with a p-dimensional covariate is converted to one with a d-dimensional covariate. The problem is to correctly identify the column subspace of a matrix β that satisfies (3.1), with the smallest number of columns. We denote this space by $S_{Y|X}$, and it is called the *central subspace* (CS). Although the central subspace problem provides the most comprehensive description of the relationship between the response vector and covariates, sometimes only certain properties of the relation are of interest. In this thesis we are interested in the dependence of the mean of Y conditional on X, that is, $E(Y|X)$. In this setting, the assumption is weakened to,

$$E(Y|X) = E(Y|\beta^T X), \text{ almost surely,} \tag{3.2}$$

for a $p \times d$ matrix β. The column space of the matrix β with the lowest dimension satisfying relation (3.2) is called the *central mean subspace* (CMS) and is denoted by $S_{E(Y|X)}$. This setting has been generalized to the *central k-th moment subspace* by Yin and Cook (2002). The central k-th moment subspace $S^{(k)}_{Y|X}$ is the column space of a $p \times d$ matrix β with the smallest number of columns d satisfying,

$$E(Y^j|X) = E(Y^j|\beta^T X), \text{ almost surely, for } j = 1, ..., k.$$

3.2 Properties

Note that we can reformulate Equation (3.2) as

$$Y = E(Y|\beta^T X) + \epsilon, \text{ with } E(\epsilon|X) = 0.$$

Since we are estimating a space instead of the more classical statistical setting in which we estimate parameters, there are some caveats which we must account for, such as existence, uniqueness, and identifiability. The existence of the smallest SDR subspace is, except for some degenerate cases, guaranteed, and it is uniquely defined (Cook (2004)). Furthermore, if $Z = V^{-1}(X - u)$ for a symmetric invertible matrix V and any p-dimensional vector u, then

$$E(Y|(V\beta)^T Z) = E(Y|\beta^T V V^{-1}(X - u)) = E(Y|\beta^T X).$$

This implies the so-called *invariance property*, which was first coined by Cook (1995):

$$S_{E(Y|X)} = V^{-1} S_{E(Y|Z)}.$$

This property implies that we can center and normalize (i.e., assume that X has mean zero and variance I_p) the covariates without loss of generality.

Identifiability

It is easily seen that β is typically not identifiable. In particular, for any full rank dxd matrix A, βA will generate the same column space as β. Ma and Zhu (2013) adopted the local coordinate system of the Grassmann-manifold to resolve this problem. By restricting our set of matrices to all $p \times d$ matrices β of the form $\beta = (I_d, C^T)^T$, where I_d is the $d \times d$ identity matrix and the lower submatrix C has dimensions $(p-d) \times d$. Then any two matrices $\beta_1 A$ and $\beta_2 A$ are different if and only if the column space of β_1 and β_2 are different. Note that the first d components of the model must contain explanatory power in this parametrization. We can rotate the order of covariates to guarantee that the first d covariates play a role in the estimation procedure. The columns of C are the relative effects of $(X_{d_0+1}, ..., X_p)^T$ compared to X_j, for $j = 1, ..., d_0$. In this parametrization the CS and CMS are essentially objects with $(p-d)d$ degrees of freedom. It should be understood that whenever we use β in the context of SDR, we use the Grassman parametrization in the rest of this thesis, i.e. $\beta = (I_d, C^T)^T$ for a parameter matrix C.

3.3 Central Mean Subspace Literature

The central mean subspace is formalized in the following definition. Let $S(\beta)$ denote the column space of a matrix β.

Definition 3.1 *(Definition 1, Cook and Li (2002)) If $Y \perp\!\!\!\perp E(Y|X)|\beta^T X$, then $S(\beta)$ is a mean dimension-reduction subspace for the regression of Y on X.*

This is equivalent to condition (3.2) by Proposition 1 in Cook and Li (2002). The notion of the smallest dimension reduction is formalized in the following definition.

Definition 3.2 *Let $S_{E(Y|X)} = \cap_m S_m$, where S_m are all the mean dimension reduction subspaces. If $S_{E(Y|X)}$ is itself a mean dimension-reduction subspace, it is called the central mean subspace (CMS).*

As the definition above suggests, the existence of the CMS is *not* guaranteed. The intersection of two mean dimension-reduction subspaces does not necessarily form another dimension-reduction subspace. However when it does exist, it is clear that it must be the smallest DR subspace. Additionally the CMS is a subspace of the central subspace, because the latter contains *all* the information that X conveys about Y, whereas the former is only focused on the conditional mean. There exist rather mild conditions under which the central subspace exists, and under similar conditions the existence of the CMS is guaranteed as well. One such condition is that if the domain of X is open and convex, then the CMS exists and is unique (Cook and Li (2002)). In the rest of this thesis we assume that the CMS exists and is unique.

3.3.1 Estimation of the CMS

We briefly review three approaches to the estimation of $\mathcal{S}_{E(Y|X)}$: Inverse regression based methods, nonparametric methods, and semi-parametric methods. The first two are not reviewed in detail, and we refer the interested reader to Ma and Zhu (2013) for a more elaborate treatment of the these approaches. A large portion of the rest of this chapter is dedicated to the approach used in Huang and Chiang (2017). We then proceed to our own estimation procedure, which is heavily inspired by the latter.

Inverse Regression Based Methods

The idea behind inverse regression is to reverse the relation between the response variable Y and the covariates X (see Li (1991)). Instead of considering the expectation of Y conditional on X, we consider the expectations of X conditional on Y. Note that this circumvents the so-called curse of dimensionality when X is high dimensional, because Y is univariate. Often additional assumptions on the relationship between the covariates X and the lower-dimensional $\beta^T X$, which include the linearity condition,

$$E(X|\beta^T X) = PX,$$

for $P = \beta(\beta^T\beta)^{-1}\beta^T$. Some of the inverse regression based methods require a constant variance condition,

$$cov(X|\beta^T X) = Q,$$

for $Q = I - P$. These assumptions only need to hold at the true value of $\beta = \beta_0$, however since β_0 is unknown, they are often strengthened to hold for all possible β. One approach in the inverse regression for estimating the CMS is Principal Hessian Directions (Li (1992)).

Nonparametric Methods

The class of nonparametric estimation is based on estimating the column space of β through minimizing a criterion that describes the fit of the DR models to the observed data. One of these methods is the minimum average variance estimation method (MAVE, Xia et al. (2002)). In this setting we require that $\beta_0^T\beta_0 = I_{d_0}$. The MAVE procedure then minimizes the objective function,

$$E(var(Y|\beta^T X)) = E(Y - E(Y|\beta^T X))^2,$$

over all matrices β such that $\beta^T\beta = I_d$. A first order Taylor expansion is then used to approximate $E(Y|\beta^T X)$, and a kernel function is used to nonparametrically approximate the first derivative of the conditional mean.

Semiparametric Methods

In the semi-parametric paradigm, methods from semi-parametric statistics are applied. The likelihood of one observation (X, Y) equals,

$$\eta_1(X)\eta_2(Y - g(\beta^T X), X), \qquad g(\beta^T X) = E(Y|\beta^T X),$$

where η_1 is the marginal density, and η_2 is the conditional density function. These two functions, and g are viewed as nuisance parameters, and we are interested solely in β. In Ma and Zhu (2014) a class of functions called the *influence functions* are derived. The complete class of influence functions provide all the possible consistent estimators. The properties of these estimators are then studied through the influence functions. Semi-parametric theory is developed in Chapter 6, and we derive the nuisance tangent spaces. We also discuss how the influence functions could be found once the nuisance tangent spaces are known. Our approach however does not rely on the nuisance tangent space to *find* estimators, we define an estimator and then prove that desirable asymptotic properties hold *using* the nuisance tangent space.

3.4 Huang and Chiang's Method

The main objective of Huang and Chiang's (2017) paper is the estimation of the central subspace. For completeness we state some of the results derived by Huang and Chiang regarding this estimation procedure. Once we are familiar with the technique, we show that similar reasoning justifies the estimator for the CMS utilized in Chapter 6. The large sample properties are discussed in Chapter 5 and Chapter 6.

3.4.1 Central Subspace Estimation

Let $1(\cdot)$ denote the indicator function. Huang and Chiang's estimation procedure is based on the following equality,

$$E(1(Y \leq y)|X = x) = P(Y \leq y|X = x), \text{for all } (x, y).$$

Thus the central subspace estimation is equivalent to integrating the estimation of the CMS of $1(Y \leq y)$ over the domain of Y, which we denote as $\mathcal{Y}$. We introduce the following notation,

$$\begin{aligned}
&\beta_d = (I_d, C_d^T)^T, \text{ with } C_d \text{ a } (p-d) \times d \text{ matrix;}\\
&F_{C_d}(y|u) = P(Y \leq y|\beta_d^T X = u);\\
&F_Y(y) = P(Y \leq y);\\
&\widehat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} 1(Y_i \leq y);\\
&\langle f_1(\cdot), f_2(\cdot)\rangle_{L_2} = \int_{\mathcal{Y}} f_1(y) f_2(y) dF_Y(y), \text{ for any function } f_1, f_2;\\
&\|f\|_{L_2} = \sqrt{\langle f(\cdot), f(\cdot)\rangle_{L_2}}, \text{ for any functions } f.
\end{aligned}$$

Proposition 1 in Huang and Chiang (2017) motivates the choice of the semi-parametric framework, and highlights the role F_{C_d} plays in the estimation.

Proposition 3.1 (Proposition 1, Huang and Chiang (2017)) *For a given β_d, $F_{C_d}(y|u)$ minimizes the mean integrated square error (MISE),*

$$E\|1(Y \leq \cdot) - G(\cdot, \beta_d^T X)\|_{L_2}^2,$$

over all $(d+1)$-variate functions $G(y, u)$. Moreover the basis matrix β of a SDR subspace minimizes $E\|1(Y \leq \cdot) - F_{C_d}(\cdot|\beta_d^T X)\|_{L_2}^2$ over all β_d.

This implies it is sensible to use the conditional distribution $F_{C_d}(y|u)$ to estimate β_0. By the existence and uniqueness of the central subspace, and the second part of Proposition 1, it follows that

$$E\|1(Y \le \cdot) - F_{C_d}(\cdot|\beta_d^T X)\|_{L_2}^2 > E\|1(Y \le \cdot) - F(\cdot|\beta_0^T X)\|_{L_2}^2$$

for all β_d such that, $S(\beta_0) \nsubseteq S(\beta_d)$. However, note that we have,

$$E\|1(Y \le \cdot) - F_{C_d}(\cdot|\beta_d^T X)\|_{L_2}^2 \begin{cases} > E\|1(Y \le \cdot) - F(\cdot|\beta_0^T X)\|_{L_2}^2, \text{ if } S(\beta_d) \nsupseteq S(\beta_0) \\ = E\|1(Y \le \cdot) - F(\cdot|\beta_0^T X)\|_{L_2}^2, \text{ if } S(\beta_d) \supseteq S(\beta_0). \end{cases}$$

Consequently, this criterion fails to distinguish the true model from overfitted ones. This implies that when the true dimensions of the central subspace (denoted by d_0) is unknown, we can't simply use the sample analogue of the MISE to estimate d_0. Therefore we use a cross-validated version of the sample MISE to determine d_0 and estimate β_0. Huang and Chiang use leave one out cross-validation (LOOCV), where one estimates the conditional distribution $F_{C_d}(\cdot|\beta_d^T X)$ based on all data minus one observation, and considers left out data point one as a "new" observation. The error is then computed for this new observation, and this is performed for all individual data points. Let $\widehat{F}_{C_d}$ denote an estimator of F_{C_d}. This approach relies on the fact that for an independent "new" observation (X_0, Y_0) we have,

$$\begin{aligned} E\|1(Y_0 \le \cdot) - \widehat{F}_{C_d}(\cdot|\beta_d^T X_0)\|_{L_2}^2 &= E\|1(Y_0 \le \cdot) - F(\cdot|\beta_0^T X_0) + F(\cdot|\beta_0^T X_0) \\ &\quad - F_{C_d}(\cdot|\beta_d^T X_0) + F_{C_d}(\cdot|\beta_d^T X_0) - \widehat{F}_{C_d}(\cdot|\beta_d^T X_0)\|_{L_2}^2 \\ &= \sigma_0^2 + b_0^2(C_d) + E\|\widehat{F}_{C_d}(\cdot|\beta_d^T X_0) - F_{C_d}(\cdot|\beta_d^T X_0)\|_{L_2}^2 \\ &\quad + 2E\langle F_{C_d}(\cdot|\beta_d^T X_0) - F(\cdot|\beta_0^T X_0), \widehat{F}_{C_d}(\cdot|\beta_d^T X_0) - F_{C_d}(\cdot|\beta_d^T X_0)\rangle_{L_2}, \end{aligned}$$

where,

$$\begin{aligned} \sigma_0^2 &= E\|1(Y_0 \le \cdot) - F(\cdot|\beta_0^T X_0)\|_{L_2}^2, \text{ and} \\ b_0^2(C_d) &= E\|F_{C_d}(\cdot|\beta_d^T X_0) - F(\cdot|\beta_0^T X_0)\|_{L_2}^2. \end{aligned}$$

The terms that equal zero are omitted.* Note that $b_0^2(C_d) = 0$ if and only if $S(\beta_d) \supseteq S_{Y|X}$.

*These are the terms that include $1(Y_0 \le \cdot) - F(\cdot|\beta_0^T X_0)$, which equals zero in expectation, in the inner product.

The conditional distribution is estimated by the Nadaraya-Watson estimator,

$$\widehat{F}_{C_d}(y|u) = \frac{\sum_{i=1}^n N_{iy} K_{q,h_d}(\beta_d^T X_i - u)}{\sum_{i=1}^n K_{q,h_d}(\beta_d^T X_i - u)},$$

where $K_{q,h_d}(u) = \frac{1}{h_{dk}} \prod_{k=1}^d K_q(\frac{u_k}{h_{dk}})$, and positive-valued bandwidth vector $h_d = (h_{d1}, ..., h_{dd})^T$. The following *pseudo sum of integrated squares* (PSIS) inspired by Chiang and Huang (2012) is proposed as a sample analogue of the MISE,

$$\begin{aligned} PSIS(C_{d_0}) &= \frac{1}{n} \sum_{i=1}^n \int (1(Y_i \leq y) - \widehat{F}_{C_{d_0}}(y|\beta_{d_0}^T X_i))^2 d\widehat{F}_Y(y) \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n (1(Y_i \leq Y_j) - \widehat{F}_{C_{d_0}}(Y_j|\beta_{d_0}^T X_i))^2. \end{aligned}$$

The *LOOCV PSIS* is then defined as,

$$\begin{aligned} CV(d, C_d, h_d) &= \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{Y}} (1(Y_i \leq y) - \widehat{F}_{C_d}^{-i}(y|\beta_d^T X_i))^2 d\widehat{F}_Y(y) \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n (1(Y_i \leq Y_j) - \widehat{F}_{C_d}^{-i}(Y_j|\beta_d^T X_i))^2. \end{aligned} \tag{3.3}$$

The minimizer of this cross-validated PSIS over all possible $(p-d) \times d$-dimensional matrices and bandwidths, which is a function in d, converges in probability to a strictly convex function with unique minimizer d_0. As a result the estimation criterion for the central subspace and conditional density can be carried out by a forward algorithm, where in the PSIS d is fixed, and is minimized by starting with $d = 0$ and increased in each iteration until the first local minimum is reached. Since in each iteration of the double sum we need to recalculate the model, leaving out the i-th observation, computation of this criterion grows at rate $O(n^3)$.

3.4.2 Central Mean Subspace Estimation

We now derive the CV-criterion for estimating the central mean subspace along the same lines. We define the conditional expectation,

$$g_{C_d}(u) = E(Y|\beta_d^T X = u). \tag{3.4}$$

The mean squared error (MSE),

$$E(Y - g_{C_d}(\beta_d^T X))^2 \begin{cases} > E(Y - E(Y|\beta_0^T X))^2, & \text{if } S(\beta_d) \not\supseteq S(\beta_0) \\ = E(Y - E(Y|\beta_0^T X))^2, & \text{if } S(\beta_d) \supseteq S(\beta_0). \end{cases}$$

This implies the sample analogue of the MSE fails to distinguish the true model from overfitted ones. Consequently we use a CV criterion to determine d_0 and estimate β_0, equivalently to the estimation of the central subspace. For a "new" independent observation (X_0, Y_0) the MSE equals,

$$\begin{aligned} E(Y_0 - \widehat{g}_{C_d}(\beta_d^T X_0))^2 = E\Big(& Y_0 - E(Y|\beta_0^T X_0) + E(Y|\beta_0^T X_0) - g_{C_d}(\beta_d^T X_0) \\ & + g_{C_d}(\beta_d^T X_0) - \widehat{g}_{C_d}(\beta_d^T X_0)\Big)^2 \\ & = E(Y_0 - E(Y|\beta_0^T X_0))^2 + E(E(Y|\beta_0^T X_0) - g_{C_d}(\beta_d^T X_0))^2 \\ & + E(g_{C_d}(\beta_d^T X_0) - \widehat{g}_{C_d}(\beta_d^T X_0))^2 \\ & + 2E\Big((Y_0 - g_{C_d}(\beta_d^T X_0))(g_{C_d}(\beta_d^T X_0) - \widehat{g}_{C_d}(\beta_d^T X_0))\Big). \end{aligned}$$

The terms that equal zero are omitted.[†] The second term equals zero if and only if $S(\beta_d) \supseteq S_{E[Y|X]}$. In Remark 3 of Huang and Chiang (2017) it is mentioned that one can estimate (d_0, C_0) in the CMS setting with the following cross-validation criterion

$$CV(d, C_d) = \frac{1}{n}\sum_{i=1}^{n}(Y - \widehat{g}_{C_d}^{-i}(\beta_d^T X_i))^2. \tag{3.5}$$

In particular they use kernel regression to now estimate the conditional mean,

$$\widehat{g}_{C_d}(u) = \frac{\sum_{i=1}^{n} Y_i K_{q,h_d}(\beta_d^T X_i - u)}{\sum_{i=1}^{n} K_{q,h_d}(\beta_d^T X_i - u)}. \tag{3.6}$$

Note that the the LOOCV makes the estimation of the central (mean) subspace costly. This is in part due to the n estimations of the model, once for each omitted case. Computation of the CV-criterion in the CMS setting grows at rate $O(n^2)$. If we write the LOOCV statistic from Equation (3.5) as,

$$CV = \frac{1}{n}\sum_{i=1}^{n} e_{[i]}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_{[i]})^2,$$

[†] These are the terms which contain $Y_0 - E(Y|\beta_0^T X_0)$, which equals zero in expectation.

where $Y_{[i]}$ is the predicted value obtained when the model is estimated with the i-th data point deleted, i.e.

$$\widehat{Y}_{[i]} = \widehat{g}_{C_d}^{-i}(\widehat{\beta}_d^T X_i).$$

3.4.3 A novel approach for the CMS estimation

Using B-splines, we can use the following linear representation for $\widehat{Y}$,

$$\widehat{Y} = \psi(\widehat{\beta}_d^T X)^T \widehat{a},$$

for the vectorized B-spline basis tensor $\psi(\widehat{\beta}_d^T X)$, and a vector of coefficients $\widehat{a} = (\psi(\beta_d^T X)\psi(\beta_d^T X)^T)^{-1}\psi(\beta_d X)^T Y$ (see Equation (2.7)). We define the *hat matrix* H as,

$$H = \psi(\beta_d^T X)^T (\psi(\beta_d^T X)\psi(\beta_d^T X)^T)^{-1}\psi(\beta_d^T X). \tag{3.7}$$

Note that the estimate of Y[‡] equals,

$$\widehat{Y} = \psi(\beta_d^T X)^T \hat{a} = HY.$$

If we let $h_1, ..., h_n$ be the diagonal values of H, the LOOCV criterion can be computed using,

$$CV(\widehat{d}, \widehat{\beta}, \widehat{g}_{C_d}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{1-h_i}\right)^2, \tag{3.8}$$

where $e_i = Y_i - \widehat{Y}_i$, and $\widehat{Y}_i$ is the predicted value when all the data is used. Thus due to the linearity of the B-spline estimation, the computation of the CV-criterion in the CMS only grows at rate $O(n)$. The computation of the hat matrix is $O(n^3)$, but for moderate n it is still very fast. For a proof of Equation (3.8), we refer the reader to Seber and Lee (2012). Another possible approach is to split the data into K clusters, and perform K-fold cross-validation. Here one trains the data on the $K-1$ clusters of data, and then computes the error on the K-th cluster. This is then performed for each individual cluster. In Chapter 6 we prove that consistency holds for such an estimator.

[‡] Which is often referred to as Y-hat in vernacular.

Chapter 4

Empirical Processes and M-estimation

Following the introduction to the problem setting and our approach to the estimation of the central mean subspace, in this chapter we discuss the techniques we use in our proofs. The theory of empirical processes is a very powerful tool in the analysis of estimators. This exposition mostly consists of a collection of results we need later on, and some of the results that are especially noteworthy are discussed in length. Consequently, this exposition is rather superficial, and in no way does justice to the many useful results that come from this area of research. This exposition is based on parts of van der Vaart and Wellner (1996). In particular we highlight sections from Chapter 2 and 3 in Section 4.2 and 4.3, respectively, and omit the exposition of the modern weak convergence theory and many measure-theoretical details. Section 4.2 is mostly an exposition of *entropy calculations*, and in Section 4.3 we attempt to highlight how we can apply these in *M-estimation*.

4.1 Setup and Notation

Let $X_1, ..., X_n$ be i.i.d. copies of a random variable X taking values in a measurable space $(\mathcal{X}, \mathcal{A})$. Let $\mathcal{F}$ be a class of measurable real-valued functions defined on a measurable space $(S, \mathcal{S})$. We assume that for any $f \in \mathcal{F}$, the absolute value of the mean is finite, i.e., $E|f(X)| < \infty$. We define an *envelope function* of $\mathcal{F}$ as any function $F(x)$ such that $|f(x)| \leq F(x)$, for every x and $f \in \mathcal{F}$. The *minimal envelope function* is $x \mapsto \sup_{f \in \mathcal{F}} |f(x)|$. The expectation of a function $f \in \mathcal{F}$ under the probability measure P is written as,

$$Pf = E_P f(X) = \int_S f dP.$$

We denote the *empirical measure* as $\mathbb{P}_n$, which gives the expectation of f under the empirical measure:

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Note that $\mathbb{P}_n$ is an estimator of $P(f)$. In fact, by the strong law of large numbers (SLLN) for each $f \in \mathcal{F}$ we have,

$$\mathbb{P}_n f \overset{a.s.}{\to} Pf.$$

The centered and scaled version of the empirical measure is given by,

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Pf(X)).$$

If we assume that additionally $Pf^2 < \infty$, a direct consequence of the central limit theorem (CLT) is that,

$$\mathbb{G}_n f \overset{d}{\to} N(0, P(f - Pf)^2).$$

To use results in empirical process theory, we typically need a stronger result than the strong law of large numbers or the central limit theorem. In order to discuss a uniform version of the LLN or CLT we introduce the notation

$$\|Q\|_{\mathcal{F}} = \sup\{|Qf| : f \in \mathcal{F}\},$$

where Q is some probability measure.

4.2 Glivenko-Cantelli classes and Entropy

We define Glivenko-Cantelli and Donsker classes formally in the following two definitions.

Definition 4.1 *A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ is called P-Glivenko-Cantelli if*

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \overset{a.s.}{\to} 0.$$

Definition 4.2 *A class $\mathcal{F}$ of measurable functions is called P-Donsker if the sequence of processes $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges in distribution to a tight limit process in the space $l^\infty(\mathcal{F})$.*

The two definitions above are not very instructive, and don't provide us a clear approach to demonstrate these properties for a class of functions. However, it turns out that whether a class of functions is Glivenko-Cantelli or Donsker depends on the size of the class. We state these results in Theorem 4.1 and 4.2. We first require a notion of measuring the size of a function class. One such measure is formulated in terms of the *entropy* of a function class. The entropy is a relatively simple way to measure the size of a class $\mathcal{F}$. The ϵ-entropy of $\mathcal{F}$ is the logarithm of the number of "balls" or "brackets" of size ϵ needed to cover $\mathcal{F}$. This is formalized in the following definitions.

Definition 4.3 *The* ***covering number*** *$N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls $\{g : \|g - f\| < \epsilon\}$ of radius ϵ needed to cover the set $\mathcal{F}$. The* ***entropy (without bracketing)*** *is the logarithm of the covering number. The centers of the balls do not need to be in $\mathcal{F}$, but they should have finite norms.*

Definition 4.4 *Given two functions $l, u : \mathcal{X} \to \mathbb{R}$ the* ***bracket*** *$[l, u]$ is the set of all functions f with $l \leq f \leq u$. An ϵ-**bracket** relative to $\|\cdot\|$ is a bracket $[l, u]$ with $\|u - l\| < \epsilon$. The* ***bracketing number*** *$N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$. The* ***entropy with bracketing*** *is the logarithm of the bracketing number. The upper and lower bounds u and l of the brackets do not need to belong to $\mathcal{F}$, but they should have finite norms.*

Note that as ϵ tends to zero, the entropy grows to infinity. The most intuitive type of sets whose metric entropy we can consider are spheres in a Euclidean space.

Lemma 4.1 *(Lemma 5, Shen and Wong (1994)) Let S be a sphere of size r in $\mathbb{R}^q$ at the origin, that is,*

$$S = \{x = (x_1, ..., x_q) \in \mathbb{R}^q : \sum_{i=1}^{q} x_i^2 \leq r^2\}.$$

Then, for any $\epsilon < r$,

$$\begin{aligned} \log N_{[]}(\epsilon, S, \|\cdot\|_1) &\lesssim q \log\left(q^{1/2}\frac{r}{\epsilon}\right), \text{ and} \\ \log N_{[]}(\epsilon, S, \|\cdot\|_2) &\lesssim q \log\left(\frac{r}{\epsilon}\right). \end{aligned} \tag{4.1}$$

Sufficient conditions for a class to be Glivenko-Cantelli or Donsker can be given in terms of the rate at which the entropy grows as ϵ tends to zero. Let P be any probability measure. The the $L_r(P)$-metric is defined as,

$$\|f\|_{L_r(P)} = (P|f|^r)^{1/r} = \left(\int_\Omega |f(x)|^r dP(x)\right)^{1/r}.$$

The space of all functions that satisfy $\|f\|_{L_r(P)} < \infty$, equipped with the $L_r(P)$ norm, is called the $L_r(P)$-space. For the $L_r(P)$ norm, for any r, the *Riesz property* holds: for a pair of functions f, g, if $|f| \leq |g|$, then $\|f\|_{L_r(P)} \leq \|g\|_{L_r(P)}$. We can use the fact that if f is in the 2ϵ-bracket $[l, u]$, it is also in a ball of radius $(l+u)/2$, and thus,

$$N(\epsilon, \mathcal{F}, \|\cdot\|_{L_r(P)}) \leq N_{[]}(2\epsilon, \mathcal{F}, \|\cdot\|_{L_r(P)}). \tag{4.2}$$

There is no general converse inequality.

Theorem 4.1 *(Glivenko-Cantelli) Let $\mathcal{F}$ be a class of measurable functions such that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$. Then $\mathcal{F}$ is $P-$Glivenko-Cantelli.*

A similar result exists that relates a class being Donsker to the rate at which the integral of the entropy with bracketing grows as $\epsilon \downarrow 0$.

Theorem 4.2 *Let $\mathcal{F}$ be a class of measurable functions such that its bracketing integral defined as,*

$$J_{[]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2(P))} d\epsilon < \infty$$

for every $\epsilon > 0$. Then $\mathcal{F}$ is P-Donsker.

Determining the covering or bracketing number can be difficult without any additional information. In Section 2.7 in van der Vaart and Wellner (1996) some very useful results are offered, on for example classes of monotone functions, classes of convex functions over closed sets, or classes that are Lipschitz in a parameter. The last one proves particularly useful in our setting, as will be demonstrated repeatedly in our entropy calculations in Chapter 6.

Theorem 4.3 *(Theorem 2.7.11, van der Vaart and Wellner (1996)) Suppose that a function class $\mathcal{F} = \{f_t : t \in T\}$ is Lipschitz in the index parameter $t \in T$, i.e.*

$$|f_s(x) - f_t(x)| \leq d(s,t)F(x) \tag{4.3}$$

for some metric d on the index set T, fixed function F on the sample space and every x. Then, for any norm $\|\cdot\|$,

$$N_{[]}(2\epsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq N(\epsilon, T, d). \tag{4.4}$$

Additionally $(\text{diam}T)F$ is an envelope function for the class $\{f_t - f_{t_0} : t \in T\}$ for any fixed t_0. The parameters of interest in our model are the coefficients of the B-splines, and the parameter matrix. Since for fixed n both of these are Euclidean, the entropy numbers of the parameters could be calculated relatively easily. The bracketing number of our function class can then be bounded by the covering numbers of the parameters we calculate if they satisfy Equation (4.3).

4.3 M-estimation

In this section we attempt to give the reader a notion of how M-estimation (M for ″maximum″) relates to our problem, and how we apply the theory discussed in the previous section in our problem. In M-estimation we consider estimators $\widehat{\theta}_n$ that maximize an empirical criterion function $\mathbb{M}_n(\theta)$. In particular we focus on criterion functions that can be written as,

$$\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta,$$

for *loss functions* m_θ. Note that although it is typically omitted from notation, these are functions of the observed data. In order to obtain the limiting distribution of M-estimators, the following steps are typically performed:

1. Establish consistency of $\widehat{\theta}_n$ to θ_0.
2. Establish a rate of convergence of $\widehat{\theta}_n$ to θ_0.
3. Derive the limiting distribution of $\widehat{\theta}_n$.

The rest of this section is dedicated to the results we use to prove consistency and rate of convergence.

4.3.1 Consistency

We can demonstrate consistency by using the following result.

Theorem 4.4 *(Corollary 3.2.3, van der Vaart and Wellner (1996)) Let $\mathbb{M}_n$ be stochastic processes indexed by a metric space Θ, and let $M : \Theta \to \mathbb{R}$ be a deterministic function. Suppose that,*

$$||\mathbb{M}_n - M||_\Theta \to 0, \tag{4.5}$$

in outer probability, and that there exists a point θ_0 such that,

$$M(\theta_0) > \sup_{\theta \notin G} M(\theta) \tag{4.6}$$

for every open set G that contains θ_0. Then any sequence $\widehat{\theta}_n$, such that $\mathbb{M}_n(\widehat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - o_P(1)$, satisfies $\widehat{\theta}_n \to \theta_0$ in outer probability.

We emphasize the presence of condition (4.5). From our discussion in the previous section, and the assumption that our data is i.i.d., it is apparent that we need to perform entropy calculations in order to show that the class of loss functions $\{m_\theta : \theta \in \Theta\}$ is Glivenko-Cantelli.

4.3.2 Rate of Convergence

In the parametric setting the rate of convergence of an estimator to the ″truth″ is typically of order $\sqrt{n}$. We, however, consider M-estimators where the dimension of the parameter is not necessarily finite. Additionally, as we discussed in Chapter 2, in the nonparametric regression there is an implicit assumption that the ″truth″ is not contained in our parameter space, i.e. there is an estimation bias present. We can reduce this bias by allowing our model become ″richer″ as the sample size grows. Thus we consider *sieved M-estimators*, the sieves Θ_n being a sequence of subsets of the parameter space. In this setting the loss functions $m_{n,\theta}$ are dependent on the sieve. The estimators $\widehat{\theta}_n$ maximize $\mathbb{M}_n$ over the sieve Θ_n.

Corresponding to the criterion functions $\mathbb{M}_n$, we define *centering functions*, M_n. Typically these are taken to be the expectation of the loss function, but this is not a requirement. We then let θ_n be the maximizer of the centering function, paralleling the maximization of $\mathbb{M}_n$. In this paradigm it is then reasonable to want the estimators $\widehat{\theta}_n$ be as close as possible to the maximizer θ_n of the centering functions M_n. We can think of $\|\theta_n - \theta_0\|$ as the distance of the sieve Θ_n to θ_0. If Θ_n is a Euclidean space, and we assume M_n is at least twice continuously differentiable at θ_n with non-singular second-derivative matrix, we can use Taylor expansions around θ_0 to find,

$$\begin{aligned} M_n(\theta) - M_n(\theta_n) &= M_n(\theta_n) + M_n'(\theta_n)(\theta - \theta_n) + \frac{1}{2}M_n''(\theta_n)(\theta - \theta_n)^2 \\ &+ O(\|\theta - \theta_n\|^3) - M_n(\theta_n) \\ &= M'(\theta_n)(\theta - \theta_n) + \frac{1}{2}M''(\theta_n)(\theta - \theta_n)^2 \\ &+ O(\|\theta - \theta_n\|^3). \end{aligned} \tag{4.7}$$

Since θ_n maximizes M_n, the first derivative must vanish at θ_n and the second derivative should be negative. Thus it is reasonable to require,

$$M_n(\theta) - M_n(\theta_n) \leq -d_n^2(\theta, \theta_n),$$

where d_n is an arbitrary non-negative map on Θ_n.

Theorem 4.5 *(Theorem 3.4.1, van der Vaart and Wellner (1996)) For each n, let $\mathbb{M}_n$ and M_n be stochastic processes indexed by a set Θ. Let $\theta_n \in \Theta$ (possibly random) and $0 \leq \delta_n < \eta$ be arbitrary, and let $\theta \mapsto d_n(\theta, \theta_n)$ be an arbitrary map (possibly random) from Θ to $[0, \infty)$. Let $[\cdot]^+ = \max\{0, \cdot\}$. Suppose that, for every n and $\delta_n < \delta \leq \eta$*

$$\sup_{\delta/2 < d_n(\theta,\theta_n) \leq \delta, \theta \in \Theta_n} M_n(\theta) - M_n(\theta_n) \leq -\delta^2,$$

$$E \sup_{\delta/2 < d_n(\theta,\theta_n) \leq \delta, \theta \in \Theta_n} \sqrt{n}[(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_n)]^+ \lesssim \phi_n(\delta), \tag{4.8}$$

for functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing on (δ_n, η), for some $\alpha < 2$. Let $r_n \lesssim \delta_n^{-1}$ satisfy

$$r_n^2 \phi_n\left(\frac{1}{r_n}\right) \leq \sqrt{n}, \qquad \textit{for every } n.$$

If the sequence $\widehat{\theta}_n$ takes its values in Θ_n and satisfies $\mathbb{M}_n(\widehat{\theta}_n) \geq \mathbb{M}_n(\theta_n) - O_P(r_n^{-2})$ and $d_n(\widehat{\theta}_n, \theta_n)$ converges to zero in outer probability, then $r_n d_n(\widehat{\theta}_n, \theta_n) = O_P(1)$. If the displayed conditions are valid for $\eta = \infty$, then the condition that $\widehat{\theta}$ is consistent is unnecessary.

Additionally the distance of the estimator sequence $\widehat{\theta}_n$ to the true function satisfies $d(\widehat{\theta}_n, \theta_0) = O_P(r_n^{-1}) + d(\theta_n, \theta_0)$ under the conditions above. Using small sieves Θ_n leads to a small *modulus of continuity* $\phi_n(\delta)$ of the centered processes $\sqrt{n}(\mathbb{M}_n - M)$ over Θ_n, and hence faster rates r_n, but then the distance of the true parameter θ_0 to the sieve will be large. This relation is reminiscent of the bias-variance trade-off, which must be balanced to obtain an optimal rate of convergence. Intuitively, this theorem gives us the rate at which the estimator $\widehat{\theta}_n$ converges to the best* estimator in a sieve θ_n. In a formula, it is the first term on the right hand side in the following display,

$$d(\widehat{\theta}_n, \theta_0) \leq d(\widehat{\theta}_n, \theta_n) + d(\theta_n, \theta_0). \tag{4.9}$$

*In the sense that it maximizes the criterion $Pm_n(\theta)$.

Thus the theorem gives us a rate of the variance, and if the bias is known, we can compute the rate of convergence of $\widehat{\theta}_n$ to θ_0.

The main challenge is to derive $\phi_n(\delta)$, the maximal inequality for the modulus of the centered processes in Equation (4.8). This involves the function class,

$$\mathcal{M}_{n,\delta} = \left\{ m_{n,\theta} - m_{n,\theta} : \theta \in \Theta_n, \frac{\delta}{2} < d_n(\theta, \theta_n) \leq \delta \right\}.$$

If $\mathcal{M}_{n,\delta}$ has a measurable envelope function $M_{n,\delta}$, we can use the following result,

$$E_P \|\mathbb{G}_n\|_{\mathcal{M}_{n,\delta}} \lesssim J_{[]}(1, \mathcal{M}_{n,\delta}, L_2(P))(PM^2_{n,\delta})^{1/2}. \tag{4.10}$$

Equation (4.10) is a consequence of Theorem 2.14.2 in van der Vaart and Wellner (1996).

Chapter 5

Asymptotic Results for Parametric CMS estimation

Following the exposition of empirical process theory, in this chapter we demonstrate how we can apply the discussed techniques. In particular we perform the entropy calculations, and show how one can demonstrate consistency for M-estimators. We consider a simplistic toy model, where the function that relates Y and $\beta_0^T X$ is known. The result is a parametric model estimation. This chapter is mostly meant for illustration purposes. Interestingly enough, the limiting distribution and rate of convergence we obtain coincide with the results derived in Huang and Chiang (2017).

5.1 Problem formulation

Suppose we observe iid random copies $(X_i, Y_i); i = 1, ..., n$ of (X, Y) drawn from,

$$Y = g(\beta_0^T X) + \epsilon, \qquad \beta_0 = (I_{d_0}, C_0^T)^T,$$

where g is a known function, and $E(\epsilon|X) = 0$. Since d_0 is assumed known, we write d for simplicity. The parameter of interest is then the $(p-d)d$ real-valued vector $\text{vec}(C_0)$. Note that the problem setting is equivalent to assuming g belongs to a parametric family of d-variate (regression) functions. We estimate the parameter using least squares,

$$\widehat{C}_n = \underset{C \in \mathbb{R}^{(p-d)\times d}}{\arg\max} -\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(\beta^T X_i))^2, \qquad \beta = (I_d, C^T)^T. \tag{5.1}$$

In the context of M-estimation, the criterion function can be written as,

$$\mathbb{M}_n(\beta) = \mathbb{P}_n m_\beta,$$

with loss functions $m_\beta(X,Y) = -(Y - g(\beta^T X))^2$, and $\beta = (I_d, C^T)^T$.

5.2 Asymptotic properties

In order to demonstrate consistency of $\text{vec}(\widehat{C}_n)$ to $\text{vec}(C_0)$, we need to show that the class of loss functions,

$$\mathcal{M} = \{m_\beta : \beta = (I_d, C^T)^T, C \in \mathbb{R}^{(p-d)\times d}\},$$

is Glivenko-Cantelli. We prove this by confirming that the function class satisfies the requirements in Theorem 4.3. The parameter space has a one to one transformation to the Euclidean space $\mathbb{R}^{(p-d)d}$. From Lemma 4.1 we know that the entropy number of spheres in $\mathbb{R}^{(p-d)d}$ is bounded if the coefficients are bounded. Consequently it makes sense to require that the true parameter is contained in a sphere, i.e.

$$\|\text{vec}(C_0)\|_2 \leq r, \text{for some non-negative constant } r. \tag{5.2}$$

Additionally we assume that the function g is Lipschitz in the parameters C, i.e. for all $C_1, C_2 \in \mathbb{R}^{(p-d)\times d}$,

$$|g(\beta_1^T X) - g(\beta_2^T X)| \leq K \cdot \|\text{vec}(C_1) - \text{vec}(C_2)\|_2, \tag{5.3}$$

for some $K > 0$, and $\beta_1 = (I_d, C_1^T)^T$, $\beta_2 = (I_d, C_2^T)^T$. Lastly, we require that the true parameter can be identified, i.e.,

$$Pm_{\beta_0} > Pm_\beta, \tag{5.4}$$

whenever $\beta_0 \neq \beta$. Note that this only means that the truth β_0 maximizes the loss function in expectation, and since we estimate β_0 by maximizing the empirical mean of the loss function, this assumption is reasonable.

Lemma 5.1 *If conditions (5.2)-(5.4) hold, and g is a non-constant bounded function, then*

$$\|vec(\widehat{C}_n) - vec(C_0)\|_2 \xrightarrow{P} 0.$$

Proof. Let $M = Pm_\beta$. We show that the conditions of Theorem 4.4 hold, starting with the uniform convergence property over the function class $\mathcal{M}$. By Theorem 4.1 it is sufficient to demonstrate that,

$$N_{[]}(\epsilon, \mathcal{M}, L_1(P)) < \infty \text{ for every } \epsilon > 0.$$

We calculate the bracketing number by showing that condition (4.3) holds. Let C_1, C_2 be parameter matrices in $\mathbb{R}^{(p-d)\times d}$, and $\beta_1 = (I_d, C_1^T)^T$, and $\beta_2 = (I_d, C_2^T)^T$, with $C_1, C_2 \neq C_0$. For any X, Y drawn from our model,

$$\begin{aligned} |m_{\beta_2}(X,Y) - m_{\beta_1}(X,Y)| &= |2Y(g(\beta_2^T X) - g(\beta_1^T X)) + g(\beta_1^T X)^2 - g(\beta_2^T X)^2| \\ &= |(2Y - g(\beta_1^T X) - g(\beta_2^T X))(g(\beta_2^T X) - g(\beta_1^T X))| \\ &\leq |2Y - g(\beta_1^T X) - g(\beta_2^T X)||g(\beta_2^T X) - g(\beta_1^T X)|. \end{aligned}$$

We use condition (5.3) to write,

$$|m_{\beta_2}(X,Y) - m_{\beta_1}(X,Y)| \leq K|2Y - g(\beta_1^T X) - g(\beta_2^T X)| \cdot \|\text{vec}(C_2) - \text{vec}(C_1)\|_2.$$

In conclusion,

$$|m_{\beta_2}(X,Y) - m_{\beta_1}(X,Y)| \leq F(X,Y) \cdot \|\text{vec}(C_1) - \text{vec}(C_2)\|_2,$$

with

$$F(X,Y) = \sup_{\beta_1,\beta_2} K|2Y - g(\beta_1^T X) - g(\beta_2^T X)|.$$

Let $B^q(u,r)$ denote a sphere in $\mathbb{R}^q$ with radius r centered at u. From Theorem 4.3 it follows that,

$$N_{[]}(2\epsilon\|F\|_{L_1(P)}, \mathcal{M}, L_1(P)) \leq N(\epsilon, B^{(p-d)d}(0,r), \|\cdot\|_2).$$

By the assumption that $E|Y| < \infty$ and g is a bounded function, it follows that

$$\|F\|_{L_1(P)} < \infty.$$

Additionally since g is non-constant, we have $\|F\|_{L_1(P)} > 0$. In conclusion, by Lemma 4.1, for any $\epsilon' = \frac{\epsilon}{2\|F\|_{L_1(P)}} > 0$,

$$N_{[]}(\epsilon', \mathcal{M}, L_1(P)) \leq (p-d)d \log\left(\frac{r}{\epsilon'}\right) < \infty. \tag{5.5}$$

By Equation (5.4)

$$M(\beta_0) > M(\beta), \text{ for all } C \neq C_0. \tag{5.6}$$

Since $\widehat{\beta}_n = (I_d, \widehat{C}_n)$ is chosen such that it maximizes $\mathbb{M}_n$, we have,

$$\mathbb{M}_n(\widehat{\beta}_n) \geq \sup_{C:\beta=(I_d,C^T)^T} \mathbb{M}_n(\beta) - o_P(1). \tag{5.7}$$

In conclusion, by Equations (5.5)-(5.7),

$$\text{vec}(\widehat{C}_n) \xrightarrow{P} \text{vec}(C_0).$$

□

Now that we have shown consistency of $\text{vec}(\widehat{C}_n)$ for $\text{vec}(C_0)$, we can use the following result to demonstrate the rate of convergence of and limiting distribution.

Theorem 5.1 *(Theorem 5.23, van der Vaart (1998)) For each θ in an open subset of a Euclidean space let $x \mapsto m_\theta(x)$ be a measurable function such that $\theta \mapsto m_\theta(x)$ is differentiable at θ_0 for P-almost every x with derivative $m'_{\theta_0}(x)$ and such that, for every θ_1 and θ_2 in a neighbourhood of θ_0 and a measurable function $\tilde{m}$ with $P\tilde{m}^2 < \infty$,*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \tilde{m}(x)\|\theta_1 - \theta_2\|. \tag{5.8}$$

Furthermore assume that the map $\theta \mapsto Pm_\theta$ admits a second-order Taylor expansion at a point of maximum θ_0 with nonsingular symmetric second derivative matrix V_{θ_0}. If $\mathbb{P}_n m_{\widehat{\theta}_n} \geq \sup_\theta \mathbb{P}_n m_\theta - o_P(n^{-1})$ and $\widehat{\theta}_n \xrightarrow{P} \theta_0$, then,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m'_{\theta_0}(X_i) + o_P(1).$$

In particular the sequence $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $V_{\theta_0}^{-1} P m'_{\theta_0} m'^T_{\theta_0} V_{\theta_0}^{-1}$.

It is fairly straight-forward to confirm that the conditions of this theorem hold.

Theorem 5.2 *If conditions (5.5)-(5.7) hold, $\|g\|_{L_2(P)} < \infty$, $E(\epsilon^2|X) < \infty$, and the loss function m admits a second-order Taylor expansion at $vec(C_0) \mapsto m_{\beta_0}$ with nonsingular symmetric second derivative matrix, then*

$$\sqrt{n}(vec(\widehat{C}_n) - vec(C_0)) \to N(0, A_0) \qquad as\ n \to \infty,$$

where

$$A_0 = \{Pg'(\beta_0^T X)^{\otimes 2}\}^{-1} \cdot [P((Y - g(\beta_0^T X))g'(\beta_0^T X))]^{\otimes 2} \cdot \{Pg'(\beta_0^T X)^{\otimes 2}\}^{-1}$$

Proof. Note that condition (5.8) has already been demonstrated to hold in our proof of the consistency, with,

$$F(X, Y) = \sup_{\beta_1, \beta_2} K|2Y - g(\beta_1^T X) - g(\beta_2^T X)|.$$

The second moment of F equals,

$$\begin{aligned}
PF^2 &= K^2 \int_{\mathcal{X},\mathcal{Y}} \sup_{\beta_1,\beta_2} |2y - g(\beta_1^T x) - g(\beta_2^T x)|^2 dP(x,y) \\
&= K^2 \int_{\mathcal{X},\mathcal{Y}} \sup_{\beta_1,\beta_2} |4y^2 - 4y(g(\beta_1^T x) - g(\beta_2^T x)) + (g(\beta_1^T x) - g(\beta_2^T x))^2| dP(x,y) \\
&\le K^2 \int_{\mathcal{X},\mathcal{Y}} \sup_{C_1,C_2} 4|y|^2 + 4K|y| \cdot \|\text{vec}(C_1) - \text{vec}(C_2)\|_2 \\
&+ 4K^2 \|\text{vec}(C_1) - \text{vec}(C_2)\|_2^2 dP(x,y) \\
&\le 4K^2 E(Y^2) + 4K^3 r E|Y| \cdot + 8K^4 r^2,
\end{aligned}$$

where we used that both $\text{vec}(C_1)$, $\text{vec}(C_2)$ lie in a sphere of radius r centered around the origin, thus their distance is bound by $2r$. By the assumption that $E(\epsilon^2|X) < \infty$, it follows that $E(Y^2) < \infty$. Thus PF^2 is finite. The estimator $\text{vec}(\widehat{C}_n)$ is assumed to maximize $\mathbb{M}_n$, $\text{vec}(\widehat{C}_n) \xrightarrow{P} \text{vec}(C_0)$ by Lemma 5.1, thus the conditions of Theorem 5.1 are satisfied.

We compute,

$$\begin{aligned}
Pm'(\beta) &= 2P((Y - g(\beta^T X))g'(\beta^T X)), \\
Pm''(\beta) &= P(-2g'(\beta^T X)^{\otimes 2} + 2(Y - g(\beta^T X))g''(\beta^T X)).
\end{aligned}$$

By conditioning on X and using that $E(Y - g(\beta_0^T X)|X) = E(\epsilon|X) = 0$, the second derivative matrix at the true parameter β_0 equals,

$$Pm''(\beta_0) = -2Pg'(\beta_0^T X)^{\otimes 2}.$$

The covariance matrix then equals,

$$-\frac{1}{2}\{Pg'(\beta_0^T X)^{\otimes 2}\}^{-1} \cdot 4P((Y - g(\beta_0^T X))g'(\beta_0^T X))^{\otimes 2} \cdot -\frac{1}{2}\{Pg'(\beta_0^T X)^{\otimes 2}\}^{-1}.$$

Thus we conclude that $\sqrt{n}(\text{vec}(\widehat{C}) - \text{vec}(C_0)) \xrightarrow{d} N(0, A_0)$. □

The covariance we obtain coincides with the covariance obtained in Huang and Chiang (2017) for the CMS estimator. Interestingly, the nonparametric estimation of the function g does not seem to affect the performance for large n. In the next chapter we see that this same type of invariance holds for our estimation procedure in the setting where g is unknown.

Chapter 6

Semiparametric Inference

In this chapter we investigate the asymptotic behaviour of our estimation technique when the function g is unknown. We initially assume that the dimension d_0 of the column space of C_0 is known, and derive a rate of convergence and limiting distribution. In Section 6.6 we discuss an efficient method to estimate the dimension d_0. Models where the parameters of interest are finite dimensional, and there is an infinite-dimensional *nuisance parameter* present, are semiparametric models.

Due to the necessity of setting up the theoretical framework, this chapter consists of two parts. Sections 6.1 to 6.3 consist of an exploration in semiparametric theory. These techniques are subsequently applied in Sections 6.4 to 6.6. Much of semiparametric theory is developed from a geometric perspective, and therefore we opted to include a brief review of Hilbert spaces in Section 6.1. In Section 6.2 we review semiparametric theory, introduce some important notions and give general exposition. We document the general framework and some of the approaches in semiparametric statistics. Moreover it consists of some technical detail, e.g. the introduction of influence functions, score functions, and tangent sets. We only treat these in a somewhat superficial manner, with the main goal being to expose the reader to the ideas. In Section 6.3 we explain our approach to prove the $\sqrt{n}$-consistency and asymptotic normality of our estimator, for which we apply Theorem 1 in Ma and Kosorok (2005). To apply this result, we need to determine the *nuisance tangent space*, which we discuss in some length in Section 6.2.

In Section 6.4 we derive the nuisance tangent space of our model. In order to prove the requirements of the result in Section 6.3, we rely heavily on the empirical process theory introduced in Chapter 4. In Section 6.6 we discuss a consistent estimator for the true dimension d_0. Section 6.1 and 6.2 are based on the first four chapters in Tsiatis (2006). Section 6.3 is based on Chapter 21 in Kosorok (2008).

6.1 Geometry of Hilbert spaces

In this section we define and briefly review some basic concepts related to the theory of Hilbert spaces. We are not including a full development of the theory, but give a sketch of some of the ideas and notions, and state some results without proof. By defining Hilbert spaces, we allow ourselves to have a notion of distance, angles, and as a consequence, orthogonality between vectors in a linear vector space. For this, we require an inner product.

Definition 6.1 *(Inner product spaces). A vector space $\mathcal{H}$ is an* ***inner product space*** *if there is an inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ which satisfies*

1. $\langle h_1, h_2 \rangle = \langle h_2, h_1 \rangle$ *for all* $h_1, h_2 \in \mathcal{H}$,
2. $\langle h_1 + h_2, h_3 \rangle = \langle h_1, h_3 \rangle + \langle h_2, h_3 \rangle$ *for all* $h_1, h_2, h_3 \in \mathcal{H}$,
3. $\langle a h_1, h_2 \rangle = a \langle h_1, h_2 \rangle$ *for all* $a \in \mathbb{R}$ *and* $h_1, h_2 \in \mathcal{H}$,
4. $\langle h, h \rangle \geq 0$ *for all* $h \in \mathcal{H}$ *and* $\langle h, h \rangle = 0$ *iff* $h = 0$.

We say that two elements $h_1, h_2 \in \mathcal{H}$ are *orthogonal* if $\langle h_1, h_2 \rangle = 0$. Note that any inner product space is a normed linear space with norm (or "length") induced by the inner product, i.e. $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$, which is the distance between a point and the origin in $\mathcal{H}$. A normed linear space $\mathcal{H}$ is called *complete* if every Cauchy sequence in $\mathcal{H}$ has a limit point in $\mathcal{H}$, or equivalently, every Cauchy sequence is convergent and has its limit in $\mathcal{H}$.

Definition 6.2 *An inner product space $\mathcal{H}$ equipped with inner product $\langle \cdot, \cdot \rangle$ that is complete with respect to the norm $\sqrt{\langle \cdot, \cdot \rangle}$ is a* ***Hilbert Space****.*

Let $\mathcal{U}$ be a (non-empty) linear subspace that is closed. Theorem 2.1 (the projection theorem) in Tsiatis (2006) shows that for every element h in $\mathcal{H}$, there exists a unique $u_0 \in \mathcal{U}$ that is closest to h, i.e.

$$\|h - u_0\| \leq \|h - u\| \text{ for all } u \in \mathcal{U},$$

and that $\langle h - u_0, u \rangle = 0$ for all $u \in \mathcal{U}$. When an element (or set) is orthogonal to an entire set $\mathcal{U}$, we say $h - u_0$ is orthogonal to $\mathcal{U}$. Then u_0 is called *the projection of h onto the space $\mathcal{U}$*, and this is denoted as $u_0 = \Pi(h|\mathcal{U})$.

Perhaps the most notorious property (if we forget the Pythagorean theorem exists for a second) of Hilbert spaces is Hölder's inequality. This will be used extensively in Section 6.5.

Theorem 6.1 *(Hölder's inequality) If $p, q \in [1, \infty]$, with $1/p + 1/q = 1$, then*

$$\|fg\|_1 \leq \|f\|_p \|g\|_q,$$

for all measurable real- or complex-valued functions f, g on a measure space. We interpret $1/\infty$ as zero in this context.

For $p = q = 2$ we have the Cauchy-Schwartz inequality. More generally, let $r \in (0, \infty]$, and $p_1, ..., p_n \in (0, \infty]$ such that,

$$\sum_{k=1}^{n} \frac{1}{p_k} = \frac{1}{r}.$$

Then, for all measurable real- or complex-valued functions $f_1, ..., f_n$,

$$\left\| \prod_{k=1}^{n} f_k \right\|_r \leq \prod_{k=1}^{n} \|f_k\|_{p_k}.$$

An example of a Hilbert space is the space of all square-integrable functions, the L_2-space, which we saw previously in Chapter 3 and have used in some entropy calculations. We give some definitions regarding operations on linear subspaces.

Definition 6.3 *Let M and N be two linear subspaces in $\mathcal{H}$. The direct sum of two linear subspaces, denoted by $M \oplus N$ is a linear subspace in $\mathcal{H}$, if every element $x \in M \oplus N$ has a unique representation of the form $x = m + n$, where $m \in M$ and $n \in N$.*

Definition 6.4 *The set of elements of a Hilbert space $\mathcal{H}$ that are orthogonal to a linear subspace M is denoted by $M^\perp$. The orthogonal complement $M^\perp$ is also a linear subspace and the entire Hilbert space is the direct sum of these two spaces, i.e.*

$$\mathcal{H} = M \oplus M^\perp.$$

The *closure* of a set S is defined as the smallest closed set that contains S. We denote this closure as $\bar{S}$. This means that all the limit points of S are in $\bar{S}$. These limits are defined in terms of the metric induced by the norm,

$$d(h_1, h_2) = \|h_1 - h_2\|.$$

L_2 spaces

In Chapter 4 we introduced the general $L_r(P)$-spaces. The $L_2(P)$-space is a Hilbert space, and it is the only value of r for which this is the case. Let the triple $(\mathcal{X}, \mathcal{A}, P)$ be a measure space, and let $\mathcal{F}$ be the set of all real-valued P-measurable functions on $\mathcal{X}$. Recall that,

$$L_r(P) = \{f \in \mathcal{F} : \int |f(x)|^r dP(x) < \infty\}.$$

For a probability measure P this equals the class of all random variables on the probability space $(\mathcal{X}, \mathcal{A}, P)$ that have finite r-th moment. This defines a normed linear space over $\mathbb{R}$ with the following norm:

$$\|f\|_{L_r(P)} = \left(\int |f(x)|^r dP(x) \right)^{1/r}.$$

The metric induced by this norm is $d(f, g) = \|f - g\|_{L_r(P)}$. Note that this normed linear space does not meet the uniqueness properties discussed in the section above. For $X \in L_2(P)$, $X = 0$ has a different meaning than usual. To retain the uniqueness of zero elements, we identify all random variables X, Y in $L_2(P)$ that are equal on a set of measure 1 with the same equivalence class. As such, we consider them to be representative of the same random variable. More formally, we define $\sim$ as the equivalence class such that:

$$f \sim g \text{ iff } f = g \text{ on } \mathcal{X} \setminus E, \text{ for } E \subset \mathcal{X} \text{ and } P(E) = 0.$$

It is easily shown that this indeed defines an equivalence relation. The space of equivalence classes $[f]$ in $L_2(P)$ form a Hilbert space, and it is understood that we generally mean equivalence classes when we speak of elements in $L_2(P)$, and uniqueness of elements means *uniqueness up to equivalence classes.* We can denote any of these equivalence classes with an arbitrary element from the equivalence class, because every element has the same integral (since they differ only on sets of measure zero).

6.2 Semiparametric Theory

6.2.1 Asymptotically linear estimators

A *statistical model* is a collection of probability measures,

$$\{P \in \mathcal{P}\},$$

on a sample space $\mathcal{X}$. Statistical models are frequently indexed by a parameter $\phi \in \Phi$, for some parameter space Φ. In parametric models Φ is finite-dimensional. Assume that $\{P \in \mathcal{P}\}$ is dominated by a measure μ. Semiparametric models are statistical models where ϕ has one or more infinite-dimensional components. Let $X_1, ..., X_n$ be i.i.d. instances of a random vector X sampled from a density p_X. We assume the density of X belongs to the class,

$$\{p_X(x; \phi), \phi = (\theta, \eta) \in \Phi\}.$$

The parameter of interest is the d-dimensional parameter θ, and the (possibly infinite-dimensional) nuisance parameter is η. An estimator $\widehat{\theta}_n$ of θ is a d-dimensional measurable function of $X_1, ..., X_n$. Many such estimators fall under the class of *asymptotically linear* estimators.

Definition 6.1 *An estimator $\widehat{\theta}_n$ of θ, for $\theta \in \mathbb{R}^d$ is asymptotically linear if there exists a function $\psi : \mathcal{X} \to \mathbb{R}^d$ such that $E\psi = 0$, and*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \sqrt{n} \sum_{i=1}^{n} \psi(X_i) + o_P(1),$$

and $E(\psi\psi^T)$ is finite and nonsingular.

Here $E\psi$ and $E(\psi\psi^T)$ are defined w.r.t. to the distribution of X. The random vector $\psi(X_i)$ is the i-th *influence function* of the estimator $\widehat{\theta}_n$ - and as the name suggests, it is the influence of the i-th observation on the estimator $\widehat{\theta}_n$. The following example is due to Tsiatis (2006).

Example 1 ***(Tsiatis (2006), Example 1)*** *Suppose $X_1, ...X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The MLE for μ and σ^2 are given by*

$$\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \text{ and}$$
$$\widehat{\sigma}_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \widehat{\mu}_n)^2.$$

The estimator $\widehat{\mu}_n$ is asymptotically linear with $\psi(X_i) = (X_i - \mu_0)$, since

$$\sqrt{n}(\widehat{\mu}_n - \mu_0) = \sqrt{n} \sum_{i=1}^{n} (X_i - \mu_0).$$

We can express the estimator for the variance as,

$$\sqrt{n}(\widehat{\sigma}_n^2 - \sigma_0^2) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[(X_i - \mu_0)^2 - \sigma_0^2 \right] + \sqrt{n}(\widehat{\mu}_n - \mu_0)^2.$$

Since $\sqrt{n}(\widehat{\mu}_n - \mu_0)$ converges to a normal distribution with mean zero, and furthermore $\widehat{\mu}_n - \mu_0$ converges in probability to zero, it follows that $n^{1/2}(\widehat{\mu}_n - \mu_0)^2$ converges in probability to zero. The influence function here equals $\psi(X_i) = ((X_i - \mu_0)^2 - \sigma_0^2)$.

By the CLT we have,

$$n^{-1/2} \sum_{i=1}^{n} \psi(X_i) \xrightarrow{D} N(0, E(\psi\psi^T)),$$

and by Slutsky's Theorem,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{D} N(0, E(\psi\psi^T)).$$

Consequently, when studying asymptotic properties of $\widehat{\theta}_n$, it suffices to consider the influence function of $\widehat{\theta}_n$. As demonstrated in Theorem 3.1 in Tsiatis (2006), an asymptotically linear estimator almost surely has a unique influence function.

6.2.2 Superefficient estimators

The variance of any unbiased estimator $\widehat{\theta}_n$ of θ must be greater than or equal to the *Cràmer-Rao lower bound*, which is defined as the inverse of the *Fisher information*, i.e.

$$\text{var}_\theta(\widehat{\theta}_n) \geq \frac{1}{n} I(\theta)^{-1},$$

where the Fisher information $I(\theta)$ is defined as,

$$I(\theta) = E_\theta \left(\frac{\partial l(x;\theta)}{\partial \theta^2} \right)^2.$$

Here l is the natural logarithm of the likelihood function for a single observation x, i.e. $l(x;\theta) = \ln p_X(x;\theta)$. The *efficiency* of an unbiased estimator $\widehat{\theta}_n$ measures how close $\widehat{\theta}_n$ comes to this lower bound. Most reasonable estimators of finite-dimensional parameters are asymptotically unbiased. Estimators whose asymptotic variance equals the Cràmer-Rao lower bound are *asymptotically efficient*. The maximum likelihood estimator under certain regularity assumptions is an example of an efficient estimator for parametric models. Interestingly enough, in 1951 Hodges produced a *super-efficient* estimator: an estimator that has asymptotic variance equal to the Cràmer-Rao lower bound for most of the parameter values, but smaller variance than the Cràmer-Rao lower bound for the other parameters. Because super-efficient estimators have undesirable local properties, certain regularity conditions are imposed on our class of estimators.

Definition 6.2 *Consider a local data generating process, where for each n, the data $X_1, ..., X_n$ are distributed according to $\phi_n = (\theta_n^T, \eta_n^T)^T$, where $\sqrt{n}(\phi_n - \phi^*)$ converges to a constant. An estimator $\widehat{\theta}_n$ is said to be regular if for each ϕ^*, $\sqrt{n}(\widehat{\theta}_n - \theta_n)$ has a limiting distribution that does not depend on the local data generating process.*

We limit our scope to estimators that are regular and asymptotically linear (RAL). Additionally we consider parametric models (i.e. ϕ is finite dimensional, and as a consequence so are θ and η) in our treatise of the geometry of influence functions for RAL estimators. These ideas are then generalized to semiparametric models.

6.2.3 Score Functions and Tangent Sets

The *score vector* $S_\phi : \mathcal{X} \to \mathbb{R}^p$ for a single observation X in a parametric model is defined as the derivative of the log-likelihood with respect to the elements of ϕ,

$$S_\phi(x; \phi_0) = \left. \frac{\partial \log p_X(x; \phi)}{\partial \phi} \right|_{\phi = \phi_0}.$$

We can partition this vector of derivatives as

$$S_\phi(X; \phi_0) = (S_\theta^T(X, \phi_0), S_\eta^T(X, \phi_0))^T.$$

The score vector has mean zero under suitable regularity conditions. We now have the necessary background to define tangent spaces. Let $\mathcal{H}$ denote the Hilbert space of all q-dimensional measurable functions of X with mean zero and finite variance, equipped with the covariance inner product, $\langle h_1, h_2 \rangle = E(h_1^T h_2)$. We define $\Lambda \subset \mathcal{H}$ as the linear subspace spanned by the score vector $S_\phi(X, \phi_0)$ as the set of all q-dimensional mean-zero random vectors consisting of

$$B^{q \times p} S_\phi(X, \phi_0)$$

for all $q \times p$ matrices B. The linear subset Λ is called the *tangent space*. In particular consider the linear subspace Λ_η spanned by the nuisance score vector $B^{q \times r} S_\eta(X, \phi_0)$ for all $q \times r$ matrices B*. This is the *nuisance tangent space*. The following corollary demonstrates two important properties of the relationship between influence functions and the tangent space. Namely, the inner product of the influence function and the tangent space spanned by the score vector S_θ has expectation 1, and the influence function $\psi_{\widehat{\theta}_n}(X)$ for $\widehat{\theta}_n$ is orthogonal to the nuisance tangent space Λ_η (i.e. $\psi \in \Lambda_\eta^\perp$). Additionally it gives the existence of a RAL estimator for every θ with that influence function. Although the proof is constructive of nature, since it demonstrates how to construct estimators that have influence functions corresponding to elements in the subspace of the Hilbert space satisfying the conditions below, we chose to omit it since it serves no purpose in our proofs. Finding a function that satisfies the conditions below, requires knowledge of the "truth" ϕ_0, which we typically lack. For more detail we refer the interested reader to Section 3.3 in Tsiatis (2006).

Corollary 6.1.1 *All RAL estimators have influence functions that belong to the subspace of our Hilbert space satisfying,*

- $E[\psi(X) S_\theta^T(X, \phi_0)] = I^q$;
- $E[\psi(X) S_\eta^T(X, \phi_0)] = 0$.

and, conversely, any element in the subspace above is the influence function of some RAL estimator.

Since RAL estimators are asymptotically normal, we can compare competing RAL estimators for θ by looking at the asymptotic variance. The asymptotic variance of a RAL estimator, in turn, is equal to the variance of its influence function. Therefore, it is sufficient to consider the variance of influence functions. The influence functions can be viewed as elements in a subspace of a Hilbert space, and the distance to the origin is the variance of the element. As a result, comparing competing estimators is equivalent to comparing elements in the subspace of the influence functions that have the smallest norms. Additionally the influence functions are orthogonal to the nuisance tangent space Λ_η. Therefore the space of influence functions is given by $h - \Pi(h|\Lambda_\eta)$ for all $h \in \mathcal{H}$, and

$$\Pi(h|\Lambda) = E(hS_\eta^T)E(S_\eta S_\eta^T)^{-1} S_\eta(X, \phi_0),$$

*Where $S_\eta : \mathcal{X} \to \mathbb{R}^r$.

and the tangent space $\Lambda = \{B^{q\times p}S_{\phi}(X, \phi_0)$ for all $B^{q\times p}\}$ can be written as the direct sum of the nuisance tangent space and the tangent space generated by the score function w.r.t. θ, i.e.

$$\Lambda = \Lambda_{\theta} \oplus \Lambda_{\eta}.$$

The *efficient influence function* $\psi_{\text{eff}}(\cdot)$ is the influence function with the smallest variance. Theorem 3.5 in Tsiatis (2006) proves that such an influence function exists, and is given by,

$$\psi_{\text{eff}}(X) = \Pi\bigg(\psi^{*}(X)\bigg|\Lambda\bigg),$$

where $\psi^{*}(X)$ is an arbitrary influence function.

6.2.4 Semiparametric Efficiency

We now have sufficient tools to extend the theory to the semiparametric case. The techniques demonstrated above are generalized in a typical mathematical fashion, by taking limits to infinity. Let $\mathcal{P}$ denote the class of densities parametrized by $\phi = (\theta, \eta)$, where θ is d-dimensional and η is infinite-dimensional. Let $p_0(x) = p(x; (\theta_0, \eta_0))$ denote the true density. We first consider *parametric submodels*, denoted by the class of densities $\mathcal{P}_{\theta,\gamma}$ with the following properties:

- $(\theta^T, \gamma^T)^T$ is a $(d + r)$-dimensional parameter. The value of the dimension of γ depends on the choice of parametric submodel.
- Every density in $\mathcal{P}_{\theta,\gamma}$ belongs to the semiparametric model $\mathcal{P}$, i.e. $\mathcal{P}_{\theta,\gamma} \subset \mathcal{P}$.
- The parametric submodel contains the true density, i.e. $p_0(\cdot) \in \mathcal{P}_{\theta,\gamma}$.

The third requirement implies that these models are only conceptual, and not instructive, since otherwise we require knowledge of the ”truth”. Our characterization of influence functions, and efficient influence functions, and RAL estimators for a parametric model obviously apply to parametric submodels. Consequently, the following properties hold:

1. Influence functions of RAL estimators for θ for a parametric submodel belong to the subspace of the Hilbert space $\mathcal{H}$ of q-dimensional, mean-zero and finite-variance, measurable functions (equipped with

the covariance inner product) that are orthogonal to the parametric submodel nuisance tangent space. The latter is given by,

$$\Lambda_\gamma = \{B^{q\times r} S_\gamma(X, Y, \theta_0, \gamma_0), \text{ for all } B^{q\times r}\},$$

and

$$S_\gamma = \frac{\partial \log p(x, y, \theta_0, \gamma_0)}{\partial \gamma}.$$

2. The efficient influence function for the parametric submodel is given by

$$\phi^{\text{eff}}_{\theta,\gamma}(X, Y) = (E(S^{\text{eff}}_{\theta,\gamma} S^{\text{eff}^T}_{\theta,\gamma}))^{-1} S^{\text{eff}}_{\theta,\gamma}(X, Y, \theta_0, \gamma_0),$$

where

$$S^{\text{eff}}_{\theta,\gamma}(X, Y, \theta_0, \gamma_0) = S_\theta(X, Y, \theta_0, \eta_0) - \Pi(S_\theta(X, Y, \theta_0, \eta_0)|\Lambda_\gamma), \text{ and}$$
$$S_\theta(X, Y, \theta_0, \eta_0) = \frac{\partial \log p(x, y, \theta_0, \eta_0)}{\partial \theta}.$$

3. The smallest asymptotic variance among RAL estimators for θ for a parametric submodel is

$$(E(S^{\text{eff}}_{\theta,\gamma} S^{\text{eff}^T}_{\theta,\gamma}))^{-1}.$$

An estimator for θ is an RAL estimator for a semiparametric model if it is an RAL estimator for every parametric submodel. Consequently, every influence function of an RAL estimator in a semiparametric model, must be an influence function of an RAL estimator within a parametric submodel. That is, if $\widehat{\theta}_n$ is a semiparametric estimator and,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0, \Sigma(\beta, \eta))$$

for all $p(x; \theta, \eta) \in \mathcal{P}$, then,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0, \Sigma(\beta, \gamma))$$

for all $p(x; \theta, \gamma) \in \mathcal{P}_{\theta,\gamma} \subset \mathcal{P}$. This allows us to characterize the influence functions of an RAL semiparametric estimator for θ as orthogonal to all parametric submodel nuisance tangent spaces. Furthermore, the variance of any RAL semiparametric influence functions must be greater than or equal to

$$(E(S^{\text{eff}}_{\theta,\gamma} S^{\text{eff}^T}_{\theta,\gamma}))^{-1}$$

for all parametric submodels $\mathcal{P}_{\theta,\gamma}$. This bound is called the *semiparametric efficiency bound*, and Theorem 6.2 emphasizes that it is reminiscent of the parametric efficiency (Cràmer-Rao) bound.

The *nuisance tangent space for a semiparametric model*, also denoted by Λ_η, is defined as the *mean-square closure* (i.e. the closure where the limits are taken w.r.t. the metric induced by the mean-square inner product) of *parametric submodel tangent spaces*. A parametric submodel nuisance tangent space is the set of elements,

$$\{B^{q\times r}S_\gamma^{r\times 1}(X,\theta_0,\eta_0)\}.$$

The *mean-square closure* of the spaces above is defined as the space $\Lambda_\eta \subset \mathcal{H}$, for

$$\Lambda_\eta = \{h^{q\times 1}(X) \in \mathcal{H} \text{ such that } E(h^T h) < \infty, \text{and there exists a sequence } B_j S_{\gamma_j}(X) \text{ such that } \|h - B_j S_{\gamma j}\|^2 \overset{j\to\infty}{\to} 0, \text{for a sequence of parametric submodels indexed by } j\},$$

where $\|h\|^2 = E(h^T h)$. We assume that Λ_η is a closed linear subspace, so that projections exist. To characterize the influence functions for semiparametric RAL estimators for θ, we need to define the semiparametric efficient score.

Definition 6.3 *The semiparametric efficient score for θ is defined as,*

$$S_{eff}(X,\theta_0,\eta_0) = S_\theta(X,\theta_0,\eta_0) - \Pi(S_\theta(X,\theta_0,\eta_0)|\Lambda_\eta).$$

Theorem 6.2 *The semiparametric efficiency bound is equal to the inverse of the variance matrix of the semiparametric efficient score, i.e.*

$$(E(S_{eff}S_{eff}^T))^{-1}.$$

The *efficient influence function* is defined as the influence function of a semiparametric RAL estimator that achieves the semiparametric efficiency bound. We formulate the semiparametric analogue of Corollary 6.1.1, so that we can characterize the efficient influence function as the unique element satisfying the conditions below, whose variance matrix equals the efficiency bound, and

$$\psi_{\text{eff}}(X,\theta_0,\eta_0) = (E(S_{\text{eff}}S_{\text{eff}}^T))^{-1}S_{\text{eff}}(X,\theta_0,\eta_0).$$

Theorem 6.3 *Any semiparametric RAL estimator for θ must have an influence function $\psi(X)$ that satisfies*

- $E(\psi(X)S_{\theta}^{T}(X,\theta_0,\eta_0)) = E(\psi(X)S_{\mathit{eff}}^{T}(X,\theta_0,\eta_0)) = I^{q}$.
- *The influence function $\psi(X)$ is orthogonal to the nuisance tangent space,* $\Pi(\psi(X)|\Lambda_{\eta}) = 0$.

In summary, the approach in semiparametric statistics is to construct estimators through deriving influence functions. We treat these influence functions as normalized elements in the orthogonal complement of a nuisance tangent set Λ_{η}. Therefore our problem reduces to deriving the orthogonal complement $\Lambda_{\eta}^{\perp}$. Although our estimation method does not rely on this paradigm, we derive the nuisance tangent space in order to prove certain asymptotic results in Section 6.4.

6.3 Semiparametric M-estimation

Assume that $X_1, ..., X_n$ are drawn from $P_{\theta,\eta}$, where $\theta \in \mathbb{R}^d$ and $\eta \in \mathcal{H}$. Suppose that the infinite dimensional space $\mathcal{H}$ has norm $\|\cdot\|$ and the true parameter is θ_0, η_0. For any fixed $\eta \in \mathcal{H}$, let $\eta_1(t), \eta_2(t)$ be smooth curves running through η at $t = 0$. Let h_1, h_2 be elements in the tangent set Λ_η. For simplicity, we write the loss functions $m(\theta, \eta; X, Y)$ as $m(\theta, \eta)$. Denote the derivatives as,

$$m_1(\theta,\eta) = \frac{\partial}{\partial\theta} m(\theta,\eta), \quad \text{and } m_2(\theta,\eta)[h_1] = \frac{\partial}{\partial t}\bigg|_{t=0} m(\theta, \eta_1(t)),$$

Furthermore denote the set of second derivatives as,

$$m_{11}(\theta,\eta) = \frac{\partial}{\partial\theta} m_1(\theta,\eta), \quad m_{12}(\theta,\eta)[h_1] = \frac{\partial}{\partial t}\bigg|_{t=0} m_1(\theta, \eta_1(t))$$

$$m_{21}(\theta,\eta)[h_1] = \frac{\partial}{\partial\theta} m_2(\theta,\eta)[h_1], \quad m_{22}(\theta,\eta)[h_1][h_2] = \frac{\partial}{\partial t}\bigg|_{t=0} m_2(\theta, \eta_2(t))[h_1],$$

For vectors $H = (h_1, ..., h_d) \in \Lambda_\eta^d$ we use the shorthand notation

$$m_2(\theta,\eta)[H] = (m_2(\theta,\eta)[h_1], ..., m_2(\theta,\eta)[h_d]).$$

Assume there exists $H^* = (h_1^*, ..., h_d^*) \in \Lambda_\eta^d$ such that for any $H \in \Lambda_\eta^d$ we have,

$$P(m_{12}(\theta_0,\eta_0)[H] - m_{22}(\theta_0,\eta_0)[H^*, H]) = 0. \tag{6.1}$$

Then let $\tilde{m}(\theta,\eta) \equiv m_1(\theta,\eta) - m_2(\theta,\eta)[H^*]$. Suppose that our estimator $(\widehat{\theta}_n, \widehat{\eta}_n)$ satisfies the following near-maximization criterion,

$$\mathbb{P}_n \tilde{m}(\widehat{\theta}_n, \widehat{\eta}_n) = o_P(n^{-1/2}). \tag{6.2}$$

The following conditions are sufficient for such $\widehat{\theta}_n$ to be asymptotically normal, and $\sqrt{n}$ consistent. We slightly strengthen condition A1 from Kosorok (2005). Whereas the original result requires mere consistency of the finite-dimensional $\widehat{\theta}_n$ to θ_0, we require a rate of convergence. By doing this, we can weaken the original stochastic equicontinuity condition (A3) by considering the supremum over all θ such that $|\theta - \theta_0| \leq \delta_n = Cn^{-c_1}$ for some constant $C > 0$, instead of arbitrary sequences $\delta_n \downarrow 0$. The proof is analogous to Theorem 1 in Kosorok (2005).

A1: (Consistency and rate of convergence) Assume

$$|\widehat{\theta}_n - \theta_0| = O_P(n^{-c_1}), \text{ and } \|\widehat{\eta}_n - \eta_0\| = O_P(n^{-c_1}),$$

for some $c_1 > 0$.
A2: (Finite variance) $0 < \det(I^*) < \infty$, where

$$\begin{aligned} I^* = \{&P(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[H^*])\}^{-1} \\ &\times P[m_1(\theta_0, \eta_0) - m_2(\theta_0, \eta_0)[H^*]]^{\otimes 2} \\ &\times \{P(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[H^*]\}^{-1}. \end{aligned}$$

A3: (Stochastic equicontinuity) For any $C > 0$,

$$\sup_{|\theta-\theta_0| \leq Cn^{-c_1}, \|\eta-\eta_0\| \leq Cn^{-c_1}} |\sqrt{n}(\mathbb{P}_n - P)(\tilde{m}(\theta, \eta) - \tilde{m}(\theta_0, \eta_0))| = o_P(1).$$

A4: (Smoothness of the model) For some $c_2 > 1$ satisfying $c_1 c_2 > 1/2$ and for all (θ, η) satisfying $\{(\theta, \eta) : |\theta - \theta_0| \leq \delta_n, \|\eta - \eta_0\| \leq Cn^{-c_1}\}$,

$$\begin{aligned} &\Bigg| P\Bigg\{ (\tilde{m}(\theta, \eta) - \tilde{m}(\theta_0, \eta_0)) - (m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[H^*])(\theta - \theta_0) \\ &- \left(m_{12}(\theta_0, \eta_0) \left[\frac{\eta - \eta_0}{\|\eta - \eta_0\|} \right] - m_{22}(\theta_0, \eta_0)[H^*] \left[\frac{\eta - \eta_0}{\|\eta - \eta_0\|} \right] \right) \|\eta - \eta_0\| \Bigg\} \Bigg| \\ &= o(|\theta - \theta_0|) + O(\|\eta - \eta_0\|^{c_2}). \end{aligned}$$

Theorem 6.4 *(Theorem 1, Kosorok (2005)) Suppose that $(\widehat{\theta}_n, \widehat{\eta}_n)$ satisfies Equation (6.2), and that Conditions A1-A4 hold, then*

$$\begin{aligned} \sqrt{n}(\widehat{\theta}_n - \theta_0) = -\sqrt{n}\{&P(m_{11}(\theta_0, \eta_0) - m_{21}(\theta_0, \eta_0)[H*])\}^{-1} \\ &\times \mathbb{P}_n(m_1(\theta_0, \eta_0) - m_2(\theta_0, \eta_0)[H^*]) + o_P(1). \end{aligned}$$

Hence $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and variance I^*.

We briefly discuss how this result can be applied to our estimator. To find $H^* \in \Lambda_g^d$ that satisfies Equation (6.1) we need to know what the nuisance tangent space looks like. We derive this space in the next section. For condition A1 we require a rate of convergence of $(\widehat{C}_n, \widehat{g}_n)$ to (C_0, g_0). We initially show that $(\widehat{C}_n, \widehat{g}_n)$ is consistent for (C_0, g_0). Both results these can be shown by performing entropy calculations and using the results for M-estimation we discussed in Section 4.3. Assumption A2 is a condition that ensures that the limiting distribution of $\sqrt{n}(\text{vec}(\widehat{C}_n) - \text{vec}(C_0))$ is non-degenerate. To demonstrate condition A3, we use maximal inequalities and entropy calculations. The last condition A4 can be verified to hold by using Taylor expansions for functionals, and resembles a typical smoothness condition.

6.4 Semiparametric estimation in the SDR model

Suppose we observe i.i.d. copies (X_i, Y_i) of a random vector $X \in \mathbb{R}^p$ and response $Y \in \mathbb{R}$, drawn from the following model,

$$Y = g_0(\beta_0^T X) + \epsilon, \qquad \text{where } \beta_0 = (I_{d_0}, C_0^T)^T,$$

with $E(\epsilon|X) = 0$. The parameters are the unknown function g_0 and the $(p - d_0) \times d_0$ parameter matrix C_0. Since d_0 is assumed to be known, we write d for simplicity in the remainder of this section. As we discussed in Chapter 2, we estimate g_0 with a B-spline, in contrast to the kernel approach used in Huang and Chiang (2017).

6.4.1 Tangent Spaces

This derivation is heavily inspired by the examples in Chapter 4 of Tsiatis (2006). We assume that Y is continuous and (X, Y) has dominating measure $\nu_X \times l_Y$. The density of one observation (X, Y) belongs to the semi-parametric model,

$$\mathcal{P} = \{p(x, y; \beta, g)\},$$

defined with respect to the dominating measure $\nu_X \times l_Y$. We denote the true density as $p_0(x, y) = p(x, y; \beta_0, g_0)$. Since there is a one-to-one transformation between (X, Y) and $(X, Y - g(\beta^T X)) = (X, \epsilon)$, we can express the density as,

$$p_{X,Y}(x, y) = p_{X,\epsilon}(x, y - g(\beta^T x)) = p_{X,\epsilon}(x, \epsilon).$$

In turn, this expression can be written as,

$$p_{X,\epsilon}(x, \epsilon) = \eta_1(x)\eta_2(x, \epsilon),$$

where $\eta_1(x) = p_X(x)$ is any non-negative function of x such that,

$$\int \eta_1(x) d\nu(x) = 1, \tag{6.3}$$

and $\eta_2(x, \epsilon) = p_{\epsilon|X}(\epsilon|x)$ is any non-negative function such that,

$$\int \eta_2(x, \epsilon) d\epsilon = 1 \text{ for all } x, \tag{6.4}$$

$$\int \epsilon \eta_2(x, \epsilon) d\epsilon = 0 \text{ for all } x. \tag{6.5}$$

The non-negativity and constraints in Equation (6.3)-(6.5) are used to characterize elements in our semi-parametric model $\mathcal{P}$ as,

$$p(x, y; \beta, \eta_1(\cdot), \eta_2(\cdot)) = \eta_1(x)\eta_2(x, y - g(\beta^T x)).$$

The true density is denoted by

$$p_0(x, y) = \eta_1^0(x)\eta_2^0(x, y - g_0(\beta_0^T x)).$$

We view η_1, η_2, and g as nuisance parameters. To derive the semiparametric nuisance tangent space, we consider parametric submodels,

$$p_X(x, \gamma_1), \qquad p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g)|x, \gamma_2), \qquad g(\beta_0^T x; \gamma_g),$$

where $\gamma_1, \gamma_2, \gamma_g$ are vectors of dimension r_1, r_2, r_g respectively. Thus $\gamma = (\gamma_1^T, \gamma_2^T, \gamma_g^T)^T$ is an r-dimensional vector, for $r = r_1 + r_2 + r_g$. This parametric submodel is given by,

$$\begin{aligned}\mathcal{P}_\gamma = \{&p(x, y; \gamma_1, \gamma_2, \gamma_g) = p_X(x, \gamma_1)p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g)|x, \gamma_2),\\ &\text{for } (\gamma_1^T, \gamma_2^T, \gamma_g^T)^T \subset \mathbb{R}^r\}.\end{aligned}$$

Note that $\mathcal{P}_\gamma$ must contain the truth $p_0(x, y)$ to be a parametric submodel, which we denote as,

$$p_0(x, y) = p_X(x, \gamma_1^0)p_{\epsilon|X}(y - g(\beta^T x; \gamma_g^0)|x, \gamma_2^0).$$

The parametric submodel nuisance score vector is given by

$$\begin{aligned}S_\gamma(x, y; \beta_0, \gamma_0) = &\left\{ \left(\frac{\partial \log p(x, y; \beta_0, \gamma)}{\partial \gamma_1}\right)^T, \left(\frac{\partial \log p(x, y; \beta_0, \gamma)}{\partial \gamma_2}\right)^T, \right.\\ &\left.\left(\frac{\partial \log p(x, y; \beta_0, \gamma)}{\partial \gamma_g}\right)^T \right\}^T \Bigg|_{\gamma = \gamma_0}\\ &= \{S_{\gamma_1}^T(x, y; \beta_0, \gamma_1^0), S_{\gamma_2}^T(x, y; \beta_0, \gamma_2^0), S_{\gamma_g}^T(x, y; \beta_0, \gamma_g^0)\}^T.\end{aligned}$$

Since,

$$\log p(x, y; \beta, \gamma) = \log p_X(x, \gamma_1) + \log p_{\epsilon|X}(y - g(\beta^T x; \gamma_g)|x, \gamma_2),$$

we obtain for the score functions evaluated at the true parameters,

$$\begin{aligned}
S_{\gamma_1}(x, y; \beta_0, \gamma_0) &= \frac{\partial \log p_X(x, \gamma_1^0)}{\partial \gamma_1} \\
S_{\gamma_2}(x, y; \beta_0, \gamma_0) &= \frac{\partial \log p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g^0)|x, \gamma_2^0)}{\partial \gamma_2} \\
S_{\gamma_g}(x, y; \beta_0, \gamma_0) &= \frac{\partial \log p_{\epsilon|X}(Y - g(\beta_0^T x; \gamma_g^0)|x, \gamma_2^0)}{\partial \gamma_g} \\
&= \frac{\partial \log p_{\epsilon|X}(y - g_0(\beta_0^T x; \gamma_g^0)|x, \gamma_2^0)}{\partial \epsilon} \cdot \frac{\partial (y - g_0(\beta_0^T x; \gamma_g^0))}{d\gamma_g} \\
&= -\frac{p'_{\epsilon|X}(\epsilon|x, \gamma_2^0)}{p_{\epsilon|X}(\epsilon|x, \gamma_2^0)} \cdot \frac{\partial \epsilon}{d\gamma_g},
\end{aligned}$$

where $\epsilon = y - g(\beta_0^T x)$. Note that we can write the score functions w.r.t. γ_2 and γ_g as $S_{\gamma_2}(x, \epsilon)$ and $S_{\gamma_g}(x, \epsilon)$ due to the one-to-one transformation of (X, Y) to (X, ϵ). An element in the parametric submodel nuisance tangent space is given by

$$B^{q\times r} S_\gamma(X, \epsilon) = B_1^{q\times r_1} S_{\gamma_1}(X) + B_2^{q\times r_2} S_{\gamma_2}(X, \epsilon) + B_g^{q\times r_g} S_{\gamma_g}(X, \epsilon),$$

for matrices of constants $B^{q\times r_1}, B^{q\times r_2}, B^{q\times r_g}$. The parametric submodel nuisance tangent space,

$$\Lambda_\gamma = \{B^{q\times r} S_\gamma(X, \epsilon), \text{ for all } B^{q\times r}\},$$

can thus be written as

$$\Lambda_\gamma = \Lambda_{\gamma_1} + \Lambda_{\gamma_2} + \Lambda_{\gamma_g},$$

where

$$\begin{aligned}
\Lambda_{\gamma_1} &= \{B_1^{q\times r_1} S_{\gamma_1}(X) \text{ for all } B^{q\times r_1}\}, \\
\Lambda_{\gamma_2} &= \{B_1^{q\times r_2} S_{\gamma_2}(X, \epsilon) \text{ for all } B^{q\times r_2}\}, \\
\Lambda_{\gamma_g} &= \{B_1^{q\times r_m} S_{\gamma_m}(X, \epsilon) \text{ for all } B^{q\times r_g}\}.
\end{aligned}$$

The semiparametric nuisance tangent space is the mean-square closure of all parametric submodel nuisance tangent spaces, i.e. the mean square closure of Λ_γ. If γ_1, γ_2, and γ_g are *variationally independent*, (i.e. proper densities in the parametric submodel can be defined by considering any combination of γ_1, γ_2, and γ_g), this would imply that the mean-square closure of the "sum" of $\Lambda_{\gamma_1}, \Lambda_{\gamma_2}$, and Λ_{γ_g}, is the "sum" of of the mean square closures of $\Lambda_{\gamma_1}, \Lambda_{\gamma_2}$, and Λ_{γ_g}. This would imply that,

$$\Lambda = \Lambda_{1s} + \Lambda_{2s} + \Lambda_{gs},$$

where

$$\begin{aligned} \Lambda_{1s} &= \{ \text{ mean square closure of all } \Lambda_{\gamma_1} \} \\ \Lambda_{2s} &= \{ \text{ mean square closure of all } \Lambda_{\gamma_2} \} \\ \Lambda_{gs} &= \{ \text{ mean square closure of all } \Lambda_{\gamma_g} \}. \end{aligned}$$

Ma and Zhu (2014) contain the explicit form of the tangent space, however there is no derivation present. We explicitly derive each of these spaces to familiarize the reader with techniques from the semiparametric paradigm. In many cases the structure of the parametric submodel nuisance tangent space allows us to make an educated guess for the semiparametric nuisance tangent space. We still need to verify our guess, which we do in Theorem 6.5.

Theorem 6.5 *The following equalities hold:*

1.

$$\Lambda_{1s} = \{f(x) : \textit{all functions } f \textit{ such that } Ef = 0 \textit{ and } Ef^2 < \infty\}.$$

2.

$$\begin{aligned} \Lambda_{2s} = \{f(x,\epsilon) &: \textit{all functions } f \textit{ such that} \\ & E(f|X) = 0 \textit{ and } E(f(X,\epsilon)\epsilon^T|X) = 0\}. \end{aligned}$$

3.

$$\Lambda_{gs} = \left\{ \frac{\frac{\partial}{\partial\epsilon}\eta_2^0(x,\epsilon)}{\eta_2^0(x,\epsilon)} f(\beta_0^T X) : \textit{ for all functions } f \right\}.$$

Proof. The proof is split in three parts, one for each of the nuisance parameters. We also explain how we arrived at our guess, and how we can prove the guess.

The space Λ_{1s}

Intuition: For any parametric submodel we have $S_{\gamma_1}(X, Y) = S_{\gamma_1}(X)$, and every score function has mean zero under certain regularity assumptions, i.e.

$$E(S_{\gamma_1}(X)) = 0. \tag{6.6}$$

This suggests that we want to consider the semiparametric nuisance tangent space for η_1 to be all functions of X that satisfy equation (6.6). Denote this space as S_1. We need to prove following two statements:

1. Any element of Λ_{γ_1}, for any parametric submodel indexed by γ_1, belongs to S_1.
2. Any element of S_1 is an element of Λ_{γ_1} for a parametric submodel or a limit of such elements.

Observe that (1) is satisfied because

$$E(B^{q\times r_1} S_{\gamma_1}(X)) = 0, \text{ for all constant matrices } B^{q\times r_1}.$$

To demonstrate the second statement, consider any bounded function $f \in S_1$. Consider the parametric submodel with density $p_X(x, \gamma_1) = p_0(x)\{1 + \gamma_1^T f(x)\}$, where γ_1 is a $q-$dimensional vector, and sufficiently small such that,

$$1 + \gamma_1^T f(X) \geq 0.$$

This enforces $p_X(x, \gamma_1)$ to be a proper density in a neighbourhood of γ_1 around zero. Since $p_X(x, \gamma_1) \geq 0$ for all x, and

$$\begin{aligned}\int p_X(x, \gamma_1) d\nu(x) &= \int p_0(x)\{1 + \gamma_1^T f(x)\} d\nu(x)\\ &= \int p_0(x) d\nu(x) + \int \gamma_1^T f(x) p_0(x) d\nu(x)\\ &= 1 + 0 = 1.\end{aligned}$$

The second term in the second-to-last equality follows from the fact that

$$0 = E(f(X)) = \int f(x) dP(x) = \int f(x) p_0(x) d\nu(x).$$

The corresponding score vector is

$$\begin{aligned}S_{\gamma_1}(X) &= \left.\frac{\partial \log(p_0(X)\{1 + \gamma_1^T f(X)\})}{\partial \gamma_1}\right|_{\gamma_1=0}\\ &= \left.\frac{p_0(X) f(X)}{p_0(X) + \gamma_1^T f(X) p_0(X)}\right|_{\gamma_1=0}\\ &= f(X).\end{aligned}$$

Since the parametric submodel nuisance tangent space consists of $B^{q \times q} S_{\gamma_1}$ for all matrices B, we can just set $B^{q \times q} = I_q$. This implies that $f(X)$ is an element of the parametric submodel nuisance tangent space we constructed above. If $f(X)$ is not bounded, it can be taken as the limit of bounded mean-zero functions of X. We conclude that all elements of S_1 are either elements parametric submodel nuisance tangent space or a limit of such elements. This proves that $\Lambda_{1s} = S_1$.

The space Λ_{2s}

Intuition: Recall that $\epsilon = y - g(\beta_0^T x)$. Since $\int p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g)|x, \gamma_2) d\epsilon = 1$ for all x, γ_2, γ_g, we have,

$$\frac{\partial}{\partial \gamma_2} \int p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g)|x, \gamma_2) d\epsilon = 0.$$

If we swap the integral and derivative, and multiply with 1, we obtain

$$\begin{aligned} 0 &= \int \frac{\partial p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g)|x, \gamma_2)}{\partial \gamma_2} \cdot \frac{p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g^0)|x, \gamma_2^0)}{p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g^0)|x, \gamma_2^0)} d\epsilon \\ &= \int \frac{\frac{\partial p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g)|x, \gamma_2)}{\partial \gamma_2}}{p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g^0)|x, \gamma_2^0)} dP(\epsilon|x). \end{aligned}$$

This implies $E(S_{\gamma_2}(X, \epsilon)|X) = 0$. Additionally,

$$\begin{aligned} 0 &= E(\epsilon|X) \\ &= \int p_{\epsilon|X}(\epsilon|x, \gamma_2) \epsilon \, d\epsilon. \end{aligned}$$

By differentiating w.r.t. γ_2 and multiplying by 1 we obtain,

$$\begin{aligned} &\int \frac{\partial p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g^0)|x, \gamma_2) \epsilon}{\partial \gamma_2} \cdot \frac{p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g^0)|x, \gamma_2^0)}{p_{\epsilon|X}(y - g(\beta^T x; \gamma_g^0)|x, \gamma_2^0)} \, d\epsilon \\ &= \int S_{\gamma_1}(x, \epsilon) \epsilon p_{\epsilon|X}(y - g(\beta^T x; \gamma_g^0)|x, \gamma_2^0) \, d\epsilon \\ &= E(S_{\gamma_2}(x, \epsilon) \epsilon | X) \\ &= 0. \end{aligned}$$

Therefore any element f of a parametric submodel nuisance tangent space must satisfy:

$$E(f(X, \epsilon)|X) = 0, \tag{6.7}$$

$$E(f(X,\epsilon)\epsilon|X) = 0. \tag{6.8}$$

We conjecture that Λ_{2s} equals the space of all functions that satisfy equation (6.7) & (6.8). Denote this space as S_2. Let $f(X,\epsilon)$ be a bounded function in S_2, and let γ_2 be a q-dimensional parameter sufficiently small so that:

$$1 + \gamma_2^T f(x,\epsilon) \geq 0 \text{ for all } x, \epsilon.$$

Consider the parametric submodel

$$p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g)|x, \gamma_2) = p_\epsilon^0(\epsilon|x)\{1 + \gamma_2^T f(x,\epsilon)\}.$$

Note that

$$\int p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g^0)|x, \gamma_2) d\epsilon = 1 \qquad \text{for all } x, \gamma_2, \gamma_g,$$

and

$$E(\epsilon|X) = 0.$$

The score vector of this parametric submodel is

$$\begin{aligned} S_{\gamma_2}(x,\epsilon) &= \left.\frac{\partial \log p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g^0)|x, \gamma_2^0)}{\partial \gamma_2}\right|_{\gamma_2=0} \\ &= \left.\frac{\partial \log(p_{\epsilon|X}^0(\epsilon|x)\{1 + \gamma_2^T f(x,\epsilon)\})}{\partial \gamma_2}\right|_{\gamma_2=0} \\ &= \left.\frac{p_{\epsilon|X}^0(\epsilon|x) f(x,\epsilon)}{p_{\epsilon|X}^0(\epsilon,x)\{1 + \gamma_2^T f(x,\epsilon)\}}\right|_{\gamma_2=0} \\ &= f(x,\epsilon). \end{aligned}$$

Choosing $B^{q\times q} = I_q$, similarly to the investigation of Λ_{1s}, allows us to conclude that $f(x,\epsilon)$ is an element of this parametric nuisance tangent space. Hence any bounded element $f(x,\epsilon)$ satisfying (6.7) and (6.8) can be obtained. Additionally any unbounded element can be obtained as a sequence of bounded $f(x,\epsilon)$ satisfying (6.7) and (6.8). Since our function space is closed, this limit is also included in Λ_{2s}. We conclude that $\Lambda_{2s} = S_2$.

The space Λ_{gs}

Intuition: There are no restrictions on the function g, so we entirely base our guess off the score vector. We know that any element in Λ_{γ_g}, say

$f(X, \epsilon)$ for any parametric submodel must be of the form

$$B^{q\times r_g} \times \frac{p'_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g)|x, \gamma_2^0)}{p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g)|x, \gamma_2^0)} \frac{\partial g(\beta_0^T X; \gamma_g)}{\partial \gamma_g}. \tag{6.9}$$

We conjecture that Λ_{gs} is equal to:

$$S_g = \left\{ \frac{\frac{\partial}{\partial \epsilon} p^0_{\epsilon|X}(\epsilon|x, \gamma_2^0)}{p^0_{\epsilon|X}(\epsilon|x, \gamma_2^0)} \cdot f(\beta_0^T X) \text{ for all functions } f \right\}. \tag{6.10}$$

Let $f(\beta_0^T x)$ be a bounded function in S_g. Consider the parametric submodel,

$$g(\beta_0^T x; \gamma_g) = g_0(\beta_0^T x) - \gamma_g^T f(\beta_0^T X). \tag{6.11}$$

By setting $\gamma_g = 0$, we see that the "truth" is contained in this parametric submodel.
The score function of this parametric submodel is then given by,

$$\begin{aligned}
S_{\gamma_g}(x, \epsilon) &= \frac{\partial \log p_X(x; \gamma_1^0) p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g)|x, \gamma_2^0)}{\partial \gamma_g}\bigg|_{\gamma_g = 0} \\
&= \frac{\partial \log p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g)|x, \gamma_2^0)}{\partial \epsilon} \cdot \frac{\partial (y - g(\beta_0^T x; \gamma_g))}{\partial \gamma_g}\bigg|_{\gamma_g = 0} \\
&= \frac{\frac{\partial}{\partial \epsilon} p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g))}{p_{\epsilon|X}(y - g(\beta_0^T x; \gamma_g))} \cdot - \frac{\partial (g_0(\beta_0^T x) - \gamma_g^T f(\beta_0^T X))}{\partial \gamma_g}\bigg|_{\gamma_g = 0} \\
&= \frac{\frac{\partial}{\partial \epsilon} p_{\epsilon|X}(y - g_0(\beta_0^T x))}{p_{\epsilon|X}(y - g_0(\beta_0^T x))} \cdot f(\beta_0^T X)
\end{aligned}$$

Choosing $B^{q\times q} = I_q$ allows us to conclude that $f(\beta_0^T X)$ is an element of this parametric nuisance tangent space. Hence any bounded function $f(\beta_0^T X)$ in S_g can be obtained from a parametric nuisance tangent space. Additionally, unbounded functions can be obtained as a sequence of bounded $f(\beta_0^T X)$ that are in a parametric nuisance tangent space. Since our function space is closed, this limit is also included in Λ_{gs}. We conclude that $\Lambda_{gs} = S_g$. □

6.4.2 Notation

Let C be a $(p-d)\times d$ real-valued matrix, and define the loss-functions as,

$$m_{\beta,g}(X,Y) = -(Y - g(\beta^T X))^2; \text{ where } \beta = (I_d, C^T)^T.$$

Corresponding to the notation in Section 6.3, we have,

$$\begin{aligned}
&\partial_1 m_{\beta,g}(X,Y) = -2(Y - g(\beta^T X))g'(\beta^T X)\\
&\partial_2 m_{\beta,g}(X,Y)[h] = -2(Y - g(\beta^T X))h\\
&\partial_{11} m_{\beta,g}(X,Y) = 2g'(\beta^T X)^{\otimes 2} - 2(Y - g(\beta^T X))g''(\beta^T X)\\
&\partial_{12} m_{\beta,g}(X,Y)[h] = 2hg'(\beta^T X)\\
&\partial_{21} m_{\beta,g}(X,Y)[h] = 2g'(\beta^T X)h\\
&\partial_{22} m_{\beta,g}(X,Y)[h_1,h_2] = 2h_2h_1.
\end{aligned}$$

Lemma 6.1 *The vector of all zeros, $H^* = 0$ satisfies Equation (6.1).*

Proof. By Theorem 6.5 we have that for $H \in \Lambda_{gs}$,

$$P\partial_{12} m_{\beta_0,g_0}[H] = 2P\left(\frac{\frac{\partial}{\partial\epsilon}p(\epsilon|X)}{p(\epsilon|X)}h(\beta_0^T X)g_0'(\beta_0^T X)\right),$$

for some arbitrary function h. By the law of total expectation we have,

$$\begin{aligned}
E\left(\frac{\frac{\partial}{\partial\epsilon}p(\epsilon|X)}{p(\epsilon|X)}h(\beta_0^T X)g_0'(\beta_0^T X)\right) &= E\left[E\left(\frac{\frac{\partial}{\partial\epsilon}p(\epsilon|X)}{p(\epsilon|X)}h(\beta_0^T X)g_0'(\beta_0^T X)\middle|X\right)\right]\\
&= E\left[E\left(\frac{\frac{\partial}{\partial\epsilon}p(\epsilon|X)}{p(\epsilon|X)}\middle|X\right)h(\beta_0^T X)g_0'(\beta_0^T X)\right].
\end{aligned}$$

The inner expectation can be computed as,

$$\begin{aligned}
E\left(\frac{\frac{\partial}{\partial\epsilon}p(\epsilon|x)}{p(\epsilon|x)}\middle|X\right) &= \int_{-\infty}^{\infty}\frac{\frac{\partial}{\partial\epsilon}p(\epsilon|x)}{p(\epsilon|x)}p(\epsilon|x)d\epsilon\\
&= p(\epsilon|x)\Big|_{\epsilon=-\infty}^{\infty}\\
&= 0.
\end{aligned}$$

Thus for $H^* = 0$,

$$P(\partial_{12} m_{\beta_0,g_0}[H] - \partial_{22} m_{\beta_0,g_0}[H^*,H]) = 0, \qquad \forall H \in \Lambda_{gs}^d, \tag{6.12}$$

□

As a consequence,

$$\begin{aligned}\tilde{m}_{\beta,g}(X,Y) &= \partial_1 m_{\beta,g}(X,Y) - \partial_2 m_{\beta,g}(X,Y)[H^*] \\ &= -2(Y - g(\beta^T X))g'(\beta^T X).\end{aligned} \tag{6.13}$$

6.5 Asymptotic results for the Central Mean Subspace estimation

We prove $\sqrt{n}$-consistency and asymptotic normality of the estimator of the central mean subspace by demonstrating that the conditions of Theorem 6.4 hold. To prove condition A1 of Theorem 6.4, we first show consistency of $(\widehat{C}_n, \widehat{g}_n)$ for (C_0, g_0). We then prove a rate of convergence of the multivariate B-spline estimator $\widehat{g}_n$ to g_0 and the parameter matrix $\widehat{C}_n$ to C_0. These are shown below in Theorem 6.6, and 6.7 respectively. We require stochastic equicontinuity for A3, and this is demonstrated in Lemma 6.5. All three results rely on entropy calculations. As such, our proofs heavily rely on the empirical process theory discussed in Chapter 4. Assumption A4 is shown to hold by using appropriate Taylor expansions for $P\tilde{m}_{\beta,g}$ at (β_0, g_0).

6.5.1 Technicalities and Model Assumptions

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the domain of X and Y respectively. We remind the reader that we assume that the function g_0 has 4 continuous (partial) derivatives, and that we estimate g_0 with a B-spline of degree 3. The implications of these two properties are discussed later on in this section. For a given β, we compute the B-spline basis functions using the Cox-de-Boor recursion formula discussed in Chapter 2. The coefficients are then estimated by Equation (2.7). In the remainder of this section, we use the product form introduced in Equation (2.3) to write a B-spline function g, i.e.,

$$g(\beta^T X) = \psi(\beta^T X)^T a,$$

where $\psi(\beta^T X)$ is the vectorized tensor product basis, and a is the vectorized coefficient matrix. We denote the "best" B-spline function with q_n^d knots as,

$$g_n(\beta^T X) = \psi(\beta^T X)^T a_n.$$

We can think of g_n as the B-spline function with q_n^d knots that lies "closest" to the true function g_0. In particular, we set (β_0, g_n) as the maximizer of the centering function M_n over their respective parameter spaces,

$$M_n(\beta, g) = Pm_{\beta,g} = -P(Y - g(\beta^T X))^2, \qquad \beta = (I_d, C^T)^T. \tag{6.14}$$

Since this expression is unknown, we estimate the coefficients a_n, and the parameter matrix C_0 by a maximizer $(\widehat{C}_n, \widehat{g}_n)$ of $\mathbb{M}_n$,

$$\mathbb{M}_n(\beta, g) = \mathbb{P}_n m_{\beta,g} = -\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(\beta^T X_i))^2, \qquad \beta = (I_d, C^T)^T, \tag{6.15}$$

where,

$$\widehat{g}_n(\widehat{\beta}_n^T X) = \psi(\widehat{\beta}_n^T X)^T \widehat{a}_n.$$

Derivative of a B-spline Curve

Let ψ_ρ denote a B-spline basis vector. A B-spline function g of degree ρ is denoted as follows,

$$g(x) = \psi_\rho(x)^T a.$$

We know from Property 5 in Section 2.3 that the derivative of a B-spline basis functions can be computed by,

$$\frac{d}{dx}\psi_{i,\rho}(x) = \rho\left\{\frac{\psi_{i,\rho-1}(x)}{t_{i+\rho} - t_i} - \frac{\psi_{i+1,\rho-1}(x)}{t_{i+\rho+1} - t_{i+1}}\right\}.$$

The derivative of the B-spline function g then equals,

$$\frac{d}{dx}g(x) = \psi_{\rho-1}(x)^T a',$$

where the coefficients a' are defined as,

$$a'_i = \frac{\rho(a_{i+1} - a_i)}{t_{i+\rho+1} - t_{i+1}}.$$

We can derive the second derivative of a B-spline function recursively by,

$$\frac{d^2}{dx^2}g(x) = \psi_{\rho-2}(x)^T a''.$$

The coefficients are defined as,

$$a''_i = (\rho - 1)\frac{a'_{i+1} - a'_i}{t_{i+\rho} - t_{i+1}}.$$

The corresponding basis functions are defined as,

$$\psi'_{i,\rho-1}(x) = \rho(\rho-1)\left\{\frac{\frac{\psi_{i,\rho-2}(x)}{t_{i+\rho-1}-t_i} - \frac{\psi_{i+1,\rho-2}(x)}{t_{i+\rho}-t_{i+1}}}{t_{i+\rho} - t_i} - \frac{\frac{\psi_{i+1,\rho-2}(x)}{t_{i+\rho}-t_{i+1}} - \frac{\psi_{i+2,\rho-2}(x)}{t_{i+\rho+1}-t_{i+2}}}{t_{i+\rho+1} - t_{i+1}}\right\}.$$

Since we assume that g_0 has at least 4 continuous partial derivatives, it seems reasonable that we restrict our parameter space to B-spline functions with at least 2 continuous partial derivatives. Thus we assume that the coefficients of the best B-spline function $g_n(x)$ is restricted in the set,

$$\begin{aligned}\mathcal{A}_n =&\{\text{vec}(a_1, ..., a_d) \in \mathbb{R}^{q_n^d} : a_i \in \mathbb{R}^{q_n}, \|a_i\|_\infty \leq A_1,\\ &\max_{1\leq j\leq q_n} |a_{i,j+1} - a_{i,j}| \leq A_2/q_n, \max_{1\leq j\leq q_n} |(a_{i,j+1} - a_{i,j}) - (a_{i,j} - a_{i,j-1})| \leq A_3/q_n^2\},\end{aligned}$$

for certain constants $A_1, A_2, A_3 > 0$. As a consequence of Equation (2.6), for functions $g(x) = \psi(x)^T a$, with $a \in \mathcal{A}_n$,

$$\|g\|_\infty = \sup_x |\psi(x)^T a| \leq \|a\|_\infty \leq A_1.$$

Additionally if we assume,

$$\frac{1}{q_n} \lesssim |t_{j+1} - t_j| \lesssim \frac{1}{q_n}, \qquad \text{for all } j, \tag{6.16}$$

a direct consequence is that,

$$\left\|\frac{d}{dx}g\right\|_\infty = \sup_x |\psi_{\rho-1}(x)^T a'| \lesssim A_2/q_n \cdot q_n \lesssim A_2.$$

In an analoguous manner, for the second derivative,

$$\left\|\frac{d^2}{dx^2}g\right\|_\infty = \sup_x |\psi_{\rho-2}(x)^T a''| \lesssim \frac{A_3}{q_n^2}q_n^2 \lesssim A_3.$$

Thus the B-spline functions have bounded partial derivatives up to the 2nd degree. A bounded derivative implies Lipschitz continuity, with the Lipschitz constant being equal to the maximum attained on the domain. We denote this as,

$$\begin{aligned}|\psi(x) - \psi(y)| &\leq K_1\|x - y\|_1\\ |\psi'(x) - \psi'(y)| &\leq K_2\|x - y\|_1.\end{aligned}$$

By the Cauchy-Schwarz inequality we have for matrices $\beta_1 = (I_d, C_1^T)^T, \beta_2 = (I_d, C_2^T)^T$, and for any X,

$$\begin{aligned}|\psi(\beta_1^T X) - \psi(\beta_2^T X)| &\leq K_1\|(\beta_1 - \beta_2)^T X\|_1\\ &\leq K_1\|\beta_1 - \beta_2\|_F\|X\|_1,\end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm, i.e.,

$$\|\beta_1 - \beta_2\|_F = \|\mathrm{vec}(\beta_1 - \beta_2)\|_2 = \|\mathrm{vec}(C_1 - C_2)\|_2.$$

Therefore,

$$|\psi(\beta_1^T X) - \psi(\beta_2^T X)| \leq K_1 \|X\|_1 \|\mathrm{vec}(C_1 - C_2)\|_2.$$

For simplicity we say that $L_1 = K_1 \sup_{X \in \mathcal{X}} \|X\|_1$. With similar reasoning we can show that for $L_2 = K_2 \sup_{X \in \mathcal{X}} \|X\|_1$,

$$|\psi'(\beta_1^T X) - \psi'(\beta_2^T X)| \leq L_2 \|\mathrm{vec}(C_1 - C_2)\|_2.$$

We briefly remind the reader of the assumptions we make for the results for the results in this section.

Assumption 1 $0 < \det(I^*) < \infty$, *where I^* is defined as,*

$$I^* = \{E[g_0'(\beta_0^T X)^{\otimes 2}]\}^{-1} E[(Y - g_0(\beta_0^T X)) g'(\beta_0^T X)]^{\otimes 2} \{E[g_0'(\beta_0^T X)^{\otimes 2}]\}^{-1}.$$

Assumption 2 *We assume that $X \in \mathcal{X}$ for some bounded subset $\mathcal{X}$ of $\mathbb{R}^p$, and $Y \in \mathcal{Y} \subset \mathbb{R}$ with,*

$$P|Y|^4 < \infty.$$

Assumption 3 *The function g_0 has at least 4 continuous (partial) derivatives.*

Assumption 4 *The number of knots is a non-decreasing sequence $q_n^d \to \infty$ such that,*

$$\frac{q_n^d \log(n)}{n} \to 0, \text{ as } n \to \infty. \tag{6.17}$$

Assumption 5

$$vec(C_0) \in \{x \in \mathbb{R}^{(p-d)d} : \|x\|_2 \leq r\} \equiv \mathcal{C},$$

for some constant r.

Assumption 6 *For each n,*

$$a_n \in \mathcal{A}_n.$$

Assumption 7 *Let $(t_{i,1}, ..., t_{i,q_n})$, for $i = 1, ..., d$ be the knot sequence along the i-th axis.*

$$1/q_n \lesssim |t_{i,j} - t_{i,j+1}| \lesssim 1/q_n.$$

Assumption 1 implies that the asymptotic distribution of $\sqrt{n}(\mathrm{vec}(\widehat{C}_n - C_0))$ is non-degenerate. Assumption 3 is a smoothness assumption on the regression function g_0. A consequence is that for two matrices $\beta_1 = (I_d, C_1^T)^T, \beta_2 = (I_d, C_2^T)^T$,

$$|g_0(\beta_1^T X) - g_0(\beta_2^T X)| \leq L_3 \|\mathrm{vec}(C_1 - C_2)\|_2, \tag{6.18}$$

for some constant $L_3 > 0$. Assumptions 5 and 6 ensure that the class of loss functions has finite entropy, and additionally due to Assumption 6 and 7 the sieves of B-splines are Lipschitz functions with Lipschitz derivatives.

6.5.2 Consistency

To use Theorem 6.4, we need to prove the consistency of $\widehat{C}_n$ for C_0. However, since our M-estimator involves the maximization over the parameter matrices C and B-spline function coefficients a, we show consistency of $(\widehat{C}_n, \widehat{a}_n)$ for (C_0, a_n). We apply Theorem 4.4. This involves demonstrating that,

$$||\mathbb{M}_n - M_n||_{\mathcal{C}\times\mathcal{A}_n} \to 0. \tag{6.19}$$

Define the class of loss functions as,

$$\mathcal{M}_n = \{m_{\beta,g}(X,Y) : g(x) = \psi(x)^T a, a \in \mathcal{A}_n, \beta = (I, C^T)^T, \mathrm{vec}(C) \in \mathcal{C}\}.$$

Note that,

$$E_P\|\mathbb{G}_n\|_{\mathcal{M}_n} = \sqrt{n}\|\mathbb{P}_n - P\|_{\mathcal{M}_n}.$$

If we show that the right-hand side of the maximal inequality in Equation (4.10) grows slower than $\sqrt{n}$, this implies that Equation (6.19) is satisfied. We prove that the entropy integral grows slower than $\sqrt{n}$ in Lemma 6.2.

Lemma 6.2 *If Assumptions 1-7 are satisfied, then,*

$$J_{[]}(1, \mathcal{M}_n, L_2) = \int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{M}_n, L_2)} d\epsilon \lesssim \sqrt{q_n^d \log(q_n)}.$$

Proof. The construction of the function class $\mathcal{M}_n$ suggests that we check whether the conditions of Theorem 4.3 hold, so that we can bound the bracketing entropy of $\mathcal{M}_n$ by the covering entropy of $\mathcal{C} \times \mathcal{A}_n$. Let $\beta_1 = (I_d, C_1^T)^T, \beta_2 = (I_d, C_2^T)^T$ be two parameter matrices in the Grassman-manifold, and $g_1(x) = \psi(x)^T a_1, g_2(x) = \psi(x)^T a_2$ be two B-spline functions of degree 3. For any (X, Y),

$$\begin{aligned}|(m_{\beta_1,g_1} - m_{\beta_2,g_2})(X,Y)| &= |2Y(g_1(\beta_1^T X) - g_2(\beta_2^T X)) + g_2(\beta_2^T X)^2 - g_1(\beta_1^T X)^2| \\ &= |(2Y - g_1(\beta_1^T X) - g_2(\beta_2^T X))(g_1(\beta_1^T X) - g_2(\beta_2^T X))| \\ &\leq (2|Y| + |\psi(\beta_1^T X)a_1| + |\psi(\beta_2^T X)a_2|)|\psi(\beta_1^T X)^T a_1 - \psi(\beta_2^T X)^T a_2|.\end{aligned}$$

By the Hölder inequality and property 2 in Section 2.3 we have for $i = 1, 2$,

$$|\psi(\beta_i^T X)^T a_i| \leq \|\psi(\beta_i^T X)\|_1 \cdot \|a_i\|_\infty \leq A_1 \tag{6.20}$$

Thus,

$$|(m_{\beta_1,g_1} - m_{\beta_2,g_2})(X,Y)| \leq 2(|Y| + A_1)|\psi(\beta_1^T X)^T a_1 - \psi(\beta_2^T X)^T a_2|.$$

We can bound the difference in B-spline functions on the right by applying the triangle inequality and then Hölder's inequality,

$$\begin{aligned}|\psi(\beta_1^T X)^T a_1 - \psi(\beta_2^T X)^T a_2| &= |\psi(\beta_1^T X)^T a_1 - \psi(\beta_1^T X)^T a_2 \\ &\quad + \psi(\beta_1^T X)^T a_2 - \psi(\beta_2^T X)^T a_2| \\ &\leq \|\psi(\beta_1^T X)\|_1 \|a_1 - a_2\|_\infty + \|\psi(\beta_1^T X) - \psi(\beta_2^T X)\|_1 \|a_2\|_\infty \\ &\leq \|a_1 - a_2\|_\infty + A_1 L_1 \|\mathrm{vec}(C_1 - C_2)\|_2.\end{aligned} \tag{6.21}$$

Let,

$$d^2((C_1, a_1), (C_2, a_2)) = \|a_1 - a_2\|_\infty + \|\mathrm{vec}(C_1 - C_2)\|_2.$$

Then,

$$|(m_{\beta_1, g_1} - m_{\beta_2, g_2})(X, Y)| \leq d((C_1, a_1), (C_2, a_2)) F(X, Y), \tag{6.22}$$

with $F(X, Y) = 2(|Y| + A_1) 2 \max\{1, A_1 L_1\} = 4(|Y| + A_1) \max\{1, A_1 L_1\}$. By applying Theorem 4.3 we obtain,

$$N_{[]}(2\epsilon \|F\|_{L_2(P)}, \mathcal{M}_n, L_2(P)) \leq N(\epsilon, \mathcal{C} \times \mathcal{A}_n, d).$$

Additionally,

$$\begin{aligned}\|F\|_{L_2}^2 &= \int_{\mathcal{X}, \mathcal{Y}} F(x, y)^2 dP(x, y) \\ &= \int_{\mathcal{X}, \mathcal{Y}} (4(|y| + A_1) \max\{1, A_1 L_2\})^2 dP(x, y) \\ &= 16 \max\{1, A_1 L_2\}^2 \int |y|^2 + 2A_1 |y|^2 + A_1^2 dP(x, y) \\ &= 16 \max\{1, A_1 L_2\}^2 (P|Y|^2 + 2A_1 P|Y| + A_1^2).\end{aligned}$$

Thus $\|F\|_{L_2} < \infty$ by Assumption 2. Furthermore the L_2-norm is lower bounded by,

$$\|F\|_{L_2}^2 \geq 16 A_1^2 \max\{1, A_1 L_1\}^2 > 0.$$

In conclusion, by Lemma 4.1, for any $\epsilon' = \frac{\epsilon}{2\|F\|_{L_2(P)}} > 0$,

$$\begin{aligned}\log N_{[]}(\epsilon', \mathcal{M}_n, L_2) &\leq \log N(\epsilon, \mathcal{C} \times \mathcal{A}_n, d) \\ &\leq \log \Big(N(\epsilon, \mathcal{C}, \|\cdot\|_2) \cdot N(\epsilon, \mathcal{A}_n, \|\cdot\|_\infty) \Big) \\ &\lesssim (p - d) d \log \left(\frac{r}{\epsilon} \right) + q_n^d \log \left(\sqrt{q_n^d} \frac{A_1}{\epsilon} \right) \\ &\lesssim q_n^d \log(q_n) + ((p - d) d + q_n^d) \log \left(\frac{\max\{r, A_1\}}{\epsilon} \right)\end{aligned}$$

We can now bound the entropy integral,

$$\begin{aligned}
J_{[]}(1, \mathcal{M}_n, L_2) &\lesssim \int_0^1 \sqrt{q_n^d \log(q_n) + ((p-d)d + q_n^d) \log\left(\frac{\max\{r, A_1\}}{\epsilon}\right)} d\epsilon \\
&\leq \sqrt{q_n^d \log(q_n)} + \sqrt{((p-d)d + q_n^d) \log(\max\{r, A_1\})} \\
&+ \sqrt{(p-d)d + q_n^d} \int_0^1 \sqrt{\log\left(\frac{1}{\epsilon}\right)} d\epsilon \\
&= \sqrt{q_n^d \log(q_n)} + \sqrt{(p-d)d + q_n^d \log(\max\{r, A_1\})} \\
&- \frac{\sqrt{\pi}}{2} \sqrt{(p-d)d + q_n^d} \\
&= O\Big(\sqrt{q_n^d \log(q_n)}\Big).
\end{aligned}$$

□

We can now readily apply the maximal inequality from Equation (4.10).

Theorem 6.6 *If the conditions of Lemma 6.2 are satisfied, we have that*

$$\|vec(\widehat{C}_n - C_0)\|_2 = o_P(1), \text{ and} \tag{6.23}$$

$$\sup_x |\widehat{g}_n(x) - g_0(x)| = o_P(1). \tag{6.24}$$

Proof. Recall that $\beta = (I_d, C^T)^T$ and $g(x) = \psi(x)^T a$ denote a parameter matrix and a B-spline of degree 3 on q_n^d knots, respectively. The second moment of the natural envelope function V_n of $\mathcal{M}_n$ has the following form,

$$\begin{aligned}
PV_n^2 &= P\left\{\sup_{m\in\mathcal{M}_n} m^2\right\} \\
&= P\left\{\sup_{C,a}(-Y^2 + 2Yg(\beta^T X) - (g(\beta^T X))^2)^2\right\} \\
&\leq P\left\{\sup_{C,a} Y^4 - 2Y^3 g(\beta^T X) + 3(Yg(\beta^T X))^2 - 2Y(g(\beta^T X))^3\right. \\
&\left.+ (g(\beta^T X))^4\right\} \\
&\leq P\left\{\sup_{C,a} |Y|^4 + |Y|^3|g(\beta^T X)| + 3|Y|^2|g(\beta^T X)|^2 + 2|Y||g(\beta^T X)|^3\right. \\
&\left.+ |g(\beta^T X)|^4\right\} \\
&\leq P\left\{|Y|^4 + |Y|^3\|a\|_\infty + 3|Y|^2 + 3|Y|^2\|a\|_\infty^2 + 2|Y|\|a\|_\infty^3 + \|a\|_\infty^4\right\},
\end{aligned} \tag{6.25}$$

where we used the bound in Equation (2.6) in the last inequality. This term does not grow in n, and is finite since $P|Y|^4 < \infty$, and $\|a\|_\infty \leq A_1$. In combination with Lemma 6.2 we conclude that,

$$\|\mathbb{P}_n - P\|_{\mathcal{M}_n} \to 0, \quad \text{as } n \to \infty.$$

Thus $\mathcal{M}_n$ is a Glivenko-Cantelli class. The other conditions of Theorem 4.4 are readily seen to hold. One of the criteria is that,

$$\begin{aligned}
M_n(\beta_0, g_n) \geq M_n(\beta, g), \text{ for all } \beta = (I_d, C^T)^T, \\
\text{vec}(C) \in \mathcal{C}; g(x) = \psi(\beta^T X)^T a, a \in \mathcal{A}_n.
\end{aligned} \tag{6.26}$$

Since (β_0, g_n) are defined as the maximizers of M_n, this holds.

The third and final requirement is that

$$\mathbb{M}_n(\widehat{\beta}_n, \widehat{g}_n) \geq \sup_{(\beta,g)} \mathbb{M}_n(\beta, g) - o_P(1). \tag{6.27}$$

Since $\widehat{\beta} = (I_d, \widehat{C}_n^T)^T$ and $\widehat{a}_n$ are found by maximizing $\mathbb{M}_n$, this requirement is satisfied. Thus we conclude that by Theorem 4.4,

$$d((\widehat{C}_n, \widehat{a}_n), (C_0, a_n)) \xrightarrow{P} 0. \tag{6.28}$$

The proof is then concluded since $\sup_x |g_n(x) - g_0(x)| \to 0$ by Equation (2.8), which is the bias of our spline estimator. □

6.5.3 Rate of Convergence

Proving consistency of $\widehat{C}_n$ for C_0 was sufficient concerning the convergence of the finite-dimensional part of our estimator. In contrast, we do require a rate of convergence of B-spline estimator to the "truth" g_0 in Theorem 6.4. In particular we show,

$$\|\widehat{g}_n - g_0\|_{L_2} = O_P(n^{-c_1}),$$

for some constant $c_1 > 0$. Since we are working with sieves[†] that grow larger in n, we first consider the rate of convergence of $\widehat{g}_n$ to the best estimator g_n in each sieve. By the triangle inequality we have,

$$\|\widehat{g}_n - g_0\|_{L_2} \leq \|\widehat{g}_n - g_n\|_{L_2} + \|g_n - g_0\|_{L_2}.$$

The second term on the right-hand side equals the estimation bias, and we know from Equation (2.8) that if the continuous partial derivatives of g_0 of order p exist,

$$\sup_x |g_n(x) - g_0(x)| = O_P(q_n^{-p}).$$

We bound the first term (the "variance", see our discussion in Section 4.3.2) by applying Theorem 4.5. This involves using the maximal inequality in Equation (4.10). Write $\beta_1 = (I_d, C_1^T)^T$, $\beta_2 = (I_d, C_2^T)^T$, and g_1, g_2 two B-spline functions of degree 3 on q_n^d knots. Define d_n as the non-negative function,

$$d_n((\beta_1, g_1), (\beta_2, g_2)) = \sqrt{\|\mathrm{vec}(C_1 - C_2)\|_2^2 + \|g_1 - g_2\|_{L_2}^2}.$$

Therefore we want to compute the entropy integral of the function class of centered loss functions,

$$\mathcal{M}_{n,\delta} = \{m_{\beta,g} - m_{\beta_0,g_n} : g = \psi(X)^T a, a \in \mathcal{A}_n, \beta = (I, C^T)^T, \mathrm{vec}(C) \in \mathcal{C}, \\ \tfrac{\delta}{2} < d_n((\beta, g), (\beta_0, g_n)) \leq \delta\}.$$

Fortunately, computing the entropy integral of $\mathcal{M}_{n,\delta}$ is easily reduced to computing the entropy integral of $\mathcal{M}_n$. We demonstrate this in the following lemma.

Lemma 6.3 *If Assumptions 1-7 are satisfied, then,*

$$J_{[]}(1, \mathcal{M}_{n,\delta}, L_2(P)) \lesssim \sqrt{q_n^d \log(q_n)}.$$

[†] The sieves correspond to the amount of knots q_n^d of the spline estimator.

Proof. Note that for any two elements $(m_{\beta_1,g_1} - m_{\beta_0,g_n}), (m_{\beta_2,g_2} - m_{\beta_0,g_n}) \in \mathcal{M}_{n,\delta}$,

$$|((m_{\beta_1,g_1} - m_{\beta_0,g_n}) - (m_{\beta_2,g_2} - m_{\beta_0,g_n}))(X,Y)| = |(m_{\beta_1,g_1} - m_{\beta_2,g_2})(X,Y)|.$$

By Equation (6.22),

$$|(m_{\beta_1,g_1} - m_{\beta_2,g_2})(X,Y)| \leq d((C_1,a_1),(C_2,a_2))F(X,Y).$$

The rest of the argument is the same as that of Lemma 6.2. □

Theorem 6.7 *We have that*

$$\sqrt{\frac{n}{q_n^d \log(q_n)}}\|\widehat{g}_n - g_n\|_{L_2} = O_P(1), \tag{6.29}$$

and

$$\sqrt{\frac{n}{q_n^d \log(q_n)}}\|vec(\widehat{C}) - vec(C_0)\|_2 = O_P(1). \tag{6.30}$$

Proof. We check that the conditions of Theorem 4.5 hold. The first condition holds if

$$\sup_{\delta/2 < d_n((\beta,g),(\beta_0,g_n)) \leq \delta} M_n(\beta,g) - M_n(\beta_0,g_n) \leq -\delta^2. \tag{6.31}$$

Let $\boldsymbol{v} = \begin{pmatrix} \text{vec}(C - C_0) \\ a - a_n \end{pmatrix}$. We can use a 3rd order Taylor expansion of $Pm_{\beta,g}$ around $(\text{vec}(C_0), a_n))$ to obtain,

$$Pm_{\beta,g} = Pm_{\beta_0,g_n} + \nabla_{\text{vec}(C),a} Pm_{\beta_0,g_n}\boldsymbol{v} + \frac{1}{2}\boldsymbol{v}^T \nabla^2_{\text{vec}(C),a} Pm_{\beta_0,g_n}\boldsymbol{v} + O(|\boldsymbol{v}|^3).$$

Since $(\text{vec}(C_0), a_n)$ maximize $Pm_{\beta,g}$, the first derivatives equal zero, and the second derivative evaluated at $(\text{vec}(C_0), a_n)$ is negative-definite. Consequently,

$$\begin{aligned}
M_n(\beta,g) - M_n(\beta_0,g_n) &= \frac{1}{2}\boldsymbol{v}^T \nabla^2_{\text{vec}(C),a} Pm_{\beta_0,g_n}\boldsymbol{v} + O(|\boldsymbol{v}|^3) \\
&= (\text{vec}(C) - \text{vec}(C_0))^T P[g_n'^{\otimes 2}(\beta_0^T X) + (Y - g_n(\beta_0^T X)) \\
&\quad \cdot g_n''(\beta_0^T X)](\text{vec}(C) - \text{vec}(C_0)) \\
&\quad - (a - a_n)^T P\psi(\beta_0^T X)\psi(\beta_0^T X)^T(a - a_n) + O(|\boldsymbol{v}|^3) \\
&= P[g_n'^{\otimes 2}(\beta_0^T X) + (Y - g_n(\beta_0^T X))g_n''(\beta_0^T X)]\|\text{vec}(C) - \text{vec}(C_0)\|_2^2 \\
&\quad - P(g(\beta_0^T X) - g_n(\beta_0^T X))^2 + O(|\boldsymbol{v}|^3) \\
&\leq -d^2((\beta,g),(\beta_0,g_n)) + O(d^3((\beta,g),(\beta_0,g_n))).
\end{aligned}$$

Denote the natural envelope function $\mathcal{M}_{n,\delta}$ by $G(X,Y)$. Then,

$$\begin{aligned} G(X,Y) &= \sup_{\beta,g}(m_{\beta,g}(X,Y) - m_{\beta_0,g_n}(X,Y)) \\ &\leq \sup_{\beta,g} d_n((\beta,g),(\beta_0,g_n))F(X,Y) \\ &\leq \delta F(X,Y). \end{aligned}$$

From the proof of Lemma 6.2, and the fact that $\delta > 0$, it follows that,

$$0 < \|G\|_{L_2}^2 \leq \delta^2 \|F\|_{L_2}^2 < \infty.$$

Combined with Lemma 6.3 it then follows that,

$$E_P\|\mathbb{G}_n\|_{\mathcal{M}_{n,\delta}} \lesssim \delta\sqrt{q_n^d \log(q_n)} = \phi_n(\delta). \tag{6.32}$$

This allows us to compute the rate of convergence r_n, which must satisfy

$$\begin{aligned} &r_n^2 \phi_n\left(\frac{1}{r_n}\right) \leq \sqrt{n}, \\ &r_n^2 \sqrt{q_n^d \log(q_n)} \frac{1}{r_n} \leq \sqrt{n}, \\ &r_n \leq \sqrt{\frac{n}{q_n^d \log(q_n)}}. \end{aligned}$$

The last two conditions we require are,

$$\mathbb{M}_n(\widehat{\beta}_n, \widehat{g}_n) \geq \mathbb{M}_n(\beta_0, g_n) - O_P(r_n^{-2}) \tag{6.33}$$

$$d_n((\widehat{\beta}_n, \widehat{g}_n), (\beta_0, g_n)) \xrightarrow{P} 0. \tag{6.34}$$

Condition (6.33) is automatically satisfied since $(\widehat{C}_n, \widehat{g}_n)$ is a maximizer of $\mathbb{M}_n$. Condition (6.34) is a consequence of Theorem 6.6. Therefore we conclude that

$$\sqrt{\frac{n}{q_n^d \log(q_n)}} d_n((\widehat{\beta}_n, \widehat{g}_n), (\beta_0, g_n)) = O_P(1). \tag{6.35}$$

□

Combining the result above and Equation (6.28) we obtain,

$$\begin{aligned} \|\widehat{g}_n - g_0\|_{L_2}^2 &\leq \|\widehat{g}_n - g_n\|_{L_2}^2 + \|g_n - g_0\|_{L_2}^2 \\ &= O_P\left(\sqrt{\frac{q_n^d \log(q_n)}{n}}\right) + O_P(q_n^{-p}). \end{aligned} \tag{6.36}$$

From the literature we know that the asymptotic variance term is typically of order $\frac{\sqrt{q_n^d}}{n}$ (see e.g. Huang (2003)). The additional $\log(q_n)$ term is due to the assumption that the best spline coefficients in each sieve $\mathcal{A}_n$ are bounded in the infinity norm. If one were to make the assumption that $\|a_n\|_2 \leq k$ for some constant k, we would obtain the optimal asymptotic variance with our approach. For our purposes the obtained rate of convergence is sufficient. Let,

$$q_n \approx n^{1/(2p+d)}.$$

Recall that we assume $p = 4$,

$$\|\widehat{g}_n - g_0\|_{L_2}^2 = O_P(n^{-2p/(2p+d)} \log(n)) = O_P(n^{-8/(8+d)} \log(n)). \tag{6.37}$$

Consequently, for any $c_1 < 4/(8+d)$ condition A1 of Theorem 6.4 is satisfied.

6.5.4 Stochastic Equicontinuity

We remind the reader that Condition A3 holds if for any $K > 0$,

$$\|\mathbb{G}_n(\tilde{m}_{\beta,g} - \tilde{m}_{\beta_0,g_n})\|_{\{\|\text{vec}(C-C_0)\|_2 \le Kn^{-c_1}, \|g-g_0\|_{L_2} \le Kn^{-c_1}\}} = o_P(1),$$

where $c_1 = 4/(8+d)$. Additionally from Equation (6.13) we know that,

$$\tilde{m}_{\beta,g} = \partial_\beta m_{\beta,g}.$$

We prove this by performing entropy calculations and using the maximal inequality in Equation (4.10). We define the function class,

$$\mathfrak{M}_n = \{\tilde{m}_{\beta,g} - \tilde{m}_{\beta_0,g_0} : \|\text{vec}(C - C_0)\|_2 \le Kn^{-c_1}, \|g - g_0\|_{L_2} \le Kn^{-c_1}\}.$$

Write the function class of the i-th component of the functions in $\mathfrak{M}_n$, $\tilde{m}^{(i)}_{\beta,g} - \tilde{m}^{(i)}_{\beta_0,g_0}$ as,

$$\mathfrak{M}^{(i)}_n = \{\tilde{m}^{(i)}_{\beta,g} - \tilde{m}^{(i)}_{\beta_0,g_0} : \|\text{vec}(C - C_0)\|_2 \le Kn^{-c_1}, \|g - g_0\|_{L_2} \le Kn^{-c_1}\}.$$

Recall that the covering number of the Cartesian product of two spaces is bound by the product of the covering number of the spaces. This implies that it is sufficient to find a bound on the entropy integral for all $\mathfrak{M}^{(i)}_n$, and the entropy integral of $\mathfrak{M}_n$ will have the same upper bound times $(p-d)d$.

In order to bound the entropy integral, we first bound the bracketing number using Theorem 4.3. We denote the parameter space of $\tilde{m}^{(i)}_{\beta,g} - \tilde{m}^{(i)}_{\beta_0,g_0}$ as,

$$\begin{aligned}\Theta_n = \{(\beta, g) : \beta = (I_d, C^T)^T; g(x) = \psi(x)^T a : \|\text{vec}(C - C_0)\|_2 \le Kn^{-c_1};\\ \|g - g_0\|_{L_2} \le Kn^{-c_1}\}.\end{aligned}$$

Lemma 6.4

$$J_{[]}(1, \mathfrak{M}_n, L_2(P)) \lesssim \sqrt{q_n^d \log(q_n)}.$$

Proof. Note that for any two elements $(\tilde{m}^{(i)}_{\beta_1,g_1} - \tilde{m}^{(i)}_{\beta_0,g_0}), (\tilde{m}^{(i)}_{\beta_2,g_2} - \tilde{m}^{(i)}_{\beta_0,g_0}) \in \mathfrak{M}^{(i)}_n$, and any X, Y,

$$|((\tilde{m}^{(i)}_{\beta_1,g_1} - \tilde{m}^{(i)}_{\beta_0,g_0}) - (\tilde{m}^{(i)}_{\beta_2,g_2} - \tilde{m}^{(i)}_{\beta_0,g_0}))(X, Y)| = |(\tilde{m}^{(i)}_{\beta_1,g_1} - \tilde{m}^{(i)}_{\beta_2,g_2})(X, Y)|.$$

From Equation (6.13) it follows that,

$$\begin{aligned}|(\tilde{m}^{(i)}_{\beta_1,g_1} - \tilde{m}^{(i)}_{\beta_2,g_2})(X, Y)| = |&-2(Y - g_1(\beta_1^T X))g_1^{(i)}(\beta_1^T X)\\ &+ 2(Y - g_2(\beta_2^T X))g_2^{(i)}(\beta_2^T X)|.\end{aligned}$$

If we write,

$$U_1 = |g_2'^{(i)}(\beta_2^T X) - g_1'^{(i)}(\beta_1^T X)| \\ U_2 = |g_1(\beta_1^T X) g_1'^{(i)}(\beta_1^T X) - g_2(\beta_2^T X) g_2'^{(i)}(\beta_2^T X)|, \tag{6.38}$$

we can bound the difference between two elements by,

$$|(\tilde{m}_{\beta_1,g_1}^{(i)} - \tilde{m}_{\beta_2,g_2}^{(i)})(X,Y)| \le 2|Y|U_1 + 2U_2.$$

Note that,

$$g'^{(i)}(\beta^T X) = \psi'^{(i)}(\beta^T X)^T a.$$

Thus,

$$\begin{aligned} U_1 &= |\psi'^{(i)}(\beta_2^T X)^T a_2 - \psi'^{(i)}(\beta_1^T X)^T a_1| \\ &= |\psi'^{(i)}(\beta_2^T X)^T a_2 - \psi'^{(i)}(\beta_2^T X)^T a_1 + \psi'^{(i)}(\beta_2^T X)^T a_1 - \psi'^{(i)}(\beta_1^T X)^T a_1| \\ &= |\psi'^{(i)}(\beta_2^T X)^T (a_2 - a_1) + (\psi'^{(i)}(\beta_2^T X) - \psi'^{(i)}(\beta_1^T X))^T a_1| \end{aligned}$$

Applying the triangle inequality allows us to consider both terms separately, and Hölder's inequality applied on both terms gives,

$$\begin{aligned} U_1 \le \; & \|\psi'^{(i)}(\beta_2^T X)\|_\infty \|a_2 - a_1\|_1 \\ & + \|\psi'^{(i)}(\beta_2^T X) - \psi'^{(i)}(\beta_1^T X)\|_1 \|a_1\|_\infty. \end{aligned}$$

Since we constructed our sieves so that the basis functions and their derivatives are Lipschitz, we can bound this by,

$$U_1 \le A_2 \|a_2 - a_1\|_\infty + A_1 L_2 \|\text{vec}(C_2 - C_1)\|_2. \tag{6.39}$$

We can bound U_2 by,

$$\begin{aligned} U_2 &= |\psi(\beta_1^T X)^T a_1 \psi'^{(i)}(\beta_1^T X)^T a_1 - \psi(\beta_2^T X)^T a_2 \psi'^{(i)}(\beta_1^T X)^T a_1 \\ &\quad + \psi(\beta_2^T X)^T a_2 \psi'^{(i)}(\beta_1^T X)^T a_1 - \psi(\beta_2^T X)^T a_2 \psi^{(i)}(\beta_2^T X)^T a_2| \\ &\le |(\psi(\beta_1^T X)^T a_1 - \psi(\beta_2^T X)^T a_2) \psi'^{(i)}(\beta_1^T X)^T a_1| \\ &\quad + |\psi(\beta_2^T X)^T a_2 (\psi'^{(i)}(\beta_1^T X)^T a_1 - \psi^{(i)}(\beta_2^T X)^T a_2)|. \end{aligned}$$

Using the Hölder inequality gives us,

$$\begin{aligned} U_2 \le \; & \|\psi(\beta_1^T X)^T a_1 - \psi(\beta_2^T X)^T a_2\|_1 \|\psi'^{(i)}(\beta_1^T X)^T a_1\|_\infty \\ & + \|\psi(\beta_2^T X)^T a_2\|_\infty \|\psi'^{(i)}(\beta_1^T X)^T a_1 - \psi'^{(i)}(\beta_2^T X)^T a_2\|_1. \end{aligned}$$

Using the bounds in Equation (6.21) and (6.39) we obtain,

$$\begin{aligned} U_2 \leq A_2 &\Big(\|a_2 - a_1\|_\infty + A_1 L_1 \|\text{vec}(C_1 - C_2)\|_2 \Big) \\ &+ A_1 \Big(A_2 \|a_2 - a_1\|_\infty + A_1 L_2 \|\text{vec}(C_2 - C_1)\|_2 \Big) \end{aligned} \tag{6.40}$$

Consequently, by collecting the terms in equation (6.39) and (6.40) we obtain,

$$|(\tilde{m}^{(i)}_{\beta_1, g_1} - \tilde{m}^{(i)}_{\beta_2, g_2})(X, Y)| \leq d((C_1, a_1), (C_2, a_2)) Q(X, Y),$$

where Q equals,

$$Q(X, Y) = \max\{2A_2 + A_1 A_2, 2A_1 L_2 + A_1 L_1\}.$$

We can now bound the bracketing number of $\mathfrak{M}^{(i)}_n$ using Theorem 4.3, and in the same vein as Lemma 6.2,

$$\log N_{[]}(\epsilon', \mathfrak{M}^{(i)}_n, L_2) \lesssim \log N(\epsilon, \mathcal{C} \times \mathcal{A}_n, d)$$

where $\epsilon' = \frac{\epsilon}{2\|Q\|_{L_2}} > 0$. Equivalently to the entropy calculation in the proof of Lemma 6.2, we obtain,

$$J_{[]}(1, \mathfrak{M}^{(i)}_n, L_2) \lesssim \sqrt{q^d_n \log(q_n)}.$$

□

Lemma 6.5 *(Stochastic equicontinuity) For $d < 8$ and any $K > 0$*

$$\sup_{\|vec(C-C_0)\|_2 \leq Kn^{-c_1}, \|g-g_0\|_{L_2} \leq Kn^{-c_1}} |\mathbb{G}(\tilde{m}_{\beta,g} - \tilde{m}_{\beta_0,g_0})| = o_P(1).$$

Proof. We know from Theorem 4.3 that $(\text{diam}\,\Theta_n)Q$ is an envelope function for the class $\mathfrak{M}$. From Theorem 6.7 it follows that

$$(\text{diam}\,\Theta)Q = O(n^{-c_1}),$$

with c_1 as defined in Equation (6.37). Using that $q_n \approx n^{1/(d+8)}$ and the maximal inequality in Equation (4.10) we obtain,

$$\|\mathbb{G}_n\|_{\mathfrak{M}} \lesssim \sqrt{n^{d/(8+d)}\log(n)} \cdot \sqrt{\frac{1}{n^{8/(8+d)}\log(n)}}.$$

Thus the condition holds as long as $d < 8$. □

6.5.5 Asymptotic Normality

We can use the results from the previous sections to demonstrate asymptotic normality and $\sqrt{n}$-consistency of $\widehat{C}_n$ for C_0.

Theorem 6.8 *If the conditions in Section 6.5.1 hold and $d < 8$, then,*

$$\begin{aligned}\sqrt{n}(vec(\widehat{C}_n) - vec(C_0)) = -\sqrt{n}(Pg_0'(\beta_0^T X)^{\otimes 2})^{-1} \\ \times \mathbb{P}_n(Y - g_0(\beta_0^T X)) + o_P(1)\end{aligned} \tag{6.41}$$

and $\sqrt{n}(vec(\widehat{C}_n) - vec(C_0))$ is asymptotically normal with mean 0 and variance $\{Pg'(\beta^T X)^{\otimes 2}\}^{-1}[P(Y - g(\beta^T X))^{\otimes 2}]\{Pg'(\beta^T X)^{\otimes 2}\}^{-1}$

Proof. We confirm that the conditions of Theorem 6.4 hold. From Theorem 6.6, Theorem 6.7, and Lemma 6.5 it follows that Condition A1 and A3 hold. Condition A2 is met by the assumption that $\sqrt{n}(\text{vec}(\widehat{C}_n) - \text{vec}(C_0))$ does not have a degenerate limit distribution.

Thus we only need to show that A4 holds, i.e., for some $c_2 > 1$ such that $c_1 c_2 > 1/2$, for all (β, g) in

$$\begin{aligned}\{(\beta,g) :& \beta = (I_d, C^T)^T, \text{vec}(C) \in \mathcal{C}, \|\text{vec}(C) - \text{vec}(C_0)\|_2 \leq Kn^{-c_1}; g(x) = \psi(x)^T a, \\ & a \in \mathcal{A}_n, \|g - g_0\|_\infty \leq Kn^{-c_1}\},\end{aligned}$$

the following condition holds,

$$\begin{aligned}&\Bigg|P\Bigg\{(\tilde{m}_{\beta,g}-\tilde{m}_{\beta_0,g_0})-(\partial_{11}m_{\beta_0,g_0}-\partial_{21}m_{\beta_0,g_0}[H^*])(\mathrm{vec}(C)-\mathrm{vec}(C_0))\\ &-\left(\partial_{12}m_{\beta_0,g_0}\left[\frac{g-g_0}{\|g-g_0\|_\infty}\right]-\partial_{22}m_{\beta_0,g_0}\left[H^*,\frac{g-g_0}{\|g-g_0\|_\infty}\right]\right)\\ &\cdot\|g-g_0\|_\infty\Bigg\}\Bigg|\\ &=o(\|\mathrm{vec}(C)-\mathrm{vec}(C_0)\|_2)+O(\|g-g_0\|_\infty^{c_2}).\end{aligned}$$

We have that,

$$P\tilde{m}_{\beta,g}=P(\partial_1 m_{\beta,g}-\partial_2 m_{\beta,g}[H^*])=P(\partial_1 m_{\beta,g}).$$

Thus if we use a Taylor expansion at (β_0,g_0), we obtain,

$$\begin{aligned}P\tilde{m}_{\beta,g}&=P(\partial_1 m_{\beta_0,g_0})+P(\partial_{11}m_{\beta_0,g_0})(\mathrm{vec}(C)-\mathrm{vec}(C_0))\\ &\quad+P(\partial_{12}m_{\beta_0,g_0}(g-g_0))[H^*]+\frac{1}{2}\partial_{111}m_{\beta_1,g_1}(\mathrm{vec}(C)-\mathrm{vec}(C_0))^2\\ &\quad+\partial_{121}m_{\beta_2,g_2}(\mathrm{vec}(C)-\mathrm{vec}(C_0))(g-g_0)[H]+\frac{1}{2}\partial_{122}m_{\beta_3,g_3}(g-g_0)^2[H_1,H_2],\end{aligned}\tag{6.42}$$

where $\beta_i=(I_d,C_i^T)^T,g_i$ are such that,

$$\|\mathrm{vec}(C_i)-\mathrm{vec}(C_0)\|_2\le Kn^{-c_1},\qquad \|g_i-g_0\|_\infty\le Kn^{-c_1},$$

and H,H_1,H_2 are proper tangents as described in Section 6.3. Consequently if we assume that the third derivatives of $Pm_{\beta,g}$ are bound in a neighbourhood of (β_0,g_0) (of the sizes in the display above), we have that Equation (6.42) equals,

$$\begin{aligned}&\Bigg|P\Bigg\{\frac{1}{2}\partial_{111}m_{\beta_1,g_1}(\mathrm{vec}(C)-\mathrm{vec}(C_0))^2+\partial_{121}m_{\beta_2,g_2}(\mathrm{vec}(C)-\mathrm{vec}(C_0))(g-g_0)\\ &+\partial_{122}m_{\beta_3,g_3}(g-g_0)^2\Bigg\}\Bigg|=O(\|\mathrm{vec}(C)-\mathrm{vec}(C_0)\|_2^2)+O(\|g-g_0\|_\infty^2)\\ &=o(\|\mathrm{vec}(C)-\mathrm{vec}(C_0)\|_2)+O(\|g-g_0\|_\infty^2).\end{aligned}$$

Thus all the conditions of Theorem 6.4 are satisfied. □

Note that the requirement $c_1 c_2 > 1/2$ is satisfied if,

$$\frac{2 \cdot 4}{8 + d} > 1/2.$$

This implies that $d < 8$. Since the computational complexity of our estimation method grows exponentially in d, it may desirable that such a condition is satisfied. If one however wants to generalize this technique, higher order B-splines (it is however necessary to assume g_0 has more continuous (partial) derivatives) can be considered. From Equation (6.37) one can then obtain the rate of convergence c_1, which allows for higher dimensions d.

6.6 A cross-validation criterion for unknown dimensions

Asymptotic normality and $\sqrt{n}$-convergence have been established when d_0 is known. In this section we propose a consistent estimator for d_0. In this setting we assume that,

$$Y = g_0(\beta_0^T X) + \epsilon, \tag{6.43}$$

where the dimension of the column space of β_0 is unknown. From Section 3.4 we know that in order to estimate the structural dimension, we require a cross-validation criterion. As we discussed in Section 3.4, the linearity of B-splines allows us to perform LOOCV, similar to Huang and Chiang (2017), with just one model estimation. As an alternative we show that a K-fold cross-validation criterion provides a consistent estimator. Let S_k, $k = 1, ..., K$, denote the index set of the elements (X_i, Y_i) in the k-th fold. For every fold $k = 1, ..., K$, we estimate $(\widehat{C}^{-k}, \widehat{g}^{-k})$ by a maximizer of,

$$\mathbb{P}_n^{(-k)} m_{d,\beta^{-k},g^{-k}} = \frac{1}{n - |S_k|} \sum_{i \notin S_k} -(Y_i - g_d^{-k}((\beta^{-k})^T X_i))^2.$$

For every fold $k = 1, ..., K$ we compute,

$$\begin{aligned}(\widehat{C}^{-k}, \widehat{g}^{-k}) &= \arg\max_{C,g} \frac{1}{n - |S_k|} \sum_{i \notin S_k} m_{\beta,g}(X_i, Y_i) \\ &= \arg\max_{C,g} \mathbb{P}_n^{(-k)} m_{\beta,g}.\end{aligned}$$

From our discussion in Section 3.4.2 we know that d_0 can be written as,

$$d_0 = \arg\max_{0 \le d \le p} M(d) = E[m_{\widehat{\beta}^{-k},\widehat{g}^{-k}}(X_j, Y_j)],$$

for any $j \in S_k$. By Theorem 6.6,

$$\max_{1 \le k \le K} d((\widehat{C}_{d_0}^{-k}, \widehat{g}_{d_0}^{-k}), (C_0, g_0)) = o_P(1). \tag{6.44}$$

We estimate d_0 by,

$$\widehat{d} = \arg\max_{0 \le d \le p} \mathbb{M}_n(d) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in S_k} m_{\widehat{\beta}_d^{-k}, \widehat{g}_d^{-k}}.$$

Theorem 6.9 *If $d_0 < 8$,*

$$\max_{0 \le d \le 8} |\mathbb{M}_n(d) - M(d)| = o_P(1), \tag{6.45}$$

and consequently $P(\widehat{d} = d_0) \to 1$.

Proof. Let $0 \le d \le \max\{p, 7\}$. Note that,

$$\begin{aligned} |\mathbb{M}_n(d) - M(d)| &= \left| \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in S_k} m_{\widehat{\beta}_d^{-k}, \widehat{g}_d^{-k}}(X_i, Y_i) - P(m_{\widehat{\beta}_d^{-k}, \widehat{g}_d^{-k}}) \right| \\ &= \left| \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|S_k|} \sum_{i \in S_k} m_{\widehat{\beta}_d^{-k}, \widehat{g}_d^{-k}}(X_i, Y_i) - P(m_{\widehat{\beta}_d^{-k}, \widehat{g}_d^{-k}}) \right| \end{aligned}$$

For each fold $k = 1, ..., K$,

$$\left| \frac{1}{|S_k|} \sum_{i \in S_k} m_{\widehat{\beta}_d^{-k}, \widehat{g}_d^{-k}}(X_i, Y_i) - P(m_{\widehat{\beta}_d^{-k}, \widehat{g}_d^{-k}}) \right| = \left| (\mathbb{P}_n^{(k)} - P) m_{\widehat{\beta}_d^{-k}, \widehat{g}_d^{-k}} \right|.$$

Analogous Lemma 6.2 it follows that,

$$\left| (\mathbb{P}_n^{(k)} - P) m_{\widehat{\beta}_d^{-k}, \widehat{g}_d^{-k}} \right| \lesssim \sqrt{\frac{q_n^d \log(q_n)}{|S_k|}}.$$

A direct consequence is that,

$$|\mathbb{M}_n(d) - M(d)| = o_P(1).$$

By Theorem 4.4 we then have $P(\widehat{d} = d_0) \to 1$, as $n \to \infty$. □

Thus we estimate d_0 by $\widehat{d}$, and we can then use the least squares technique to estimate the $(p - \widehat{d})\widehat{d}$ parameter matrix $\widehat{C}$. Since the column space of $\widehat{\beta}$, $S(\widehat{\beta}) = (I_{\widehat{d}}, \widehat{C}^T)^T$, has a random dimension, it makes sense to consider large sample properties of the projection matrix P_β, which is defined as the orthogonal projection operator onto $S(\beta)$, i.e.,

$$P_\beta = \beta(\beta^T \beta)^{-1} \beta^T.$$

Let $S_{\beta_0} = (Y - g(\beta_0^T X)) g'(\beta_0^T X)$, $V_{\beta_0} = E[g'(\beta_0^T X)^{\otimes 2}]$, and $\text{vec}(A_0) = V_{\beta_0}^{-1} S_{\beta_0}$. As a consequence the asymptotic variance of $\widehat{C}_{d_0}$ can be written as $E[(\text{vec}(A_0))^{\otimes 2}]$. We determine the asymptotic distribution of the projection matrix of $\widehat{\beta}_{d_0}$ in the following theorem.

Theorem 6.10 *If we let 0_{d_0} denote the $d_0 \times d_0$ matrix with only zero entries, we have that,*

$$\sqrt{n}(vec(P_{\widehat{\beta}_{d_0}} - P_{\beta_0})) \xrightarrow{d} N(0, \Sigma_0), \qquad as\ n \to \infty, \tag{6.46}$$

where $\Sigma_0 = E[(vec((I_p - P_{\beta_0})(0_{d_0}, A_0^T)^T(\beta_0\beta_0)^{-1}\beta_0^T + \beta_0(\beta_0^T\beta_0)^{-1}(0_{d_0}, A_0^T)(I_p - P_{\beta_0})))^{\otimes 2}]$.

Proof. Let $C \in \mathbb{R}^{(p-d)\times d}$. The projection operator P maps C to,

$$C \mapsto \beta(\beta^T\beta)^{-1}\beta^T,$$

where $\beta = (I_d, C^T)^T$. Theorem 6.8 gives,

$$\sqrt{n}\text{vec}(\widehat{C}_{d_0} - C_0) \to N(0, E[(\text{vec}(A_0))^{\otimes 2}])$$

Consequently,

$$\sqrt{n}\text{vec}(\widehat{\beta}_{d_0} - \beta_0) \xrightarrow{d} N(0, E[(\text{vec}((0_{d_0}, A_0)))^{\otimes 2}]).$$

The asymptotic distribution achieved in Huang and Chiang (2017) differs slightly from ours due to the asymptotic covariance we find for $\widehat{\beta}_{d_0} - \beta_0$. The derivative of the projection operator with respect to β is given by,

$$\begin{aligned} C \mapsto\ & (\beta^T\beta)^{-1}\beta^T - \beta(\beta^T\beta)^{-1}(\beta + \beta^T)(\beta^T\beta)^{-1}\beta^T + \beta(\beta^T\beta)^{-1} \\ &= (\beta^T\beta)^{-1}\beta^T - \beta(\beta^T\beta)^{-1}P_\beta - P_\beta(\beta^T\beta)^{-1}\beta^T + \beta(\beta^T\beta)^{-1} \\ &= (I_p - P_\beta)(\beta^T\beta)^{-1}\beta^T + \beta(\beta^T\beta)^{-1}(I_p - P_\beta). \end{aligned}$$

If we then apply the multivariate Delta method, we obtain,

$$\begin{aligned} \sqrt{n}\text{vec}(P_{\widehat{\beta}_{d_0}} - P_{\beta_0}) \xrightarrow{d} N(0, E[(\text{vec}((I_p - P_{\beta_0})(0_{d_0}, A_0^T)^T(\beta_0^T\beta_0)^{-1}\beta_0^T \\ + \beta_0(\beta_0^T\beta_0)^{-1}(0_{d_0}, A_0^T)^T(I_p - P_{\beta_0})))^{\otimes 2}]). \end{aligned}$$

□

If we let E denote the event,

$$E = \{\widehat{d} = d_0\}.$$

Since $1(E) + 1(E^c) = 1$, for all $\epsilon > 0$, and

$$P(\|\sqrt{n}\text{vec}(P_{\widehat{\beta}} - P_{\beta_0})\|_2 1(E^c) > \epsilon) \leq P(E^c),$$

as a consequence of Theorem 6.10 we have,

$$\sqrt{n}\text{vec}(P_{\widehat{\beta}} - P_{\beta_0}) = \sqrt{n}\text{vec}(P_{\widehat{\beta}_{d_0}} - P_{\beta_0})1(E) + o_P(1). \tag{6.47}$$

Consequently, $P_{\widehat{\beta}} - P_{\beta_0}$ has the same asymptotic distribution as $P_{\widehat{\beta}_{d_0}} - P_{\beta_0}$.

Chapter 7

Finite Sample Performance

In this chapter we assess the performance of our estimation method in various models, and discuss some of the drawbacks and advantages. The following regression models are used for the benchmark:

$$
\begin{aligned}
&\text{M1.} \qquad Y = (\beta_0^T X)^3 + \epsilon \text{ with } \epsilon \sim N(0, 0.2),\\
&\text{M2.} \qquad Y = \cos(2X_1) - \cos(X_2) + \epsilon \text{ with } \epsilon \sim N(0, 0.2),\\
&\text{M3.} \qquad Y = X_1 + \cos(X_2) + e^{X_3} + \epsilon \text{ with } \epsilon \sim N(0, 0.2).
\end{aligned}
$$

Let e_j denote the j-th basis vector in $\mathbb{R}^{10}$. The corresponding basis matrices are then respectively $(1,1,1,0,...,0)^T \in \mathbb{R}^{10}$, $(e_1, e_2) \in \mathbb{R}^{10\times 2}$, $(e_1, e_2, e_3) \in \mathbb{R}^{10\times 3}$. The first and third model are chosen arbitrarily, and the second model corresponds to model M2 in Section 4.2 of Huang and Chiang (2017). Since they also estimate the CMS for this model, it allows us to compare the two techniques in terms of performance and computation time. In these simulations, we used $p = 10$ features. We let $X_i = (X_{i1}, ..., X_{ip})$, where

$$X_i = \Sigma_X^{1/2} w_i.$$

Σ_X is a $p \times p$ matrix with (i, j)-th entry $.5^{|i-j|}$ and $w_i = (w_{i1}, ..., w_{ip})^T$ has entries w_{ij} independently standard normally distributed (Normal) or uniformly on $[-\sqrt{3}, \sqrt{3}]$ (Uniform). The third distribution used for X is a mixture normal distribution according to $N(2 \cdot e_j, \Sigma_X)$, $j = 1, 10$, with probability $1/2$ each. We ran 1000 simulations of each model combination and different distributions of X for sample sizes $n = 100, 200, 400$. Due to the random dimensions of $\beta_{\hat{d}}$, we evaluated the estimation accuracy using the projection matrices, i.e.

$$\Delta(\beta_d, \beta_0) = \|P_{\beta_d} - P_{\beta_0}\|_2,$$

where $\|\cdot\|_2$ here denotes the spectral norm.

7.1 Computation scheme

In Table 7.1 and 7.2 we summarized the results from the 1000 simulation studies we performed. In Table 7.3 we included the results from the simulation studies performed by Huang and Chiang (2017). The minimization of $CV(d, C_d, g_n)$ was carried out by the following steps:

1. Set $CV(0) = \sum_{i=1}^{n}(Y_i - \overline{Y}^{-i})^2$ and $CV(p+1) = \infty$.
2. Set $d = 1$, and compute
$$(\widehat{C_d}, \widehat{g}_n) = \underset{C_d, g}{\arg\min}\, CV(d, C_d, g)$$
Set $CV(d) = CV(d, \widehat{C_d}, \widehat{q_n^d})$.
3. If $CV(d) < CV(d-1)$, set $d = d+1$ and go to step 1. If $CV(d) \geq CV(d-1)$, estimate $(d_0, C_0, q_n^{d_0})$ by $(d-1, \widehat{C}_{d-1}, \widehat{q_n^{d-1}})$. If $d = 0$, estimate $E(Y|X)$ by the sample mean, $\overline{Y}$.

For large d_0 or n, this scheme can be computationally intensive. The optimization of this scheme, in particular over the number of knots, is out of the scope of this thesis. As such, the results in the tables below are mostly meant to demonstrate that in a reasonable amount of time, one can obtain satisfactory results. Additionally a one-to-one comparison is hard to perform, however we can see that for $d_0 = 2$ our estimation of the true dimension performs adequately compared to that of Huang and Chiang. They do seem to achieve lower errors, but this could be due to the convergence criterion applied. Other factors at play, are the initial points for C and bandwidth at which the optimization scheme begins. In our own findings we realized that the number of degrees of freedom seems to play a large role in the performance of the optimization scheme. The computation times in the simulations performed in Huang and Chiang for the CMS estimation are not mentioned. It is also noteworthy that the estimation of d_0 is worse when d_0 is larger. Whereas for $d_0 = 1$, a sample size of $n = 100$ seems to perform satisfactory for the normal and uniformly distributed X, for $d_0 = 3$ we need at least $n = 200$ to obtain a comparable performance. Additionally all estimations seem to drop in efficiency when X is a mixture of normal distributions. The performance does steadily improve as the sample size grows, both in terms of structural dimension estimates and estimation accuracies.

			Proportions of $\hat{d}$						
Model	dist. of X	n	0	1	2	3	4	5	
M1 ($d_0 = 1$)	Normal	100	0	969	31	0	0	0	
		200	0	988	12	0	0	0	
		400	0	1000	0	0	0	0	
	Uniform	100	0	1000	0	0	0	0	
		200	0	1000	0	0	0	0	
		400	0	999	1	0	0	0	
	Mixture normal	100	0	647	349	4	0		
		200	0	848	152	0	0	0	
		400	0	858	142	0	0	0	
M2 ($d_0 = 2$)	Normal	100	1	6	893	101	0	0	
		200	0	54	945	1	0	0	
		400	0	33	967	0	0	0	
	Uniform	100	0	117	819	64	0	0	
		200	0	64	934	2	0	0	
		400	0	23	977	0	0	0	
	Mixture normal	100	0	202	685	113	0	0	
		200	0	118	754	128	0	0	
		400	0	84	823	93	0	0	
M3 ($d_0 = 3$)	Normal	100	0	510	7	483	0	0	
		200	0	269	0	730	1	0	
		400	0	84	0	884	32	0	
	Uniform	100	0	259	2	739	0	0	
		200	0	58	0	942	0	0	
		400	0	26	0	974	0	0	
	Mixture normal	100	0	487	7	506	0	0	
		200	0	236	2	759	3	0	0
		400	0	91	0	884	25	0	0

Table 7.1: *The proportions of 1000 structural dimension estimates of 1000 CMS estimates under models M1-M3*

Model	dist. of X	n	$\Delta(\beta_{\hat{d}}, \beta_0)$	$\Delta(\beta_{d_0}, \beta_0)$	t (seconds)
M1 ($d_0 = 1$)	Normal	100	0.449 (0.16)	0.442 (0.15)	12
		200	0.328 (0.14)	0.324 (0.13)	18
		400	0.254 (0.11)	0.254 (0.11)	34
	Uniform	100	0.363 (0.13)	0.363 (0.13)	9
		200	0.258 (0.11)	0.258 (0.11)	16
		400	0.229 (0.10)	0.229 (0.10)	34
	Mixture normal	100	0.57 (0.353)	0.37 (0.199)	30
		200	0.33 (0.306)	0.23 (0.148)	52
		400	0.24 (0.207)	0.15 (0.091)	248
M2 ($d_0 = 2$)	Normal	100	0.53 (0.208)	0.47 (0.135)	22
		200	0.36 (0.195)	0.321 (0.121)	63
		400	0.34 (0.177)	0.27 (0.120)	115
	Uniform	100	0.608 (0.22)	0.516 (0.13)	19
		200	0.416 (0.11)	0.392 (0.10)	51
		400	0.340 (0.133)	0.324 (0.09)	94
	Mixture normal	100	0.73 (0.214)	0.61 (0.21)	24
		200	0.78 (0.191)	0.56 (0.137)	48
		400	0.58 (0.231)	0.48 (0.137)	106
M3 ($d_0 = 3$)	Normal	100	0.618 (0.39)	0.409 (0.26)	36
		200	0.429 (0.35)	0.319 (0.11)	75
		400	0.303 (0.15)	0.215 (0.05)	239
	Uniform	100	0.41 (0.357)	0.197 (0.060)	50
		200	0.22 (0.202)	0.17 (0.059)	89
		400	0.18 (0.145)	0.16 (0.057)	160
	Mixture normal	100	0.59 (0.398)	0.21 (0.062)	34
		200	0.39 (0.199)	0.20 (0.039)	81
		400	0.27 (0.179)	0.17 (0.083)	201

Table 7.2: *The means (standard deviations) of 1000 estimation accuracies, and the median computation times (seconds) of 1000 CS estimates under models M1-M3.*

			Proportions of $\widehat{d}$					
Model	dist. of X	n	0	1	2	3	4	5
M2 ($d_0 = 2$)	Normal	100	23	187	716	54	13	7
		200	0	20	944	36	0	0
		400	0	0	979	21	0	0
	Uniform	100	0	61	741	135	42	21
		200	0	0	971	29	0	0
		400	0	0	999	1	0	0
	Mixture normal	100	48	184	739	24	1	4
		200	0	32	938	30	0	0
		400	0	0	981	19	0	0
Model	dist. of X	n	$\Delta(\beta_{\hat{d}}, \beta_0)$	$\Delta(\beta_{d_0}, \beta_0)$				
M2 ($d_0 = 2$)	Normal	100	0.42 (0.375)	0.24 (0.172)				
		200	0.19 (0.211)	0.15 (0.103)				
		400	0.14 (0.135)	0.12 (0.067)				
	Uniform	100	0.42 (0.348)	0.24 (0.120)				
		200	0.17 (0.151)	0.15 (0.067)				
		400	0.08 (0.041)	0.08 (0.028)				
	Mixture normal	100	0.38 (0.379)	0.22 (0.173)				
		200	0.18 (0.225)	0.14 (0.119)				
		400	0.12 (0.133)	0.10 (0.064)				

Table 7.3: *Central mean subspace estimation summary of 1000 estimates under model M2, using the semiparametric estimator from Huang and Chiang (2017).*

Conclusion and Discussion

In this thesis, we developed a semi-parametric estimation method to estimate the SDR model. In particular we estimate the dimension of the smallest DR space using a cross-validation criterion, and then estimate the central mean subspace and conditional mean simultaneously. The biggest advantage of our estimation technique is that one can use the faster computation of the estimation criterion to obtain asymptotic results equivalent to the existing literature. In simulations the performance of the estimation method has been demonstrated to be satisfactory with reasonable computation time. The problem setting allowed us to use techniques from semiparametric statistics, and we utilized results on semiparametric M-estimation in order to prove our main result. The broadness and applicability of the literature leaves a lot of avenues unexplored which might be interesting for future research. The main one being the computational complexity - which plays a rather small role in our theoretical developments of the method but is important for practical use. Due to the exponential growth of the computations in the number of dimensions, for large d_0, p or n problems are anticipated. When d_0 is large, the curse of dimensionality plays a role. A large value of p or n are typically seen in big data settings. An efficient optimization algorithm is necessary to overcome these computational difficulties. It might be of interest to generalize the estimation of the central mean subspace to that of the central subspace. Here one would use the same reasoning as we invoked in Chapter 3.2 in order to derive the cross-validation criterion, and the same one used by Huang and Chiang (2017). The semiparametric framework we introduced in Chapter 6 could then be used to prove asymptotic results, accompanied with the M-estimation theory discussed in Chapter 4.

Bibliography

[1] C.-T. Chiang and M.-Y. Huang, "New estimation and inference procedures for a single-index conditional distribution model," *Journal of Multivariate Analysis*, vol. 111, pp. 271–285, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X12000942

[2] R. D. Cook, *Regression Graphics: Ideas for Studying Regressions through Graphics*. Springer, New York, NY, 1995.

[3] ——, "Testing predictor contributions in sufficient dimension reduction," *Ann. Statist.*, vol. 32, no. 3, pp. 1062–1092, 06 2004. [Online]. Available: https://doi.org/10.1214/009053604000000292

[4] R. Cook and B. Li, "Dimension reduction for conditional mean in regression," *The Annals of Statistics*, vol. 30, no. 2, pp. 455 – 474, 2002. [Online]. Available: https://doi.org/10.1214/aos/1021379861

[5] C. de Boor, "Splines as linear combinations of b-splines," *Approximation Theory II*, 01 1976.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[7] J. Z. Huang, "Local asymptotics for polynomial spline regression," *The Annals of Statistics*, vol. 31, no. 5, pp. 1600 – 1635, 2003. [Online]. Available: https://doi.org/10.1214/aos/1065705120

[8] M.-Y. Huang and C.-T. Chiang, "An effective semiparametric estimation approach for the sufficient dimension reduction model," *Journal of the American Statistical Association*, vol. 112, no. 519,

pp. 1296–1310, 2017. [Online]. Available: https://doi.org/10.1080/01621459.2016.1215987

[9] M. Kosorok, *Introduction to Empirical Processes and Semiparametric Inference*, ser. Springer Series in Statistics. Springer-Verlag New York, 2008.

[10] K.-C. Li, "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, vol. 86, no. 414, pp. 316–327, 1991. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1991.10475035

[11] ——, "On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma," *Journal of the American Statistical Association*, vol. 87, no. 420, pp. 1025–1039, 1992. [Online]. Available: http://www.jstor.org/stable/2290640

[12] S. Ma and M. R. Kosorok, "Robust semiparametric m-estimation and the weighted bootstrap," *Journal of Multivariate Analysis*, vol. 96, no. 1, pp. 190 – 217, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0047259X04001812

[13] Y. Ma and L. Zhu, "A review on dimension reduction," *International Statistical Review / Revue Internationale de Statistique*, vol. 81, no. 1, pp. 134–150, 2013. [Online]. Available: http://www.jstor.org/stable/43298809

[14] ——, "On estimation efficiency of the central mean subspace," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 5, pp. 885–901, 2014. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12044

[15] ——, "Efficient estimation in sufficient dimension reduction," *Ann. Statist.*, vol. 41, no. 1, pp. 250–268, 02 2013. [Online]. Available: https://doi.org/10.1214/12-AOS1072

[16] I. J. Schoenberg, *Contributions to the Problem of Approximation of Equidistant Data by Analytic Functions*. Boston, MA: Birkhäuser Boston, 1988, pp. 3–57. [Online]. Available: https://doi.org/10.1007/978-1-4899-0433-1_1

[17] G. Seber and A. Lee, *Linear Regression Analysis*, ser. Wiley Series in Probability and Statistics. Wiley, 2012. [Online]. Available: https://books.google.nl/books?id=X2Y6OkXl8ysC

[18] X. Shen and W. H. Wong, "Convergence rate of sieve estimates," *Ann. Statist.*, vol. 22, no. 2, pp. 580–615, 06 1994. [Online]. Available: https://doi.org/10.1214/aos/1176325486

[19] A. A. Tsiatis, *Semiparametric Theory and Missing Data*, ser. Springer Series in Statistics. Springer, New York, NY, 2006.

[20] A. W. van der Vaart, *Asymptotic Statistics*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[21] A. W. van der Vaart and J. A. Wellner, *Weak Convergence*. New York, NY: Springer New York, 1996. [Online]. Available: https://doi.org/10.1007/978-1-4757-2545-2_3

[22] L. Wasserman, *All of Nonparametric Statistics*, ser. Springer Series in Statistics. Springer, New York, NY, 2005.

[23] Y. Xia, H. Tong, W. K. Li, and L.-X. Zhu, "An adaptive estimation of dimension reduction space," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 363–410, 2002. [Online]. Available: http://www.jstor.org/stable/3088779

[24] X. Yin and R. Cook, "Dimension reduction for the conditional kth moment in regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, pp. 159 – 175, 05 2002.