Model Selection, Uniform Inference and Nonparametric Regression

Alexis De Boeck

Faculty of Economics

University of Cambridge

This thesis is submitted for the degree of Doctor of Philosophy

November 2019

UNIVERSITY OF CAMBRIDGE

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution.

It does not exceed the prescribed word limit for the Faculty of Economics Degree Committee.

Acknowledgments

I would like to thank my supervisor, Prof. Oliver Linton, for his guidance. I am indebted to the Economic and Social Research Council for fully funding me with an ESRC Award during three out of the four years of the PhD programme.

A big thank you goes out to the staff in the Faculty of Economics, in particular to Cherie, Dawn, Nathan and Silvana. Without them the paperwork would have been impossible to navigate and the Faculty itself would not have been such a bright place to work.

Finally, I would like to thank my parents, David and Caroline, for their endless support and encouragement. With lots of love, I dedicate this thesis to you.

Alexis De Boeck — Cambridge, November 2019

Abstract

Model Selection, Uniform Inference and Nonparametric Regression

Alexis De Boeck

Model selection in the nonparametric regression model is inevitable since any nonparametric estimator requires tuning parameters to be specified in order for it to be feasible. It is, however, standard practice to carry over the theory of nonparametric estimators when the model is fixed to the case where the tuning parameters are no longer fixed, but chosen by, possibly, data-driven model selection algorithms. This theory is not necessarily valid as the model selection step is not taken into account. This thesis contributes to the nonparametric econometrics and statistics literature and, in particular, to the theory of series estimators, by showing that such estimators have desirable properties and that valid inference is possible even when a model-selection step precedes estimation.

The first chapter is concerned with K-fold cross-validation and shows that the cross-validated least-squares estimator predicts the response equally well as the unfeasible best-linear predictor whose dimension may diverge with the sample size. This property, known as risk consistency, is uncommon in econometrics, but it has the benefit that

it holds under few and very weak conditions. The risk-consistency result crucially relies on the non-asymptotic analysis of the difference between the prediction error of the cross-validated estimator and the best-linear predictor. As the dimension of the parameters may diverge, this set-up analyses both the high-dimensional linear model as well as the nonparametric regression model which reduces the need for duplicate theories. An extensive Monte Carlo experiment corroborates the theoretical results by showing that the non-asymptotic bound becomes arbitrarily small as the sample size diverges.

The second chapter returns to more classical statistics and econometrics by studying the uniform consistency of the series estimator for the conditional mean function and its linear functionals. The uniformity holds both in the support of the covariates as well as the models considered. Under high-level assumptions, a non-asymptotic linearisation result delivers uniform rates of convergence for the series estimator. By verifying the high-level assumptions, case-specific rates can easily be derived. For example, the series estimator attains, up to a small logarithmic penalty, the minimax rate of convergence for functions lying in a Hölder ball.

The results from the second chapter form the basis for the inference procedure proposed in the final chapter in order to construct valid uniform confidence bands for the series estimator. The uniform confidence bands are valid in the sense that they control the asymptotic size for the conditional mean function, or its linear functionals, seen as a process in the covariates and the models considered. Given that the results hold uniformly over the models considered, the inference procedure is valid regardless of which model-selection algorithm delivers the final model used to estimate the parameters of interest. The key quantity is the maximal t-statistic correctly studentised using an estimator for the standard error. The theory relies on the uniform linearisation result from chapter two and the concept of strong approximations, or couplings, as the limit distribution of the

maximal t-statistic does not exist. A Monte Carlo study establishes that the uniform confidence bands have the correct coverage even in finite samples. The chapter concludes with an application testing for shape restrictions on the demand function for gasoline in the US using a cross-validated series estimator.

Contents

List of Figures

List of Tables

1 Introduction

At the core of this thesis lies the nonparametric regression model

$$Y = g(X) + \varepsilon, \tag{1.1}$$

where Y is a real-valued response, X is a set of covariates and ε is the residual. Series estimators, as explained in Section 1.2, offer an attractive method to estimate the nonlinearity in (1.1) if a linear parametric specification is not appropriate. Before introducing the concept of series estimation, it is important to set out the notation which is used throughout.

1.1. Notation

The consistency of the notation is vital in order to make the arguments in the text and proofs easy to follow. Unless explicitly mentioned, all quantities are indexed by the sample size n, but this dependence is dropped in order to avoid repeatedly making use of double subscripts. Much of the notation is, in fact, consistent with the notation in the empirical process literature. For $X_1, \ldots, X_n \in \mathcal{X}$ i.i.d. random variables drawn from $X \sim P$, let $\mathcal{F}$ be a sequence of classes of functions $\mathcal{X} \to \mathbf{R}$. The empirical mean indexed by $\mathcal{F}$ is

$$\mathbb{E}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i), \tag{1.2}$$

and denote the empirical process by

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(X_i) - \mathrm{E}\, f(X_i). \tag{1.3}$$

The randomised empirical process is denoted by

$$\mathbb{G}_n^o f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varsigma_i f(X_i), \tag{1.4}$$

where ς_i is an i.i.d. Rademacher random variable with $P(\varsigma_i = 1) = P(\varsigma_i = -1) = \frac{1}{2}$.

The set of natural numbers and real numbers are respectively written as $\mathbf{N}$ and $\mathbf{R}$. The model set $\mathcal{M}_n \subseteq \mathbf{N}$ is loosely defined as $\{\underline{m}, \cdots, \overline{m}\}$ where $\underline{m}$ and $\overline{m}$ are respectively the minimum and maximum of $\mathcal{M}_n$ and its cardinality is $\tilde{m} = |\mathcal{M}_n|$. Its members will be made explicit in the text when necessary. The letter m, without any accent marks nor sub- or superscripts, is used to either index the model set or to denote a generic $m \in \mathbf{N}$, but its meaning will be clear from the context. This is to avoid confusion between the various letters m introduced in the text. The indexing set $\mathcal{I}_n$ is exclusively used to denote the Cartesian product of the support $\mathcal{X}$ and the model set $\mathcal{M}_n$. The unit sphere in $\mathbf{R}^m$ is denoted by $\mathbb{S}^{m-1}$.

The notation $\|a\|$ or $\|A\|$ denotes the ℓ_2-norm of a vector a or the spectral norm of a matrix A such that $\|A\| = \sqrt{\mathrm{tr}\, A'A}$. The minimal and maximal eigenvalues of a matrix A are denoted by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$. For any real-valued random vector $x \in \mathbf{R}^m$, $\|x\|_p = (\mathrm{E}\,|x|^p)^{1/p}$ where for $p = \infty$ this is $\|x\|_\infty = \mathrm{E} \max_{1 \le j \le m} |x_j|$. Additionally, let $\|x\|_{p,n} = (\mathbb{E}_n |x|^p)^{1/p}$ where the expectation is taken under the empirical measure and let $\|f\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |f|$ for some arbitrary function $f \in \mathcal{F}$. The covariance matrix is denoted by $\Sigma_m = \mathrm{E}[Z_{i,m} Z_{i,m}']$ for an i.i.d. random vector $Z_{i,m} \in \mathbf{R}^m$. Its sample analogue is written as $\hat{\Sigma}_m$.

I write $a \lesssim b$ if there exists a universal $C > 0$ such that $a \leq Cb$. Similarly, $a \lesssim_P b$ if for some $C > 0$, $a \leq Cb$ on an event with probability $1 - o(1)$. Let $a \asymp b$ denote that $c < a/b < C$. The letters c and C are constants which do not depend on n, but may change meaning depending on where in the text they are used. Let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Finally, non-random sequences converging to zero which are only used as prove devices are given by δ_n.

1.2. Series Estimation

Let me conclude the introduction with a brief overview of series, or linear sieve, estimation of the regression model in (1.1). The method of sieves (Grenander 1981) relies upon approximating the (possibly) infinite-dimensional parameter $g \in \mathcal{G}$ by finite-dimensional parameters whose dimension may grow with the sample size. The sieve space is

$$\mathcal{G}_{m,n} = \left\{ g(\cdot) : \sum_{j=1}^{m} \beta_j^m Z_j^m(\cdot), \beta \in \mathbf{R}^m \right\}, \tag{1.5}$$

where $Z_j^m : \mathcal{X} \to \mathbf{R}$ and $Z_m(\cdot) = (Z_1^m(\cdot), \ldots, Z_m^m(\cdot))'$ is a system of orthonormal basis transformation on $\mathcal{X}$. Equation (1.5) reveals that the tuning parameter for this estimator is the number of series terms m. The structure of $\mathcal{G}_{m,n}$ restricts the estimator in such a way that estimation boils down to a straightforward least-squares problem

$$\beta_m = \underset{\beta \in \mathbf{R}^m}{\arg\min}\, \mathrm{E}\left(Y_i - Z_m(X_i)'\beta\right)^2, \tag{1.6}$$

and its sample analogue

$$\hat{\beta}_m = \underset{\beta \in \mathbf{R}^m}{\arg\min}\, \mathbb{E}_n\left(Y_i - Z_m(X_i)'\beta\right)^2. \tag{1.7}$$

The approximating function mapping $\mathcal{X}$ into $\mathbf{R}$ is

$$g_m(x) = Z_m(x)'\beta_m,$$

and its plug-in estimator is

$$\hat{g}_m(x) = Z_m(x)'\hat{\beta}_m.$$

Define the approximation error as

$$r_m(x) := g(x) - g_m(x), \tag{1.8}$$

which leads to the decomposition

$$\hat{g}_m(x) - g(x) = Z_m(x)'(\hat{\beta}_m - \beta_m) - r_m(x).$$

Series estimators are especially attractive as they immediately offer estimators for linear functionals of the conditional mean function. Consider the linear operator $T : \mathcal{G} \to \mathbf{R}$ applied to the decomposition in (1.8)

$$\theta(x) := (Tg)[x] = (TZ_m)[x]'\beta_m + (Tr_m)[x],$$

which suggests using the plug-in estimator

$$\hat{\theta}_m(x) = (TZ_m)[x]'\hat{\beta}_m, \tag{1.9}$$

as an estimator for any linear functional of g. The approximation error of linear functionals will similarly to (1.8) depend on the number of series term m. In Chapters 2 and 3 this will be heavily exploited as the theory for series estimators of g carries over to plug-in estimators of the linear functional θ with minimal effort under suitable conditions. For ease of notation, write $\alpha_m(x)$ for $(TZ_m)[x]$. A non-exhaustive list of linear operators and functionals which are within the scope of this thesis are:

1. the identity operator $\theta(x) = Tg[x] = g(x)$ with

$$x \mapsto \alpha_m(x) = Z_m(x) \text{ and } x \mapsto r_{\theta,m}(x) = r_m(x);$$

2. the differential operator $\theta(x) = Tg[x] = \partial^j g(x)$ with

$$x \mapsto \alpha_m(x) = \partial^j Z_m(x) \text{ and } x \mapsto r_{\theta,m}(x) = \partial^j r_m(x);$$

3. the integral operator $\theta(x) = Tg[x] = \int g(x)\,\mathrm{d}F(x)$ with

$$x \mapsto \alpha_m(x) = \int Z_m(x)\,\mathrm{d}F(x) \text{ and } x \mapsto r_{\theta,m}(x) = \int r_m(x)\,\mathrm{d}F(x).$$

1.2.1. Examples of Basis Functions

Below, I review some of the most commonly used basis functions in the literature. For a more in depth reference see Chen (2007) or Belloni et al. (2015).

Example 1.1 (Polynomials)**.** The space of polynomials of degree $m-1$ is given by

$$Z_m(x) = (1, x, x^2, \ldots, x^{m-1})'.$$

Example 1.2 (Fourier)**.** The space of Fourier series of degree $(m-1)/2$ for an odd m is given by

$$Z_m(x) = (1, \cos(2\pi jx), \sin(2\pi jx))' \qquad \text{for } j = 1, \ldots, (m-1)/2.$$

Example 1.3 (Splines)**.** Let N be a positive integer and $t_1, \ldots, t_N$ be a real knot sequence with $t_i < t_j$ for $i < j$. The space of splines of degree p with knot sequence $\{t_i\}_{i=1}^N$ is given by

$$Z_m(x) = \left(1, x, x^2, \ldots, x^p, (x-t_1)_+^p, \ldots, (x-t_N)_+^p\right)',$$

where $m = 1 + p + N$.

Example 1.4 (B-Splines)**.** Let N be a positive integer and $t_0, \ldots, t_N$ be a real knot sequence with $t_i < t_j$ for $i < j$ and define the first-order B-Spline as

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

and higher-order B-Splines by

$$B_{i,k+1}(x) = \frac{x - t_i}{t_{i+k} - t_i} B_{i,k}(x) + \frac{x - t_{i+k+1}}{t_{i+k+1} - t_{i+1}} B_{i+1,k}(x),$$

with the convention that any $B_{i,k+1}(x) = 0$ for $x < t_0$ or $x > t_N$. This recursion is known as the Cox-de Boor recursion (De Boor 1978). The space of B-Splines of order p is given by

$$Z_m(x) = (B_{i,k}(x))' \qquad \text{for } i = 1, \ldots, N,\ k = 1, \ldots, p-1,$$

with $m = N + p$.

1.2.2. Multivariate Series Estimators

The examples above only work for univariate covariates, i.e. $d = 1$. It is easy to construct basis transformations for $d > 1$ from these univariate basis functions. This is generally referred to as a tensor product of basis functions. Let $x = (x_1, \ldots, x_d) \in \mathcal{X} \subset \mathbf{R}^d$ and let $Z_{m_i,i}(x_i)$ be a univariate basis transformation of x_i for $i = 1, \ldots, d$. The tensor product of basis functions is

$$Z_m^d := Z_{m_1,1}(x_1) \otimes \cdots \otimes Z_{m_d,d}(x_d),$$

which forms a basis system for $\mathcal{G}_{m,n}^d = \otimes_{i=1}^d \mathcal{G}_{m,n,i}^d$ with dimension

$$\dim \mathcal{G}_{m,n}^d = \prod_{i=1}^d m_i =: m.$$

The estimation problem using a tensor product of basis functions in (1.7) becomes

$$\hat{\beta}_m = \underset{\beta \in \mathbf{R}^m}{\arg\min}\, \mathbb{E}_n\left(Y_i - Z_m^d(X_i)'\beta\right)^2.$$

The subscript on m in each of the univariate basis functions indicates that the number of series terms for each one needs to be chosen. There are a total of d tuning parameters to be specified. This results in the number of parameters to be estimated to grow exponentially with the dimension of the covariates. Finally, there is no need for each of the dimensions to use the same basis functions. It is perfectly possible to, say, specify B-Splines for one covariate and polynomials for another. Tensor products make the implementation and the theory of series estimators straightforward when the dimension of the regressors is greater than one.

2 Risk-Consistency with Many Regressors After Cross-validation

2.1. Introduction

This chapter considers choosing the regressors in high-dimensional linear models by K-fold cross-validation (KFCV). I show that the cross-validated least-squares estimator predicts the outcome Y equally well as the best linear predictor (BLP) of dimension m_* under a minimal set of assumptions. Let $(Y, X) \in \mathbf{R} \times \mathbf{R}^d$ be a real-valued response and a vector of covariates such that there exists some relationship between Y and X captured by an unknown function $g \in \mathcal{G}$ up to some disturbance term ε

$$Y = g(X) + \varepsilon, \tag{2.1}$$

which is estimated by the series estimator as introduced in Chapter 2. Given a triangular array $(Y_{i,n}, X_{i,n}) \in \mathbf{R} \times \mathcal{X}$, I analyse the properties of the cross-validated estimator in two main cases of interest simultaneously:

(1) **Nonparametric model** (NP): the researcher transforms the observed independent variables X using basis functions $Z_m = Z_m(X)$ where $Z_m : \mathcal{X} \to \mathbf{R}^m$. The basis functions form an orthonormal system on $\mathcal{X} \subset \mathbf{R}^d$. The resulting estimator $\hat{\beta}_m$ yields the series estimator $Z_m(X)'\hat{\beta}_m$ which is a finite-dimensional approximation to the infinite-dimensional function $g \in \mathcal{G}$. In this case, the covariates are fixed for

fixed n, but KFCV chooses the number of series terms m to enter into the regression which are contained by the model set $\mathcal{M}_n$.

(2) **High-Dimensional Linear model** (HDL): Contrary to the NP model, the regressors are such that $Z = X$. Here, the function g is finite-dimensional and the goal is to explicitly let KFCV choose which regressors from the model set to enter into the model. The number of regressors is allowed to but does not need to diverge as $n \to \infty$, which means that even though g is finite-dimensional its dimension may change with the sample size.

By estimating the model using the series estimator, the estimation problem reduces to a least-squares problem such that for $Z_{i,m} = Z_m(X_i) = (Z_1^m(X_i), \ldots, Z_m^m(X_i))'$ and $D_i = (Y_i, Z_{i,m})_{i=1}^n$ a random sample of observations drawn from P yields the least-squares estimator

$$\hat{\beta}_m := \arg\min_{\beta \in \mathbf{R}^m} \mathbb{E}_n \left(Y_i - Z_{i,m}' \beta \right)^2. \tag{2.2}$$

In many situations the choice of regressors or the number of variables to enter into the model is unknown. Data-driven procedures can be very helpful in choosing an appropriate model, but common techniques use the same data for model selection as for estimation. It is vitally important to derive the statistical properties of these estimators by taking into account the model-selection step rather than assuming that the usual properties carry over from the case where the model is fixed. The goal in this chapter is to study the properties of the least-squares estimator when the number of regressors can grow with the sample size, but the final dimension is chosen by KFCV.

The well-known KFCV procedure in the least-squares setting is as follows. Define a sequence of models $\mathcal{M}_n = \{\underline{m}, \ldots, \overline{m}\}$, called the model set, where the researcher chooses

$\underline{m}$ and $\overline{m}$ beforehand. For any $m \in \mathcal{M}_n$ partition the data into K independent sub-samples, S_k. For each fold $k \in \{1, \ldots, K\}$ compute the least-squares estimator $\hat{\beta}_{m,-k}$ withholding the data in S_k such that the cross-validated choice of m is

$$m^{CV} = \arg\min_{m \in \mathcal{M}_n} R_K(m), \tag{2.3}$$

for

$$R_K(m) := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|S_k|} \sum_{i \in S_k} \left(Y_i - Z'_{i,m} \hat{\beta}_{m,-k} \right)^2.$$

After obtaining m^{CV}, estimate the least squares estimator in (2.2) using these regressors.

I establish the optimality of the cross-validated estimator in terms of risk consistency which is defined below. Let $D_{n+1} = (Y_{n+1}, Z_{n+1,m})$ be a new observation drawn from P and denote the predictive risk conditional on the data used for estimation by

$$R(\beta_m) := \mathrm{E}_{D_{n+1}|D_1,\ldots,D_n}[(Y_{n+1} - Z'_{n+1,m}\beta_m)^2].$$

I want to compare the cross-validated estimator to the BLP of dimension m_*, whose particular role will be discussed in more detail after Definition 2.1, by studying the excess risk between these two quantities

$$\mathcal{E}(m^{CV}, m_*) := R(\hat{\beta}_{m^{CV}}) - R(\beta^{BLP}_{m_*}), \tag{2.4}$$

where,

$$\beta^{BLP}_{m_*} := \arg\min_{\beta \in \mathbf{R}^{m_*}} \mathrm{E} \left(Y - Z'_{m_*} \beta \right)^2. \tag{2.5}$$

Using these concepts, risk consistency is then defined as follows.

Definition 2.1 (Greenshtein and Ritov 2004)**.** A sequence of estimators $\hat{\beta}_m$ is risk consistent if, for any sequence of $P \in \mathcal{P}$

$$\mathcal{E}(m, m_*) \xrightarrow{p} 0, \tag{2.6}$$

where m is not necessarily equal to m_*.

Any estimator that satisfies this definition asymptotically predicts the outcome equally well as the theoretically optimal but practically infeasible best linear predictor (BLP) using m_* regressors. This criterion is uncommon in econometrics, but there are several advantages of using risk consistency. Firstly, far fewer assumptions are maintained to establish optimality in this sense compared to the usual ℓ_2-distance between $\hat{\beta}$ and some β. For example, there is no need for exogeneity assumptions nor any assumptions on the conditional heteroskedasticity of the data. The exact assumptions maintained in this chapter are set out below. Secondly and more importantly, it allows for the comparison between the data-driven estimator $\hat{\beta}_{m^{CV}}$ and the infeasible choice $\beta^{BLP}_{m_*}$. This makes a direct comparison between quantities of different dimensions possible and straightforward. Finally, establishing the properties of estimators in terms of risk consistency is useful and meaningful in applications where $R(\beta_m) \xrightarrow{p} 0$ is unlikely, as pointed out in Greenshtein and Ritov (2004).

Therefore, I do not explicitly assume the existence of a true model. The goal is to use the information in X to predict Y, rather than to conduct inference on g. Hence, it is not assumed that g is correctly specified and thus does not necessarily have the usual interpretation of the conditional mean function. Both cases are, however, covered by the theory. For this reason, the definition of m_* is left as general as possible on purpose. However, the interpretation of m_* should depend on the application at hand. In the NP model, it is natural to consider a sequence of m_* such that it diverges at a rate

at which the researcher believes the series estimator attains the optimal rate of convergence. The leading constant in nonparametric rates of convergence is often unknown. The researcher could specify $\underline{m}$,, $\overline{m}$ and m_* to diverge at the same rate, but starting from different levels. Therefore, it would be perfectly possible that m_* is not a member of the model set for some n. In contrast, it would make more sense in the HDL model that $m_* \in \mathcal{M}_n$ with the understanding that the researcher usually has some prior knowledge of which regressors are good predictors for Y. It is then natural to study the performance of the cross-validated estimator to the BLP of some dimension $m_* \in \mathcal{M}_n$. The simulation results in Section 2.3 show how this comparison could work in practice by setting m_* to the true data-generating process in the HDL model and to the optimal L^2-rate of convergence in the NP model. The generality also offers the additional benefit that it is possible to give upper bounds on how complex the estimated model can be given the maintained assumptions. The rates of convergence derived on the excess risk, and in particular in Corollary 2.1.2, provide a useful indication of how large an m_* one can choose before KFCV breaks down. Condition 2.4 formally states how $\mathcal{M}_n$, through $\overline{m}$, relates to m_* which, given the discussion above, is left as general as possible for the theory to go through whilst leaving room for practitioners to make the comparisons they find most appropriate and relevant.

The main result establishes a non-asymptotic bound on the excess risk under minimal assumptions on the data. Using this non-asymptotic bound I derive a rate of convergence for the excess risk and establish the risk consistency of the cross-validated estimator as the sample size grows. Like Shao (1993), I show that leave-one-out cross-validation (LOOCV) is not necessarily optimal, i.e. not risk consistent, as this is equivalent to KFCV with $K = n$. In fact this holds for any cross-validation method for which $K \asymp n$. There is a trade-off between how fast K and m can grow with the sample size. This is the first such

result in a high-dimensional linear setting. To truly show that LOOCV is sub-optimal, one would need to establish a lower bound on the risk consistency which is not within the scope of the theory presented here.

Remark 2.1. The results presented in the chapter continue to hold for model choices $\check{m}$ and some non-random sequence $\tau_n > 0$ such that

$$R_K(\check{m}) \leq R_K(m^{CV}) + \tau_n.$$

This would come at the cost of carrying the term τ_n around in Theorem 2.1 and the proofs with the assumption that $\tau_n \to 0$ as $n \to \infty$. This is especially relevant when the cross-validation objective function $R_K(m)$ is flat in a neighbourhood around m^{CV}. In such cases, researchers may wish to select the most parsimonious model within a certain tolerance level around $R_K(m)$.

2.1.1. Related literature

This chapter mainly contributes to the literature on cross-validation which has been extensively studied in statistic and econometrics. Much of the work, especially on LOOCV, dates back to the 1970s starting with Allen (1974), Stone (1974), Geisser (1975) and Wahba and Wold (1975). In a seminal paper, Li (1987) established the optimality of LOOCV in terms of asymptotic efficiency, i.e.

$$\frac{L(\hat{\beta}_{m^{CV}})}{\inf_{m \in \mathcal{M}_n} L(\hat{\beta}_m)} \xrightarrow{p} 1, \tag{2.7}$$

where $L(\beta_m)$ is the L_2-loss with m regressors. Andrews (1991b) extended these results to the heteroskedastic case. Hansen (2014), in turn, extended their work to the non-parametric setting where cross-validation is used to select the number of series terms to

enter into the regression. However, his results are limited to LOOCV and to nested models which are truly infinite dimensional. Neither of which is assumed here. The question answered in this chapter is fundamentally different from the ones in the cited literature as I do not establish that KFCV can approach the optimal $m \in \mathcal{M}_n$ in mean-squared error sense, but compare how well cross-validation does compared to the BLP whose dimension is a sequence m_* which need not be a member of the model set $\mathcal{M}_n$.

Several optimality criteria exist and the various flavours of cross-validation may be optimal according to one and sub-optimal according to another. Shao (1993) showed that LOOCV is inconsistent if a true finite-dimensional model exists, but leave-K-out cross-validation is consistent which is in line with this chapter but for another optimality criterion. Lastly, Arlot and Celisse (2010) contains and extensive review of the theoretical results on various flavours of cross-validation combined with practical guidelines for researchers using these techniques in applied work.

This chapter also contributes to the literature where risk consistency is used as the (main) property of interest. Despite it being uncommon in econometrics, it has been used quite successfully in the statistical literature. In particular, it is popular in establishing optimality in penalised estimators. For example, Greenshtein and Ritov (2004) for the Lasso under L_2 loss and van de Geer (2008) for the Lasso under Lipschitz loss functions. Hsu, Kakade and Zhang (2014) establish the risk consistency of the least-squares and ridge regression estimators.

These authors established the risk consistency property under weak assumptions, but the main drawback is that they do not take into account that the tuning parameters are usually chosen by some data-driven procedure. Homrighausen and McDonald (2013) filled in this gap by proving the risk consistency of the Lasso where the penalty is chosen

by KFCV. Since the use of these regularisation estimators in practice usually includes a model-selection step it is important to establish their properties rather than ignoring the model-selection step as is all too common.

2.2. Risk Consistency

Before presenting the main results – the non-asymptotic bound on the excess risk and the risk consistency of $\hat{\beta}_{\hat{m}^{CV}}$ – I list the assumptions which are maintained to deliver these results.

Condition 2.1 (Nested models)**.** *For the set of regressors $\mathcal{M}_m$ with dimension m, we have that $m < m'$ implies that $\mathcal{M}_m \subset \mathcal{M}_{m'}$.*

Condition 2.1 simply states that the models are nested, which means that the regressors are nested for increasing values of m. In the HDL model this literally means that the set of regressors is nested, but in the NP model this states that the series expansions are nested after the regressors are transformed by some basis transformation. For example, this is the case when the transformations are splines of a fixed degree, but the knot sequences are nested as m increases.

Condition 2.2 (Sample)**.** *The data $D_i = (Y_i, Z_{i,m})_{i=1}^n$ is a random sample from $P \in \mathcal{P}$.*

Condition 2.3 (Distribution)**.** *Let $\mathcal{P}$ be the set of distributions P for which the following holds: (i)* $\mathrm{E}\left[\max_{1\leq i\leq n}|Y_i|^2\right] + \mathrm{E}\left[\max_{1\leq i\leq n}\|Z_{i,m_*}\|^2\right] =: \xi_{m_*}^2 < \infty$*; (ii) for some $C < \infty$,* $\max_{j\leq \overline{m}\vee m_*} \mathrm{E}\,|Z_{ij}Y_i|^3 \leq C$*; (iii)* $\lambda_{\min}(\Sigma_n) \geq c > 0$ *for* $\Sigma_n := \mathrm{E}[D_i D_i']$.

The second set of assumptions, Conditions 2.2 and 2.3, summarises the characteristics of the distributions over which the results hold uniformly. These impose mild conditions

on the moments of the data. In particular, Condition 2.3(i) does not require the support of the data to be bounded almost surely which is a common simplifying assumption in the literature. Of course, all results will be valid if this is indeed the case. Condition 2.3(ii) states that the covariance matrix of the regressors of dimension m is invertible, but under weak conditions on the density of the data this can always be achieved (Belloni et al. 2015). The design is random, but without loss of generality $\Sigma_n = I_n$.

Condition 2.4 (Dimensions)**.** *The relationship between $\overline{m}$ and m_* is such that $\overline{m} \asymp m_*$.*

Condition 2.4 is a high-level assumption laying out the relationship between $\overline{m}$ and m_*. It basically states that $\overline{m}$ may not be 'too far' from m_* and incorporates the two situations of interest: $\overline{m}$ and m_* are fixed or $\overline{m}$ and m_* diverge with the sample size. There should ideally be a way to choose $\overline{m}$ such that it adapts to the rate in Corollary 2.1.1. However, such a method should not introduce extra nuisance parameters as this would defeat the point. How to choose $\overline{m}$ is still very much an open question and beyond the scope of the research presented here.

The quantity $\mathcal{E}(m^{CV}, m_*)$ is a random variable since KFCV is a data-driven procedure which means that m^{CV} and m_* may differ. It is possible for the excess risk to be negative, but Lemma 2.1 below provides some high-level conditions under which this event is unlikely to happen, at least asymptotically.

Lemma 2.1. *Suppose that Condition 2.1 holds and that $\overline{m} = o(m_*)$. Then,*

$$\mathcal{E}(m^{CV}, m_*) \geq 0, \tag{2.8}$$

with probability at least $1 - o(1)$. The inequality in (2.8) holds almost surely if $\overline{m} = m_*$.

Let me introduce the quantities below which simplify the exposition of the main results

and the proofs:

$$\sqrt{A} := \sqrt{m_*} + \sqrt{\xi_{m_*}^2 \log m_*/n}$$
$$\epsilon_1 := c + \sqrt{m_*} + 2\sqrt{K}$$
$$\epsilon_2 := (c' + 4\sqrt{K})A,$$

for constants $c = 3 + 2\sqrt{2}$ and $c' = c + \sqrt{2}$.

Theorem 2.1. *Assume that Conditions 2.2 and 2.3 and Condition 2.4 hold and that* $n \geq C\xi_{m_*}^2 \log m_*$ *for* $C > 0$ *sufficiently large. Then, uniformly in* $P \in \mathcal{P}$,

$$\mathcal{E}(m^{CV}, m_*) \lesssim_P (\epsilon_1 + \epsilon_2)\sqrt{\frac{\xi_{m_*}^2 \log m_*}{n}}. \tag{2.9}$$

The non-asymptotic bound in Theorem 2.1 shows that under minimal assumptions the excess risk is bounded for finite n with high probability. Theorem 2.1 delivers a rate of convergence on the excess risk under the following growth condition.

Condition 2.5 (Growth rates)**.** $K\xi_{m_*}^2 m_*^2 \log m_* = o(n)$.

Condition 2.5 rules out a meaningful upper bound for LOOCV which hints at its sub-optimality similar to the results in Shao (1993). However, it does provide a justification for the usual wisdom that 3 or 5-fold cross-validation works well in prediction-type problems. Theorem 2.1 combined with Condition 2.5, a mild growth condition on the dimension of the infeasible BLP, shows that the cross-validated estimator is risk consistent according to Definition 2.1. It thus shows that the cross-validated estimator can predict the outcome equally well asymptotically as the infeasible BLP.

Corollary 2.1.1. *Assume the conditions in Theorem 2.1 and Condition 2.5, then*

$$\mathcal{E}(m^{CV}, m_*) = O_P\left(\sqrt{K\xi_{m_*}^2 m_*^2 \log m_*/n}\right) = o(1).$$

Corollary 2.1.1 only uses the high-level conditions as stated above, but it is straightforward to apply it to obtain a rate of convergence on the excess risk for the HDL and NP model.

Corollary 2.1.2. *Assume that the conditions in Corollary 2.1.1 hold. Furthermore, assume that* $\mathrm{E}[|Y_i|^s] \lesssim 1$ *for some* $s > 2$ *and that* $\|X_i\| \lesssim 1$ *for any* n*. In the HDL model with bounded regressors, i.e.* $Z = X$*, it follows that*

$$\mathcal{E}(m^{CV}, m_*) = O_P\left(\sqrt{Km_*^2 \log m_* / n^{(s-2)/s}}\right) = o(1).$$

Furthermore, in the NP model

$$\mathcal{E}(m^{CV}, m_*) = O_P\left(\left(n^{1/s} \vee \sqrt{m_*}\right)\sqrt{Km_*^2 \log m_* / n}\right) = o(1),$$

if the basis transformations $Z : \mathcal{X} \to \mathbf{R}^m : x \mapsto Z(x)$ *are tensor products of B-Splines on* $\mathbf{R}^d$ *or*

$$\mathcal{E}(m^{CV}, m_*) = O_P\left((n^{1/s} \vee m_*)\sqrt{Km_*^3 \log m_* / n}\right) = o(1),$$

if the basis transformations Z *are tensor products of polynomials on* $\mathbf{R}^d$*.*

For polynomials and B-Splines, respectively, it holds that $\|Z_m\| \lesssim m$ and $\|Z_m\| \lesssim \sqrt{m}$ for any m. Hence, it only remains to combine a bound on ξ_m in Condition 2.3 with Corollary 2.1.1 to derive the rate of convergence of the excess risk. The bound on ξ_m will inevitably depend on the tails of Y and Z. The tails of the response will differ from application to application, but bounds on $\|Z_m\|$ are readily available for many basis transformations, see e.g. DeVore and Lorentz (1993).

2.3. Monte Carlo Experiment

In this section, I conduct an extensive Monte Carlo experiment to study the finite-sample properties of the excess risk $\mathcal{E}(m^{CV}, m_*)$. Two main designs are considered: one relating to the HDL model and one relating to the NP model. For each of the designs the experiment is replicated 100,000 times for the sample sizes 50, 100, 250, 500 and 1000. Each experiment uses $K = 5$ which relates to 5-fold cross-validation (5FKV). In addition to studying the behaviour of KFCV, I include the results from LOOCV and the Akaike Information Criterion (AIC) for comparison.

Design 1 is inspired by the Wild Bootstrap (Wu 1986; Liu 1988). The linear model is

$$Y_{i,n} = X'_{i,n}\beta_n + \varepsilon_n, \tag{2.10}$$

where $(Y_{i,n}, X_{i,n})$ is an observed response and a vector of regressors. The data-generating process used in the simulation study is

$$Y^*_{i,n} = X'_{i,n}\hat{\beta}_n + u_{i,n},$$

where $\hat{\beta}$ is the estimator for β in (2.10) and $u_{i,n} \sim \mathcal{N}(0, 1)$. The data used to calibrate the model is a real-world dataset from Riphahn, Wambach and Million (2003) on German health care.[1] This dataset collects data from a panel study, but I fix the year to 1988 to obtain a cross-sectional dataset. The independent variable is the log of household income and 'true' model is chosen from the available regressors. Note that the subscript n in (2.10) implies that the model changes for each $n = 50, 100, 250, 500, 1000$. Specifically, the

[1] This dataset can be downloaded from https://dx.doi.org/10.1002/jae.680 together with a 'readme' document fully explaining the dataset.

dimension of the regressors used to estimate β_n diverges as $n \to \infty$. This yields a high-dimensional fixed-regressors design with Gaussian noise. The comparisons are made versus m_* which is set to the true model generating the data. See Appendix A.2 for the specification of the true models and the model set from which the three model-selection techniques can choose a model for estimation.

The model generating the data in the NP case (Design 2) is:

$$Y = \log(|6x - 3|)\,\mathrm{sgn}(x - 0.5) + \varepsilon.$$

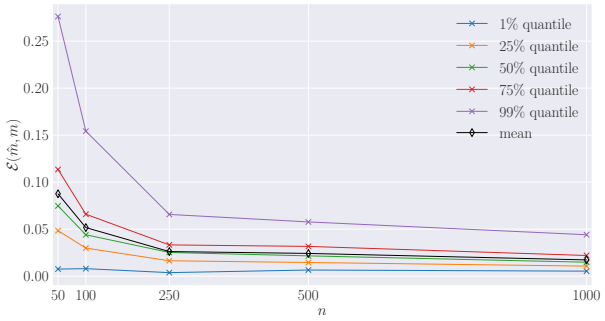
The random variable X is a uniform random variable on $(-3, 3)$ and ε is a centred Gaussian random variable. This function is a rescaled version of the one used in Newey and Powell (2003) and Chen and Christensen (2015a).

There are three sub-cases: a low and high-variance homoskedastic case (Design 2.1 and 2.2) and a heteroskedastic case (Design 2.3). In Design 2.1, $\sigma^2 = \sqrt{0.5}$, in Design 2.2 $\sigma^2 = 2\sqrt{2}$ and in the heteroskedastic Design 2.3 $\sigma^2 = \sqrt{0.5X}$. Note that in the final case, the error variance is not bounded away from zero. The unknown conditional mean function g is approximated using cubic splines with a sequential knot sequence $\mathcal{S}_n$ on $(-3, 3)$ where the knots are multiple of 0.25:

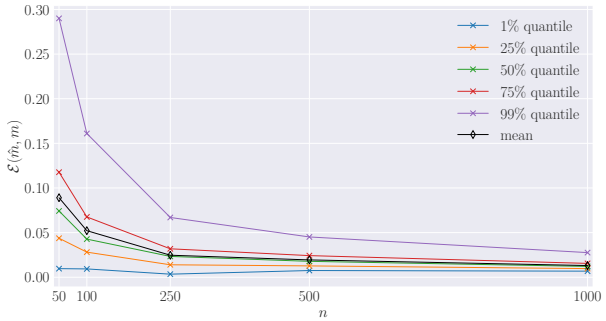
$$\mathcal{S}_n^k = \bigcup_{j=0}^{k} \left\{ \pm \frac{j}{4} \right\}.$$

For Designs 2.1-2.3, the knot sequences are set with $k = 2, 3, 5, 7, 9$ for the oracle and $k_{\max} = 4, 6, 10, 14, 18$ for $n = 50, 100, 250, 500, 1000$. The latter is also the number of models since the lower bound for the model set is zero for each n. This translates to the following number of parameters $m_* = 9, 12, 15, 19, 23$ and $\overline{m} = 13, 17, 25, 33, 41$. The sequence m_* ensures that the infeasible $\hat{g}_{m_*} = Z'_{m_*}\hat{\beta}_{m_*}$ attains the optimal L^2-rate, see Belloni et al. (2015).

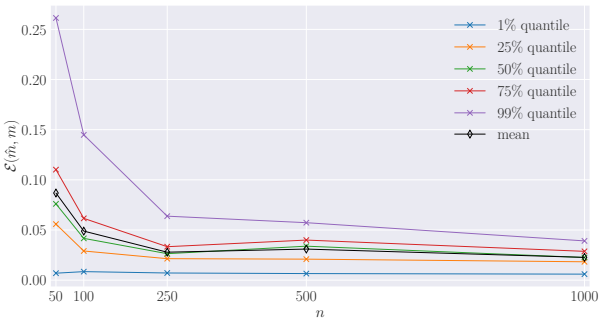
Figures 2.1 to 2.4 summarise the results. They accompany the theoretical results nicely by showing that 5FCV works well for predicting the unknown response, but that LOOCV does not perform as consistently as hinted at by Theorem 2.1. This is especially apparent in Figure 2.2. Surprisingly, the AIC works well in all cases despite the fact that Andrews (1991b) showed that optimality breaks down in heteroskedastic designs albeit for a different optimality criterion. This suggests that the AIC deserves more attention even in heteroskedastic designs. More results showing the deciles of $\mathcal{E}(m^{CV}, m_*)$ for each of the sample sizes are collected in Appendix A.2.



(a) 5FCV

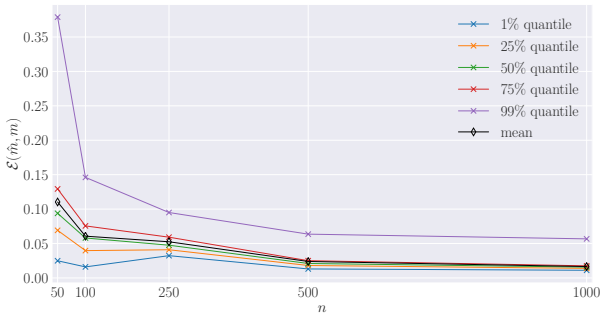


(b) AIC

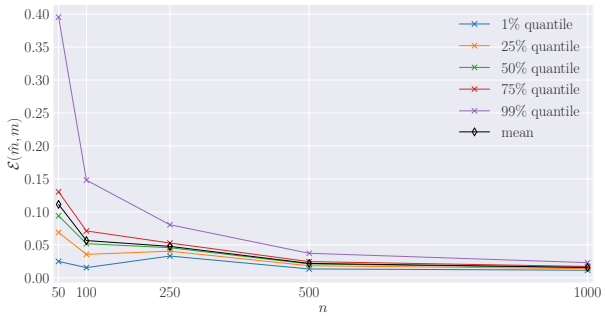


(c) LOOCV

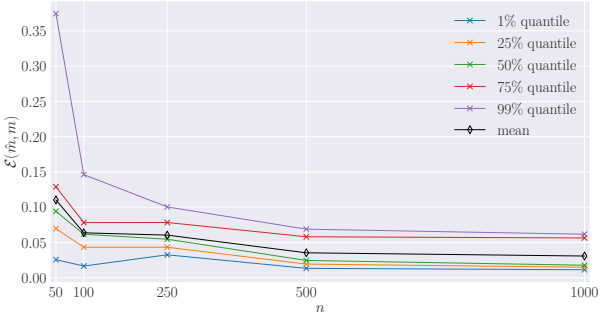
Figure 2.1: Quantiles of $\mathcal{E}(m^{CV}, m_*)$ in Design 1.



(a) 5FCV

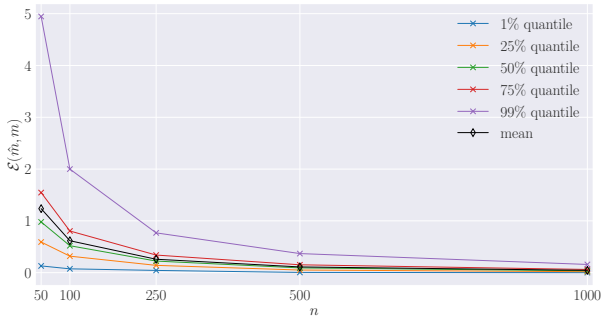


(b) AIC

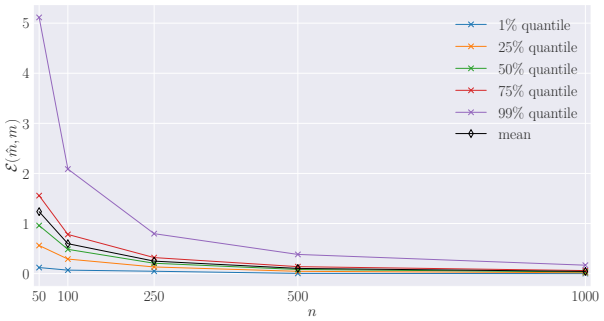


(c) LOOCV

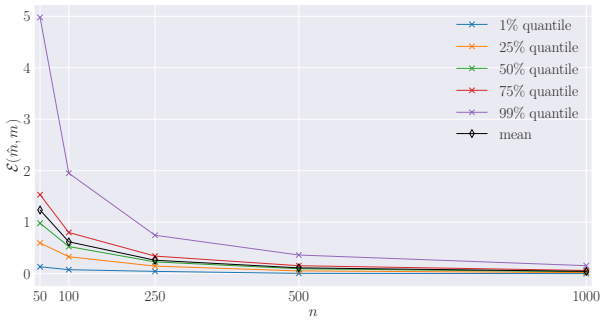
Figure 2.2: Quantiles of $\mathcal{E}(m^{CV}, m_*)$ in Design 2.1.



(a) 5FCV

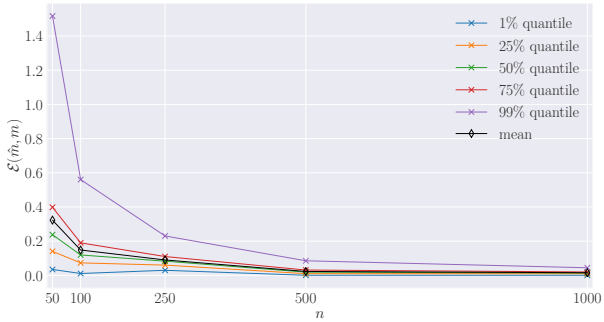


(b) AIC

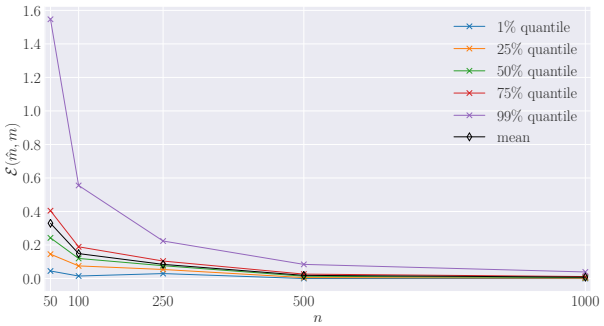


(c) LOOCV

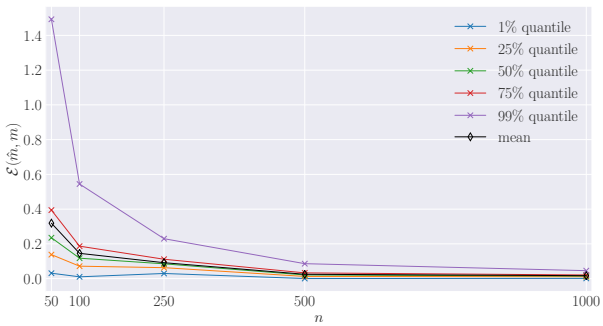
Figure 2.3: Quantiles of $\mathcal{E}(m^{CV}, m_*)$ in Design 2.2.



(a) 5FCV



(b) AIC



(c) LOOCV

Figure 2.4: Quantiles of $\mathcal{E}(m^{CV}, m_*)$ in Design 2.3.

3 Uniform Convergence of Series Estimators and its Linear Functionals

3.1. Introduction

Chapter 2 showed that the cross-validated series estimator predicts the outcome Y well in high-dimensional linear or nonparametric models. In this chapter, I return to a more classical optimality criterion and turn the attention to the conditional mean function $g \in \mathcal{G}$ and its linear functionals $\theta : \mathcal{G} \to \mathbf{R}$ in the nonparametric regression model in (1.1). The main result in this chapter is the uniform consistency with a rate of convergence of the series estimator where the uniformity holds both over the support of the covariates $\mathcal{X}$ as well as the model set $\mathcal{M}_n$, i.e. it holds for any $(x, m) \in \mathcal{X} \times \mathcal{M}_n =: \mathcal{I}_n$. For functions in a Hölder ball, the series estimator attains the minimax rate of Stone (1982) up to a small logarithmic factor due to the uniformity over the model set $\mathcal{M}_n$.

A non-asymptotic bound together with an undersmoothing condition on the approximation error is central in establishing the uniform consistency result. These results are of independent interest, but more importantly they lay the foundation for the theory presented in Chapter 4. The non-asymptotic bound requires deriving properties on high-dimensional covariance matrices. Therefore, this chapter concludes by showing that the heteroskedasticity-consistent standard errors for $\hat{\theta}_m$ are uniformly consistent in $\mathcal{M}_n$.

3.1.1. Related Literature

The literature on the consistency and asymptotic normality of sieve and series estimators is mature. Classical results can be found in Andrews (1991a), Stone (1994), Newey (1997), Huang (2003) and an excellent review chapter in Chen (2007) summarising results which are still relevant today. More recent results are in Chen and Christensen (2015b), Belloni et al. (2015) and Hansen (2015). Belloni et al. (2015) were the first authors to study the consistency and inference uniformly in the support of the regressors and extended the theory of series estimators in several directions. However, their results only hold uniformly in $\mathcal{X}$.

As pointed out in Chapter 2, there is also an extensive literature on the optimality of data-driven series estimators, see Li (1987), Andrews (1991b) and Hansen (2014). However, many of these concentrate on proving that the chosen m delivers a model which minimises some variant of the mean-squared error, yet inference on the 'true' g or θ is often the end goal in econometrics. These papers fail to address the issue that data-driven procedures also deliver uniformly consistent estimators, let alone valid inference. An important exception to these works is Chetverikov, Liao and Chernozhukov (2019) who derive the rate of convergence of the cross-validated Lasso estimator when $m \gg n$. Their theory could be adapted to the nonparametric case, but this is not a trivial exercise. The set-up here circumvents the problem of handling each model-selection procedure on a case-by-case basis. Due to the uniformity over $\mathcal{M}_n$, it holds that the data-driven series estimator is consistent regardless of the algorithm used in choosing the number of series terms.

The uniform convergence results presented in this chapter fill a gap in the series literature which has long been filled in the kernel literature. Einmahl and Mason (2005) proved

the uniform-in-bandwidth consistency of kernel estimators in density estimation and in nonparametric regression. The techniques used in their paper are unfortunately not transferable due to the fact that the choice of the smoothing parameter is inherently discrete in series estimation. Hence, this research complements the kernel literature and tries to bridge the gaps in the results already proved in that area.

3.2. Model Framework

The model is as in (1.1)

$$Y = g(X) + \varepsilon, \tag{3.1}$$

where $g : \mathcal{X} \to \mathbf{R}$ is the unknown conditional mean function. With a sample (Y_i, X_i) drawn from $(Y, X) \sim P$, the model can be decomposed into three components

$$Y_i = Z'_{i,m}\beta_m + r_{i,m} + \varepsilon_i,$$

for $Z_{i,m} = Z_m(X_i) \in \mathbf{R}$ and $r_{i,m} = r_m(X_i)$ and any $m \in \mathcal{M}_n$. The quantities Y_i, X_i and g may all change with n, but this notation is suppressed for simplicity. The model set

$$\mathcal{M}_n := \left\{ m \in \mathbf{N} : m \in [\underline{m}, \overline{m}] \right\}, \tag{3.2}$$

collects the models under consideration and contains the number of series terms to appear in the expansion (1.5). The vector β_m is the best linear predictor which solves (1.6) such that $Z'_{i,m}\beta_m$ is the BLP of g ignoring the deterministic bias r_m. Thus, for any $m \in \mathcal{M}_n$

$$\hat{g}_m - g = Z'_m(\hat{\beta}_m - \beta_m) - r_m,$$

where $\hat{\beta}_m$ is the least-squares estimator solving (1.7), which motivates using $\hat{g}_m = Z_{i,m}'\hat{\beta}_m$ as an approximation for g. This decomposition conveniently carries over to linear functionals using the notation from Section 1.2

$$\hat{\theta}_m - \theta = \alpha_m'(\hat{\beta}_m - \beta_m) - r_{\theta,m}.$$

The approximation error, $r_{\theta,m}$ depends on the functional θ as well as the order of the approximation m. The dependence on θ will be assumed and will not be made explicit in the notation. This allows me to handle the case for the conditional mean function and linear functionals simultaneously without the need for duplicate theories for both cases. Linear functionals will be of particular interest in Section 4.3.2 where I use the theory derived in this chapter to derive uniform inference methods to test the monotonicity of the demand function for US gasoline.

Remark 3.1. The popular partially linear model (Blundell, Chen and Kristensen 2007; Cattaneo, Jansson and Newey 2018a,b) is fully supported in the framework as described above. Writing this model with $X = (X_1, X_2) \in \mathcal{X}_1 \times \mathcal{X}_2$ as

$$Y = g(X_1) + X_2'\gamma + \varepsilon.$$

Then, define $Z_m(x_1, x_2) = (Z_{1,m}(x_1)', x_2')'$ where the basis transformations are $Z_{1,m} : \mathcal{X}_1 \to \mathbf{R}^m$ and $\beta_m = (\beta_{1,m}', \gamma')'$. The theory continues to hold if the assumptions described below are adapted appropriately.

Within this set-up, various quantities are left to the researcher to specify. The first choice is which basis functions to use to approximate g and θ well. This will ultimately depend on the nature of the observed data and the application at hand. The theory is presented to accommodate frequently used basis functions or tensor products thereof, see Sections 1.2.1 and 1.2.2. The most important quantity is the model set $\mathcal{M}_n$ as defined in

(3.2). At an abstract level, the researcher needs to specify the minimum $\underline{m}$ and maximum $\overline{m}$ number of series terms to appear in the approximation of the conditional mean function. Gaps in $\mathcal{M}_n$ are allowed, but the size of the model set $\tilde{m} := |\mathcal{M}_n|$ will no longer be $\overline{m} - \underline{m} + 1$ in that case. This will be assumed implicitly, but it is not a material assumption, see e.g. Corollary 3.3.1.

The uniformity results depend on the tail behaviour of the additive noise ε. The quantity $\nu_n := \sqrt{\mathrm{E}[\max_{1 \le i \le n} \varepsilon_i^2]}$ will, therefore, feature frequently. In the description of the assumptions below, fix $m \in \mathcal{M}_n$ to avoid repeatedly having to use 'uniformly over $m \in \mathcal{M}_n$'.

Condition 3.1. *(Data)* $(X_i, Y_i)_{i=1}^n$ *is an i.i.d sample from* P *and satisfies* (1.1)*. Furthermore, the support of* X*,* $\mathcal{X}$*, is a bounded set in* $\mathbf{R}^d$ *for any* n*.*

Condition 3.2. *(Sieve space) The basis transformations are such that* $\Sigma_m = \mathrm{E}[Z_{i,m} Z_{i,m}']$ *is a positive definite matrix with* $\lambda_{\max}(\Sigma_m) \lesssim 1$*.*

Condition 3.3. *(Loadings) The loadings* α_m *in the approximation to the linear functional* θ *are such that* $\xi_m := \sup_{x \in \mathcal{X}} \|\alpha_m(x)\| \le C_n$*,* $0 < c \le \xi_m$ *and* $\log \xi_m \lesssim \log m$*.*

Condition 3.1 is a standard assumption. The support $\mathcal{X}$ is bounded, but its diameter is allowed to depend on the sample size n. Condition 3.2 implies without loss of generality that $\Sigma_m = I_m$. It is always possible under this assumption to rotate the basis transformations such that they are uncorrelated and have unit variance, see Proposition 2.1 in Belloni et al. (2015). Hence, the design is random with an unknown covariance matrix. The growth condition on ξ_m depends on the number of regressors m and is generally increasing in m. Therefore, the results are stated in terms of its upper bound $\xi_{\overline{m}}$ as $\xi_m \le \xi_{\overline{m}}$ for any $m \in \mathcal{M}_n$ if the basis transformation remain the same across the different m. If this is not the case then $\xi_{\overline{m}}$ can be replaced by $\xi_{\tilde{m}} := \max_{m \in \mathcal{M}_n} \xi_m$. The bounded-support

condition and ξ_m can be relaxed to assumption on the moments of X at the cost of greater technicalities. See Chapter 2 or Hansen (2015) who replaces this assumption by showing that the growth rate of m is invariably linked to the number of moments of X.

Condition 3.3 is a regularity condition stating which linear functionals are allowed. This makes it possible to normalise α_m to lie in the unit sphere $\mathbb{S}^{m-1}$ without loss of generality. Of course, with $\alpha_m(x) = Z_m(x)$ these bounds immediately follow from the properties of the actual basis functions and are readily available in many situations as discussed in Example 3.1.

Condition 3.4. *(Noise) The errors satisfy: (i)* $\mathrm{E}[\varepsilon_i|X] = 0$*; (ii)* $c < \mathrm{E}[\varepsilon_i^2|X]$ *a.s. for some* $c > 0$*; (iii)* $\mathrm{E}[\varepsilon_i^s|X] \lesssim 1$ *a.s. for some* $s > 2$*.*

Condition 3.5. *(Bias) The approximation errors satisfy* $\|r_{i,m}\|_\infty \leq b_m \leq b_{\tilde{m}}$ *where* $b_{\tilde{m}} := \max_{m \in \mathcal{M}_n} b_m < \infty$*.*

Condition 3.6. *(Growth rates) Let the following growth rates be: (i)* $b_{\tilde{m}} = o(1)$*; (ii)* $(v_n \vee b_{\tilde{m}})$ $\xi_m \tilde{m}^2 = o(\sqrt{n})$*; (iii)* $(v_n \vee b_{\tilde{m}}) \xi_{\underline{m}}^2 \tilde{m} \log \overline{m} \log \tilde{m} = o(\sqrt{n})$*.*

Condition 3.4 is a mild condition stating that the errors are exogenous with conditional variance bounded from above and below. However, I do not need to maintain that they are homoskedastic. The estimators of the variance of $\hat{\theta}_m$ in Section 3.3.1 is the well-known heteroskedasticity-consistent estimator from the parametric setting. This implies that a heteroskedastic design is assumed by default as this estimator is valid under heteroskedasticity as well as the lack thereof. Condition 3.5 relates to the bias term r_m which is deterministic but does depend on the chosen model m. The results are most useful if $b_{\tilde{m}} \to 0$ which surely happens when $\underline{m} \to \infty$ and Condition 3.1 is satisfied.

It is helpful to keep the following example in mind for the results on uniform conver-

gence. This example also aids in making the abstract theorems more concrete and practical as evidenced in Theorem 3.3 and Corollary 3.3.1.

Example 3.1. Recall the definition of a Hölder ball in $\mathcal{X}$, $H^p \equiv H^p(\mathcal{X})$. Let $C_u(\mathcal{X})$ be the space of uniformly continuously differentiable functions on $\mathcal{X}$ then define the Hölder ball as

$$H^p := \left\{ f \in C_u(\mathcal{X}) : \left\| f^{([p])} \right\|_\infty + \sup_{x \neq \check{x}, x, \check{x} \in \mathcal{X}} \frac{\left| D^{[p]} f(x) - D^{[p]} f(\check{x}) \right|}{\|x - \check{x}\|^{p-[p]}} \leq c \right\}.$$

for $p > 0$ and $[p]$ the integer part of p for some radius $0 < c < \infty$. Suppose that the linear functional θ lies in H^p, that $X \in \mathcal{X} \subset \mathbf{R}^d$ and that $Z_m : \mathbf{R}^d \mapsto \mathbf{R}^m$ are tensor product B-Splines of order p_0. Standard approximation theory then states that

$$\xi_{\overline{m}} \lesssim \sqrt{\overline{m}} \text{ and } b_{\tilde{m}} \lesssim \underline{m}^{-(p \wedge p_0)/d}.$$

Approximation results for many linear and non-linear sieve spaces are widely available and do not need to be derived specially for bounding b_m or $b_{\tilde{m}}$, see e.g. DeVore and Lorentz (1993), Huang (2003) and Chen (2007).

The example above already highlights that the rates will depend on how $\underline{m}$ and $\overline{m}$ grow with n and on the unknown smoothness of the conditional mean function. It is worth re-iterating that the techniques in this chapter do not adapt to the unknown smoothness of θ.

3.3. Uniform Linearisation and Convergence

The uniform linearisation in Theorem 3.1 establishes the connection between $\theta_m(x) - \theta(x)$, $Z_{i,m}$, ε_i and $r_{i,m}$. It is a non-asymptotic result and thus holds for some n large enough

with high probability. This non-asymptotic bound on $\theta_m(x) - \theta(x)$ uniformly in $(x, m) \in \mathcal{I}_n$ is the foundation for the main result in this chapter: the uniform convergence in $(x, m) \in \mathcal{I}_n$ under high-level conditions.

Theorem 3.1. *Under Conditions 3.1 to 3.5 it follows that*

$$\sqrt{n}\alpha_m(x)'(\hat{\beta}_m - \beta_m) = \alpha_m(x)'\mathbb{G}_n Z_{i,m}\varepsilon_i + R_{1,n}(\alpha_m(x)) + R_{2,n}(\alpha_m(x)),$$

with

$$R_{1,n}(\alpha_m(x)) \lesssim_P (\nu_n + b_{\tilde{m}})\sqrt{\frac{\xi^2_{\overline{m}}(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}} =: \bar{r}_{1n},$$

summarising the impact of estimation, and

$$R_{2,n}(\alpha_m(x)) \lesssim_P b_{\tilde{m}}\sqrt{\overline{m}\log\overline{m}\log\tilde{m}} =: \bar{r}_{2n},$$

summarising the impact of the approximation error, uniformly over $(x, m) \in \mathcal{I}_n$.

Remark 3.2. The $\overline{m}\log\overline{m}$ terms in the bounds of Theorem 3.1 can be improved to $\log\overline{m}$ by maintaining an extra assumption on $\alpha(x)$, namely a Lipschitz condition

$$\|\alpha(x) - \alpha(\check{x})\| \leq L\|x - \check{x}\|,$$

where L potentially depends on n such that it may diverge as $n \to \infty$. This relaxes the bounds on the covering numbers used in the proofs, see e.g. Definition D.1 and Lemma B.6. The reason for this is that the covering numbers under a Lipschitz condition depend on d, the dimension of X which is fixed, rather than the dimension of Z_m. See Chernozhukov, Chetverikov and Kato (2014b), Belloni et al. (2015) and Belloni, Chernozhukov and Fernandez-Val (2019) where this condition has been successfully used to improve the rates. However, I prefer to present a slightly worse rate which circumvents an extra condition that would have to be verified in practice.

The uniform linearisation result in Theorem 3.1 is a key result in order to obtain a non-asymptotic bound on the L^∞-norm of the processes $\alpha_m(x)'(\hat{\beta}_m - \beta_m)$ and $\hat{\theta}_m(x) - \theta(x)$ where the latter quantity takes into account the bias term.

Theorem 3.2. *Assume that the conditions from Theorem 3.1 hold. Then,*

$$\sup_{(x,m)\in\mathcal{I}_n} \left|\alpha_m(x)'(\hat{\beta}_m - \beta_m)\right| \lesssim_P \frac{\xi_{\overline{m}}}{\sqrt{n}}\left(\sqrt{\overline{m}\log\overline{m}\log\tilde{m}} + \overline{r}_{1n} + \overline{r}_{2n}\right), \tag{3.3}$$

and

$$\sup_{(x,m)\in\mathcal{I}_n} \left|\hat{\theta}_m(x) - \theta(x)\right| \lesssim_P \frac{\xi_{\overline{m}}}{\sqrt{n}}\left(\sqrt{\overline{m}\log\overline{m}\log\tilde{m}} + \overline{r}_{1n} + \overline{r}_{2n}\right) + b_{\tilde{m}}. \tag{3.4}$$

Theorem 3.2 is the second main result and immediately provides a non-asymptotic bound in the sup-norm on the estimator of the conditional mean function as well as its linear functionals. This abstract result can readily be used to derive rates of convergence for these quantities of interest where the rate will ultimately depend on the application at hand. Theorem 3.2 is the first one to consider uniform convergence in both the support of the regressors as well as the number of series terms. Below, I show how to derive a more familiar rate of convergence for linear functionals which lie in a Hölder ball with finite radius as in Example 3.1.

Theorem 3.3. *Let Conditions 3.1 to 3.5 hold. In addition assume that $\mathcal{X} \subset \mathbf{R}^d$, $\theta \in H^p$ and $(\nu_n + b_{\tilde{m}})\tilde{m}\xi_{\overline{m}} \lesssim \sqrt{n}$. If the vector of approximating functions Z consists of a tensor product of polynomials of order p, then*

$$\sup_{(x,m)\in\mathcal{I}_n} \left|\hat{\theta}_m(x) - \theta(x)\right| \lesssim_P \sqrt{\frac{\overline{m}^3\log\overline{m}\log\tilde{m}}{n}} + \underline{m}^{1-p/d}. \tag{3.5}$$

Also, if the vector of approximating functions Z consists of a tensor product of B-Splines of order p_0, then

$$\sup_{(x,m)\in\mathcal{I}_n} \left|\hat{\theta}_m(x) - \theta(x)\right| \lesssim_P \sqrt{\frac{\overline{m}^2\log\overline{m}\log\tilde{m}}{n}} + \underline{m}^{-(p_0\wedge p)/d}. \tag{3.6}$$

The conditions used to derive the rate in Theorem 3.3 are completely in line with the results in Belloni et al. (2015) up to the $\tilde{m}$ term, but the rate itself is slower due to the uniformity over $\mathcal{M}_n$.

Corollary 3.3.1. *In the setting of Theorem 3.3 with the vector of approximating functions a tensor product of B-Splines of order $p_0 \geq p$ assume the Lipschitz condition on the loadings α_m as in Remark 3.2 with $L \lesssim 1$. Then, choosing for $c < C$*

$$\underline{m} = c\left(\frac{\log n}{n}\right)^{-d/(2p+d)} \quad \text{and} \quad \overline{m} = C\left(\frac{\log n}{n}\right)^{-d/(2p+d)},$$

it follows that

$$\sup_{(x,m)\in\mathcal{I}_n} \left|\hat{\theta}_m(x) - \theta(x)\right| = O_P\left(\sqrt{\log n - \log\log n}\left(\frac{\log n}{n}\right)^{p/(2p+d)}\right) = o(1), \tag{3.7}$$

if $p/d > 1/2$ and $s > 2 + \frac{2}{p/d-0.5}$.

The $\sqrt{\log n - \log\log n}$ term is the price to pay for uniformity over the model set. Up to this term, the rate derived in Corollary 3.3.1 is the optimal rate of Stone (1982) in this problem and matches the one derived in Belloni et al. (2015). The Lipschitz condition as discussed in Remark 3.2 makes it possible to compare the rate presented here with those derived in the literature on equal terms. Note that Theorem 3.3 and Corollary 3.3.1 implicitly assume Condition 3.6(i), i.e. $b_{\tilde{m}} \to 0$. It is a necessary condition that the approximation error vanishes as $n \to \infty$ since $\alpha_m'\beta_m$ is only an approximation to the the infinite-dimensional linear functional θ. Despite the fact that $\underline{m}$ and $\overline{m}$ grow at the same rate in Corollary 3.3.1, the size of the model set $\tilde{m}$ diverges with the sample size.

3.3.1. Estimation of the Variance

This section provides non-asymptotic bounds on the estimation error of the covariance matrices which ultimately leads to the consistency and rate of convergence of the vari-

ance of $\hat{\theta}_m$ uniformly in $m \in \mathcal{M}_n$. The central covariance matrix is

$$\Sigma_m := \mathrm{E}[Z_{i,m} Z_{i,m}'],$$

with its natural estimator

$$\hat{\Sigma}_m := \mathbb{E}_n[Z_{i,m} Z_{i,m}'].$$

These quantities play an important role in the least-squares estimator

$$\beta_m = \Sigma_m^{-1} \mathrm{E}[Z_{i,m} Y_i] \quad \text{and} \quad \hat{\beta}_m = \hat{\Sigma}_m^{-1} \mathbb{E}_n[Z_{i,m} Y_i].$$

Theorem D.1 in Appendix D, which has also been extensively used in Chapter 2, provides the following non-asymptotic bound on high-dimensional random matrices. These results depend on an important inequality due to Rudelson (1999), see Lemma D.1. There is a rich literature on bounds of this type for random matrices, see Tropp (2015) for a extensive summary and Belloni et al. (2015) and Chen and Christensen (2015b) for applications in econometrics.

Corollary 3.3.2. *Under Conditions 3.1 to 3.4, it follows that*

$$\|\hat{\Sigma}_m - \Sigma_m\| \lesssim_P \nu_n \sqrt{\frac{\xi_{\overline{m}}^2 \log m}{n}} \quad \text{and} \quad \|\hat{\Sigma}_m^{-1} - \Sigma_m^{-1}\| \lesssim_P \nu_n \sqrt{\frac{\xi_{\overline{m}}^2 \log m}{n}}. \tag{3.8}$$

Furthermore,

$$\max_{m \in \mathcal{M}_n} \|\hat{\Sigma}_m - \Sigma_m\| \lesssim_P \nu_n \sqrt{\frac{\xi_{\overline{m}}^2 \tilde{m}^2 \log \overline{m}}{n}}, \tag{3.9}$$

and if the models indexed by $m \in \mathcal{M}_n$ are nested

$$\max_{m \in \mathcal{M}_n} \|\hat{\Sigma}_m - \Sigma_m\| \lesssim_P \nu_n \sqrt{\frac{\xi_{\overline{m}}^2 \log \overline{m}}{n}}. \tag{3.10}$$

Corollary 3.3.2 establishes that under very mild conditions, I can consistently estimate both Σ_m as well as its inverse which has the same rate. Similarly, the estimator $\hat{\Sigma}_m$ is uniformly consistent with a small $\tilde{m}$ penalty term. However, this penalty term disappears when the models are nested since $\max_{m \in \mathcal{M}_n} \|\hat{\Sigma}_m - \Sigma_m\| = \|\hat{\Sigma}_{\overline{m}} - \Sigma_{\overline{m}}\|$. Crucially the rate requires that

$$v_n^2 \xi_{\overline{m}}^2 \log \overline{m}/n \to 0,$$

for pointwise convergence and the stronger condition

$$v_n^2 \xi_{\overline{m}}^2 \tilde{m}^2 \log \overline{m}/n \to 0,$$

for uniform convergence. Note that this does not impose the need for any $m \in \mathcal{M}_n$ to diverge with the sample size. Next, I turn to the estimation of the variance of $\hat{\theta}_m$. For this purpose define the matrices

$$\Xi_m := \mathrm{E}[(\varepsilon_i + r_{i,m})^2 Z_{i,m} Z_{i,m}'] \quad \text{and} \quad \Omega_m := \Sigma_m^{-1} \Xi_m \Sigma_m^{-1},$$

and their plug-in estimators

$$\hat{\Xi}_m := \mathbb{E}_n[\hat{\varepsilon}_{i,m}^2 Z_{i,m} Z_{i,m}'] \quad \text{and} \quad \hat{\Omega}_m := \hat{\Sigma}_m^{-1} \hat{\Xi}_m \hat{\Sigma}_m^{-1}.$$

The matrix $\hat{\Omega}_m$ is the well-known heteroskedasticity-consistent covariance estimator from the linear regression model which shows, once more, how sieve estimation can be viewed through the lens of the high-dimensional linear model. The important distinction in the nonparametric model is that $\hat{\varepsilon}_{i,m}$ is an estimator for $\varepsilon_i + r_{i,m}$ such that it implicitly nests the bias.

These matrices act as estimators of the the variance of $\hat{\theta}_m$ defined as

$$\sigma_m^2(x) := \frac{\alpha_m(x)' \Omega_m \alpha_m(x)}{n}, \tag{3.11}$$

with estimator

$$\hat{\sigma}_m^2(x) := \frac{\alpha_m(x)'\hat{\Omega}_m\alpha_m(x)}{n}. \tag{3.12}$$

This $\hat{\sigma}_m^2$ is used in the construction of the maximal t-statistic T_n in (4.4) in Chapter 4. The third and fourth main results establish a non-asymptotic bound on the spectral norm between $\hat{\Omega}_m$ and Ω_m and the ratio $\hat{\sigma}_m/\sigma_m$ uniformly over $m \in \mathcal{M}_n$.

Theorem 3.4. *Assume Conditions 3.1 to 3.5 hold. If* $(\nu_n \vee b_{\tilde{m}})\tilde{m}\xi_{\overline{m}} \lesssim \sqrt{n}$, *then*

$$\max_{m\in\mathcal{M}_n} \|\hat{\Xi}_m - \Xi_m\| \lesssim_P (\nu_n + b_{\tilde{m}})\sqrt{\frac{\xi_{\overline{m}}^2(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}}, \tag{3.13}$$

and

$$\max_{m\in\mathcal{M}_n} \|\hat{\Omega}_m - \Omega_m\| \lesssim_P (\nu_n + b_{\tilde{m}})\sqrt{\frac{\xi_{\overline{m}}^2(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}}. \tag{3.14}$$

Theorem 3.4 allows for uniformly consistent estimation of the covariance matrix Ω_m under stronger conditions than for Σ_m, namely that

$$(\nu_n^2 \vee b_{\tilde{m}}^2)(\xi_{\overline{m}}^2\tilde{m}^2\overline{m}\log\overline{m}\log\tilde{m})/n \to 0,$$

as stated in Condition 3.6(iii). Not surprisingly this requirement depends on the tails of ε_i through ν_n and on the bias $b_{\tilde{m}}$ due to the use of $\hat{\varepsilon}_{i,m}$ which accounts for both the noise as well as the bias term. This rate carries over into the uniform consistency of $\hat{\sigma}_m$ which is a direct consequence of Theorem 3.4.

Corollary 3.4.1. *With the same set-up as in Theorem 3.4 together with Condition 3.6(iii), for any* $\alpha \in \mathbb{S}^{m-1}$

$$\max_{m\in\mathcal{M}_n} \left|\frac{\hat{\sigma}_m(x)}{\sigma_m(x)} - 1\right| \lesssim_P (\nu_n + b_{\tilde{m}})\sqrt{\frac{\xi_{\overline{m}}^2(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}} = o(1). \tag{3.15}$$

Remark 3.3. The consistency is in terms of the ratio such that $\hat{\sigma}_m/\sigma_m = 1 + o_p(1)$ rather than the more common $|\hat{\sigma}_m - \sigma_m| = o_p(1)$. However, this will be sufficient for the inference procedure presented in Chapter 4.

4 Uniform Inference and Model Selection

4.1. Introduction

Consistency of series estimators is often a means to an end resulting in valid inference on a statistic of interest. Using the set-up and the results from Chapter 3, I suggest an inference procedure which delivers valid uniform confidence bands for θ capturing both the conditional mean function or its linear functionals in the nonparametric regression model. As in Chapter 3, the uniformity holds simultaneously over the space of the covariates as well as the model set. Uniform confidence bands for series estimators are interesting in their own right, but these bands have the added benefit that they are valid *regardless* of the model-selection technique used to decide on the dimension of the series expansion. Therefore, the inference procedure yields confidence bands which cover the process $\{\theta(x) : x \in \mathcal{X}\}$ and $\theta(x)$ for any fixed $x \in \mathcal{X}$. The confidence bands are asymptotically exact for the whole process, whereas they are conservative in the latter case. This is the price to pay for the generality of the set-up presented here.

The inference procedure described in Algorithm 4.1 yields a feasible critical value $\tilde{c}_n^*(\alpha)$ such that the uniform confidence band

$$\tilde{\mathcal{C}}_{n,m} := \left\{ \left[\hat{\theta}_m(x) - \tilde{c}_n^*(\alpha)\hat{\sigma}_m(x), \hat{\theta}_m(x) + \tilde{c}_n^*(\alpha)\hat{\sigma}_m(x) \right] : (x, m) \in \mathcal{I}_n \right\}, \tag{4.1}$$

is valid for the whole process $\{\theta(x) : x \in \mathcal{X}\}$ in the following sense

$$P\Big(\theta(x) \in \tilde{\mathcal{C}}_{n,m} \text{ for all } (x, m) \in \mathcal{I}_n\Big) = 1 - \alpha + o(1), \tag{4.2}$$

and valid, yet conservative, for any fixed $(x, m) \in \mathcal{I}_n$

$$P\Big(\theta(x) \in \tilde{\mathcal{C}}_{n,m}\Big) \geq 1 - \alpha + o(1). \tag{4.3}$$

A key assumption is the undersmoothing condition, Condition 3.6(i), which states that the bias in the uniform linearisation in Theorem 3.1 vanishes as the sample size diverges. Without this assumption, the uniform confidence bands asymptotically control the size for the pseudo-true process $\{\theta_m(x) : (x, m) \in \mathcal{I}_n\}$ or are conservative for $\theta_m(x)$ for any fixed $(x, m) \in \mathcal{I}_n$. See Remark 4.1 below for further discussion on the distinction between θ and θ_m.

In light of (4.2), the proposed critical value also allows for pointwise hypothesis testing

$$\mathrm{H}_0 : \theta(x) = \theta_0(x),$$

as well as testing for shape restrictions

$$\mathrm{H}_0 : \sup_{x \in \mathcal{X}} |\theta(x)| \leq 0.$$

The latter has received a lot of attention lately in both statistics and econometrics, see e.g. the review article by Chetverikov, Santos and Shaikh (2018). Sections 4.2.1 and 4.3.2 explore the theory and an application of testing for shape restrictions.

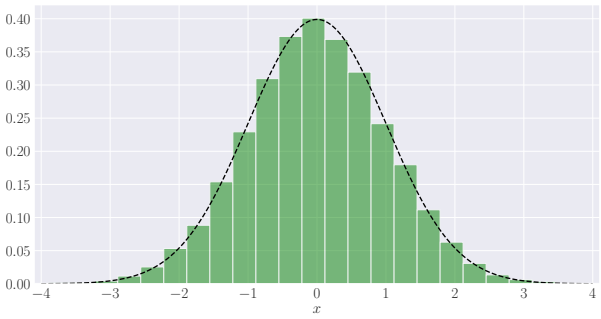
In order to obtain a valid critical value, I study the studentised empirical process

$$T_n := \sup_{(x,m) \in \mathcal{I}_n} \sqrt{n} \left| \frac{\hat{\theta}_m(x) - \theta}{\hat{\sigma}_m(x)} \right|, \tag{4.4}$$

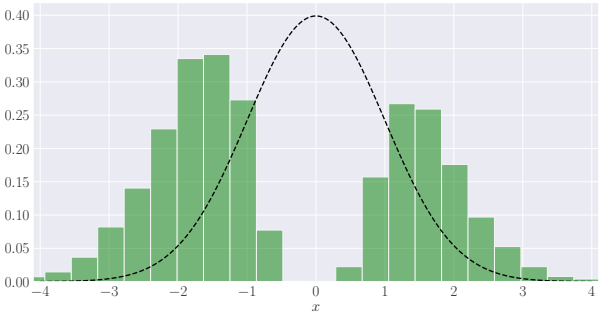
where $\hat{\sigma}_m^2(x)$ is the estimator of the variance of $\hat{\theta}_m(x)$ introduced in (3.12). Notice that this quantity is the usual t-statistic where the supremum is taken over the support and the set $\mathcal{M}_n$. The main results are two strong approximation theorems. The first one

shows that there is a random variable T_n^*, whose distribution could be simulated if σ_m were known, which is sufficiently close to T_n. The second main result shows that there exists another random variable $\tilde{T}_n^*$ sufficiently close to T_n^*. The advantage of using $\tilde{T}_n^*$ is that its distribution can easily be approximated by simulation. Hence, the inference procedure asymptotically controls the size when using a plug-in estimator for σ_m and when the critical value is estimated by simulation.

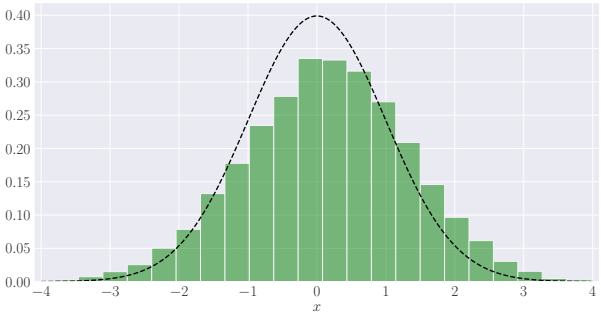
In real-world applications the 'true' model is always unknown which implies that researchers often search over multiple candidate models. This amounts to choosing the number of series terms to use in the approximation to θ. The uniformity over all the models considered is of paramount importance. To explain why this is the case, consider the following thought experiment. Suppose that researchers are interested in testing the hypothesis $\mathrm{H}_0 : g(x) = g_0(x)$ on the conditional mean function with the prior assumption that the hypothesis is false. They search over a multiple number of series terms and report the results using the expansion which delivered the highest t-statistic. This practice increases the probability of rejecting the null since the model is no longer fixed and the distribution of the t-statistics are no longer Gaussian, even asymptotically, as predicted by standard theory on series estimators (Newey 1997; Chen 2007). Figure 4.1 shows the empirical distribution of the t-statistics when m is fixed, m is chosen by K-fold cross-validation and m is the number of series terms which yielded the highest absolute t-value. The data comes from the Monte Carlo experiment in Section 4.3.1. It is clear that the usual Gaussian critical values are not appropriate when the number of series terms is chosen in a data-driven manner. The theory presented here does away with this problem by offering an inference procedure which is valid *regardless* of which algorithm is used to decide on the final dimension of the estimated model.



(a) Fixed m



(b) Maximal t-statistic



(c) 5-Fold Cross-Validation

Figure 4.1: Three examples of the empirical distribution of the t-statistics from the simulation study in Section 4.3.1 evaluated at $x = 0.5$ overlaid with the standard Gaussian density function where (a) keeps m fixed; (b) is the maximal t-statistic over all $m \in \mathcal{M}_n$; (c) picks an $\hat{m}$ using 5-fold cross-validation. The sample size in each case is 500 with 10,000 replications.

4.1.1. Related Literature

The paper closest to this chapter is Kang (2019) which established similar uniform inference procedures. An important contribution of his paper is that uniform-in-m confidence bands are correctly sized for fixed $x \in \mathcal{X}$. The uniform confidence bands presented here

will be conservative in the pointwise case which shows that it is important to continue studying pointwise inference procedures. Furthermore, he also proved the validity of the multiplier bootstrap in order to approximate the distribution of T_n. This is especially desirable when the estimator of V – the central covariance function in this chapter as defined in (4.6) – is non-singular. However, his paper did not incorporate the extension to shape restrictions which can be found in Section 4.2.1.[2]

The theory on the construction of uniform confidence bands in nonparametric statistics dates back to at least Bickel and Rosenblatt (1973). More recent contributions are Claeskens and Van Keilegom (2003), Giné and Nickl (2010) and Horowitz and Lee (2012). These works all rely on limit theorems of the statistic under investigation. In many applications, as is the case here, such limit distributions may not exist for models of increasing complexity as $n \to \infty$. Chernozhukov, Chetverikov and Kato (2014a) recognised this and developed the theory to construct uniform confidence bands through strong approximations which is beneficial in situations where a limit distribution does not exist or is difficult to simulate. The proofs of the main results heavily rely on their apparatus, see e.g. Theorem D.3 in Appendix D.

The use of uniform confidence bands for controlling multiple inference is not new. This problem has an extensive history in both statistics and econometrics, see e.g. White (2000) and Lehmann and Romano (2005). However, few papers make the link between uniform inference and data-driven tuning parameters. An important exception is Armstrong and Kolésar (2018). They established similar results to the ones derived in this chapter for uniform inference on kernel estimators of the conditional mean function. The benefit of their set-up is that they can tabulate critical values which only depend on

[2] At the time of writing I only knew of Kang (2018) and was unaware of the new version Kang (2019). The latter is a major improvement and provides many extensions over the earlier working paper version.

the kernel and the ratio of the maximal and minimal bandwidths $\overline{m}/\underline{m}$. As is the case here, their critical values lead to conservative confidence bands when the researcher wants to cover $\theta_m(x)$ for some fixed $(x, m) \in \mathcal{I}_n$ rather than the whole process $\{\theta_m(x) : (x, m) \in \mathcal{I}_n\}$.

There is a large and growing literature on post-selection inference which is related, but logically distinct from the results presented here. This body of research was initiated by the impossibility theorems in Leeb and Pötscher (2006). Seminal papers are Belloni, Chernozhukov and Hansen (2010, 2014). The former established valid inference in instrumental variable models after the first-stage regressors are chosen by regularisation. The latter derived methods for valid inference on high-dimensional parameters by orthogonalising the estimation equations with respect to the nuisance parameters. Both of these papers heavily rely on the theory of the Lasso and a full review on these techniques is available in Chernozhukov, Hansen and Spindler (2015). Another strand on research originating with Berk et al. (2013) developed valid post-selection inference, not on the fixed true parameter of interest, but on some pseudo-true parameter conditional on a model selection event. Loftus (2015) and Markovic, Xia and Taylor (2017) extended their work for cross-validated estimators using various estimation techniques such as the Lasso and forward stepwise regression. However, these techniques only apply to high-dimensional parameters in sparse models and they are computationally very intensive making them nearly impossible to solve when the number of regressors is greater than twenty.

4.2. Uniform Inference

The reason for using strong approximations instead of appealing to conventional convergence-in-distribution type arguments is the following. The complexity of the classes of functions indexing the Gaussian process in (4.4) above or (4.5) below,

$$\mathcal{F} = \bigcup_{m \in \mathcal{M}_n} \mathcal{F}_m,$$

for

$$\mathcal{F}_m = \left\{ (Z, \varepsilon) \mapsto \alpha_m' Z \varepsilon, \alpha_m \in \mathbb{S}^{m-1} \right\},$$

increases too quickly as the dimension of Z_m diverges with $n \to \infty$. This process is, therefore, not asymptotically equicontinuous which renders standard Donsker-type theorems as in van der Vaart and Wellner (1996) useless. Hence, it is impossible to establish a result of the form

$$T_n \Rightarrow \sup_{f \in \mathcal{F}} |G_\infty f|,$$

for some fixed, centred Gaussian process G_∞. The use of strong approximations circumvents this problem as it is a key step in order to show that the distribution of T_n can be approximated by a *sequence* of distributions of T_n^* or its feasible version $\tilde{T}_n^*$ which are defined below. In turn, this establishes that the uniform confidence bands using their $(1-\alpha)$-quantiles, respectively denoted by c_n^* or $\tilde{c}_n^*$, as critical values asymptotically control the size.

A strong approximation, or coupling, is a result which shows that T_n is sufficiently close to a sequence of random variables T_n^* on the same probability space. Define this random variable as

$$T_n^* := \sup_{(x,m) \in \mathcal{I}_n} \left| G_{n,m}(x) \right|, \tag{4.5}$$

where $G_n = G_{n,m}(x)$ is a tight, centred Gaussian process in $\ell^\infty(\mathcal{I}_n)$ with covariance function

$$V[(x,m),(\check{x},\check{m})] := \mathrm{E}[G_{n,m}(x)G_{n,\check{m}}(\check{x})'] = \frac{\alpha_m(x)'\Sigma_m^{-1}\,\mathrm{E}\left[Z_{i,m}Z'_{i,\check{m}}\varepsilon_i^2\right]\Sigma_{\check{m}}^{-1}\alpha_{\check{m}}(\check{x})}{\sigma_m(x)\sigma_{\check{m}}(\check{x})}. \tag{4.6}$$

All the conditions maintained here are the same as the ones stated in Section 3.2. Since the goal is uniform inference rather than uniform convergence there are stricter requirements on the number of moments of the errors in (1.1). Condition 4.1 below replaces Condition 3.4.

Condition 4.1. *(Noise) The errors satisfy: (i)* $\mathrm{E}[\varepsilon_i|X] = 0$*; (ii)* $c < \mathrm{E}[\varepsilon_i^2|X]$ *a.s. for some* $c > 0$*; (iii)* $\mathrm{E}[\varepsilon_i^s|X] \lesssim 1$ *a.s. for some* $s > 4$*.*

Theorem 4.1. *Assume that Conditions 3.1 to 3.3, 3.5 and 4.1 hold. Furthermore, if Condition 3.6(i), (iii),* $\bar{r}_{1n} + \bar{r}_{2n} \lesssim \delta_n^{-1}$ *and*

$$\frac{\xi_{\overline{m}}^2(\overline{m}\log\overline{m}\log\tilde{m})^\gamma}{n^{1-1/s}} \to 0,$$

for some $\gamma > 0$ *hold, then*

$$\left|T_n - T_n^*\right| = o_P\left(\delta_n^{-1}\right), \tag{4.7}$$

for $\delta_n = \sqrt{\overline{m}\log\overline{m}\log\tilde{m}}$*.*

The approximation $T_n \approx T_n^*$, as a result of Theorem 4.1, heuristically suggests using the distribution of T_n^* to approximate the distribution of T_n. The accuracy of the approximation crucially depends on how quickly the bias and the estimation error, $\bar{r}_{1n} + \bar{r}_{2n}$ from Theorem 3.1, vanish. The strong approximation is an important result, but does not in itself deliver that c_n^* is, even asymptotically, a valid critical value. Theorem 4.1 yields, under suitable conditions, a bound in the Kolmogorov distance which is crucial in the proof of Theorem 4.2 which does establish its validity.

Corollary 4.1.1. *With the same set-up as in Theorem 4.1*

$$\sup_{t\in\mathbf{R}}\left|P(T_n \leq t) - P(T_n^* \leq t)\right| = o(1). \tag{4.8}$$

Theorem 4.2. *Under the conditions in Theorem 4.1 it follows that*

$$P\left(T_n \leq c_n^*\right) = 1 - \alpha + o(1), \tag{4.9}$$

such that

$$P\left(\theta(x) \in \mathcal{C}_{n,m} \text{ for all } (x,m) \in \mathcal{I}_n\right) = 1 - \alpha + o(1), \tag{4.10}$$

and that for any fixed $(x,m) \in \mathcal{I}_n$

$$P\left(\theta(x) \in \mathcal{C}_{n,m}\right) \geq 1 - \alpha + o(1). \tag{4.11}$$

Remark 4.1. Theorem 4.2, and Theorem 4.4 below, continue to hold when replacing θ in the definition of T_n with the pseudo-parameter θ_m, which is $\alpha_m'\hat{\beta}_m$ such that $r_{\theta,m} := r_m = \theta - \theta_m$. In fact, it would be more correct to do so, as inference on θ inevitably depends on the undersmoothing condition in Condition 3.6(i). The undersmoothing condition ultimately depends on multiple factors which are nearly impossible to verify not in the least because the smoothness of the function class to which θ belongs is unknown. Adaptive inference where the choice of the tuning parameter adapts to the unknown smoothness, see e.g. Chapter 8 in Giné and Nickl (2016), still requires the choice of tuning parameters such that it does not solve the problem of how to choose m. It is still very much an open problem on how to satisfactorily deal with the bias. In practice, inference is always conducted on θ_m with the hope that the undersmoothing condition holds and r_m vanishes as $n \to \infty$. If the bias is of no importance to practitioners and they are happy to conduct inference on pseudo-true parameters then the inference procedure and its theory carry over to the HDL model described in Section 2.1 as well. This showcases once more the

flexibility of series estimators. When there is one true model – fixed or dependent on n – this no longer holds. The bias in such a setting does not necessarily need to vanish as any chosen model in $\mathcal{M}_n$ can be arbitrarily bad.

Theorem 4.2 shows that c_n^* is a critical value which delivers valid inference on the whole process $\{\theta(x) : x \in \mathcal{X}\}$ or is conservative for $\theta(x)$ and any fixed $x \in \mathcal{X}$. As discussed in Remark 4.1 the use of c_n^* is also correct for the pseudo-true process $\{\theta_m(x) : (x, m) \in \mathcal{I}_n\}$ under milder assumptions, i.e. the usual undersmoothing condition is not necessary. The catch, however, is that c_n^* is not feasible as the residuals, Ω_m or $\sigma_{x,m}$ are unknown. For this purpose, define

$$\tilde{T}_n^* = \sup_{(x,m)\in\mathcal{I}_n} \left|\tilde{G}_{n,m}(x)\right|, \tag{4.12}$$

where $G_{n,m}$ conditional on the data $Z_m = (Z_{1,m}, \ldots, Z_{n,m})'$ is a centred Gaussian process with covariance function

$$\hat{V}[(x, m), (\check{x}, \check{m})] = \frac{\alpha_m(x)'\hat{\Sigma}_m^{-1}\sum_{i=1}^n\sum_{j=1}^n[Z_{i,m}Z'_{j,\check{m}}\hat{\varepsilon}_{i,m}\hat{\varepsilon}_{j,\check{m}}]\hat{\Sigma}_{\check{m}}^{-1}\alpha_{\check{m}}(\check{x})}{\hat{\sigma}_m(x)\hat{\sigma}_{\check{m}}(\check{x})}. \tag{4.13}$$

Theorem 4.3 solves this problem by showing that $\tilde{T}_n^*$, whose distribution can be simulated by using plug-in estimators for the unknown quantities, is close to the sequence T_n^*.

Theorem 4.3. *Under Conditions 3.1 to 3.3, 3.5, 3.6 and 4.1 and* $\overline{r}_{1n} + \overline{r}_{2n} = o_P(1/\sqrt{\overline{m}\log m \log \tilde{m}})$, *for some* $\tau_n \to 0$ *it holds that*

$$P\left(\left|\tilde{T}_n^* - T_n^*\right| \geq \frac{\tau_n}{\sqrt{\overline{m}\log\tilde{m}}}\right) = o(1). \tag{4.14}$$

Theorem 4.4 using Theorem 4.3 establishes that $\tilde{c}_n^*$ also asymptotically controls the size of the uniform confidence bands. Algorithm 4.1 describes how to compute this critical value by simulation.

Theorem 4.4. *Under the conditions in Theorem 4.1 it follows that*

$$P\Big(T_n \leq \tilde{c}_n^*(1-\alpha)\Big) = 1 - \alpha + o(1), \tag{4.15}$$

such that

$$P\Big(\theta(x) \in \tilde{\mathcal{C}}_{n,m} \text{ for all } (x,m) \in \mathcal{I}_n\Big) = 1 - \alpha + o(1), \tag{4.16}$$

and that for any fixed $(x,m) \in \mathcal{I}_n$

$$P\Big(\theta(x) \in \tilde{\mathcal{C}}_{n,m}\Big) \geq 1 - \alpha + o(1). \tag{4.17}$$

I finish this section by describing the algorithm to compute $\tilde{c}_n^*$ by simulation on a fine, but discrete grid $\mathcal{X}^{grid}$ on $\mathcal{X}$. The maximum, as all sets are finite, is computed over the set $\mathcal{J}_n := \mathcal{X}^{grid} \times \mathcal{M}_n$ with cardinality $|\mathcal{J}_n| = \tilde{m}|\mathcal{X}^{grid}|$.

Algorithm 4.1.

(1) Compute $\hat{V}$ with typical elements as described in (4.13) for any $(x,m) \in \mathcal{J}_n$;

(2) For $b = 1, \ldots, B$, take i.i.d draws of the $|\mathcal{J}_n|$-dimensional Gaussian random vector

$$\left(\tilde{G}_{m,x}^{n,b}\right)_{(x,m)\in\mathcal{J}_n} \sim \mathcal{N}(0, \hat{V});$$

(3) Compute the maximal t-statistics

$$T_n^b = \max_{(x,m)\in\mathcal{J}_n} \left|\tilde{G}_{m,x}^{n,b}\right|;$$

(4) Estimate the critical value $\tilde{c}_n^*(\alpha)$ by taking the $(1-\alpha)$-sample quantile of $\{T_n^b : 1 \leq b \leq B\}$;

(5) Use $\tilde{c}_n^*(\alpha)$ to construct the uniform confidence band in (4.1) or to test the desired hypothesis as usual.

4.2.1. Testing for Shape Restrictions

The procedure laid out in this chapter can also be used for testing shape restrictions on linear functionals θ of the conditional mean function g. Section 1.2 gives some examples of linear functionals: partial derivatives and integrals are of particular interest in economics and econometrics as discussed in Section 4.3.2 below. The hypothesis of interest is

$$\mathrm{H}_0 : \sup_{x\in\mathcal{X}} \theta(x) \leq 0 \quad \text{vs.} \quad \mathrm{H}_1 : \sup_{x\in\mathcal{X}} \theta(x) > 0. \tag{4.18}$$

For this purpose, define the one-sided test statistic

$$T_{n,m}^{shape} = \sup_{x\in\mathcal{X}} \sqrt{n}\frac{\hat{\theta}_m(x)}{\hat{\sigma}_m}, \tag{4.19}$$

Observe, that under H_0

$$\sup_{x\in\mathcal{X}} \sqrt{n}\frac{\hat{\theta}_m(x)}{\sigma_m} \leq \sup_{x\in\mathcal{X}} \sqrt{n}\frac{\hat{\theta}_m(x)-\theta(x)}{\sigma_m} = T_n^{os},$$

for T_n^{os} the one-sided equivalent of T_n. Let $\tilde{c}_n^{os}$ be the $(1-\alpha)$-quantile the distribution of T_n^{os} conditional of the data. Reject the null in favour of the alternative when $T_{n,m}^{shape} > \tilde{c}_n^{os}(1-\alpha)$. This critical value can be estimated in the same way as described in Algorithm 4.1 by removing the absolute values.

Theorem 4.5. *Assume that Conditions 3.1 to 3.3, 3.5, 3.6 and 4.1 hold. Then,*

$$P\Big(T_{n,m}^{shape} < \tilde{c}_n^{os}(1-\alpha)\Big) \geq 1-\alpha+o(1),$$

if $\overline{r}_{1n} + \overline{r}_{2n} = o_P(1/\sqrt{\overline{m}\log\overline{m}\log\tilde{m}})$.

Theorem 4.5 shows that using $\tilde{c}_n^{os}(1-\alpha)$ asymptotically controls the size of tests for shape restrictions as in (4.18). The construction of the critical values for testing inequality hypotheses does not take into account whether the linear functional θ is on the boundary

or not. Hence, combining this critical value with the method of moment selection (Chernozhukov, Hong and Tamer 2007; Andrews and Soares 2010) could increase the power of the test when all inequalities are indeed slack.

4.3. Numerical Results

4.3.1. Monte Carlo Experiment

To assess the finite sample performance of the proposed inference procedure, I conduct a small Monte Carlo experiment. The model is

$$Y_i = \arctan\left[\left(2X_i + \frac{1}{2}\right)\log\left(2X_i + \frac{1}{2}\right)\right] + \varepsilon_i,$$

where X_i and ε_i are i.i.d samples drawn from a uniform distribution on $[0, 1]$ and a standard Gaussian distribution for each replication. The approximating functions are B-Splines of order 2 or 3 with equally spaced knot sequences in $[0, 1]$.

Table 4.1: Summary of the simulation study parameters.

n	repetitions	order	knots (#)	$\underline{m}$	$\overline{m}$	$\lvert\mathcal{M}_n\rvert$
100	10,000	$\{2, 3\}$	$\{1, 2\}$	3	4	4
250	10,000	$\{2, 3\}$	$\{2, \ldots, 4\}$	4	7	6
500	10,000	$\{2, 3\}$	$\{2, \ldots, 6\}$	4	9	10
1000	10,000	$\{2, 3\}$	$\{2, \ldots, 8\}$	4	11	14

The model set is the Cartesian product of the orders of the B-Splines and the knot sequences considered. Hence, the size of $\mathcal{M}_n$ can grow with the sample size. The minimum number of parameters $\underline{m}$ remains fixed, but $\overline{m}$ grows at an increasing rate as n diverges.

The parameters are summarised in Table 4.1. Figure 4.2 plots the average coverage using the standard Gaussian critical values (Naive) and the ones suggested in this chapter based on the maximal t-statistic (Max. t-stat) for 10,000 replications and $n = 100, 250, 500, 1000$. I consider two confidence bands to cover g_m:

1. using the estimator with the largest absolute t-statistic for all $m \in \mathcal{M}_n$

$$\left\{ \max_{m \in \mathcal{M}_n} \hat{g}_m(x) - c_n^{(j)}(1-\alpha)\hat{\sigma}_m(x), \max_{m \in \mathcal{M}_n} \hat{g}_m(x) + c_n^{(j)}(1-\alpha)\hat{\sigma}_m(x) \right\},$$

2. using the estimator chosen with 5-fold cross-validation from the set $\mathcal{M}_n$

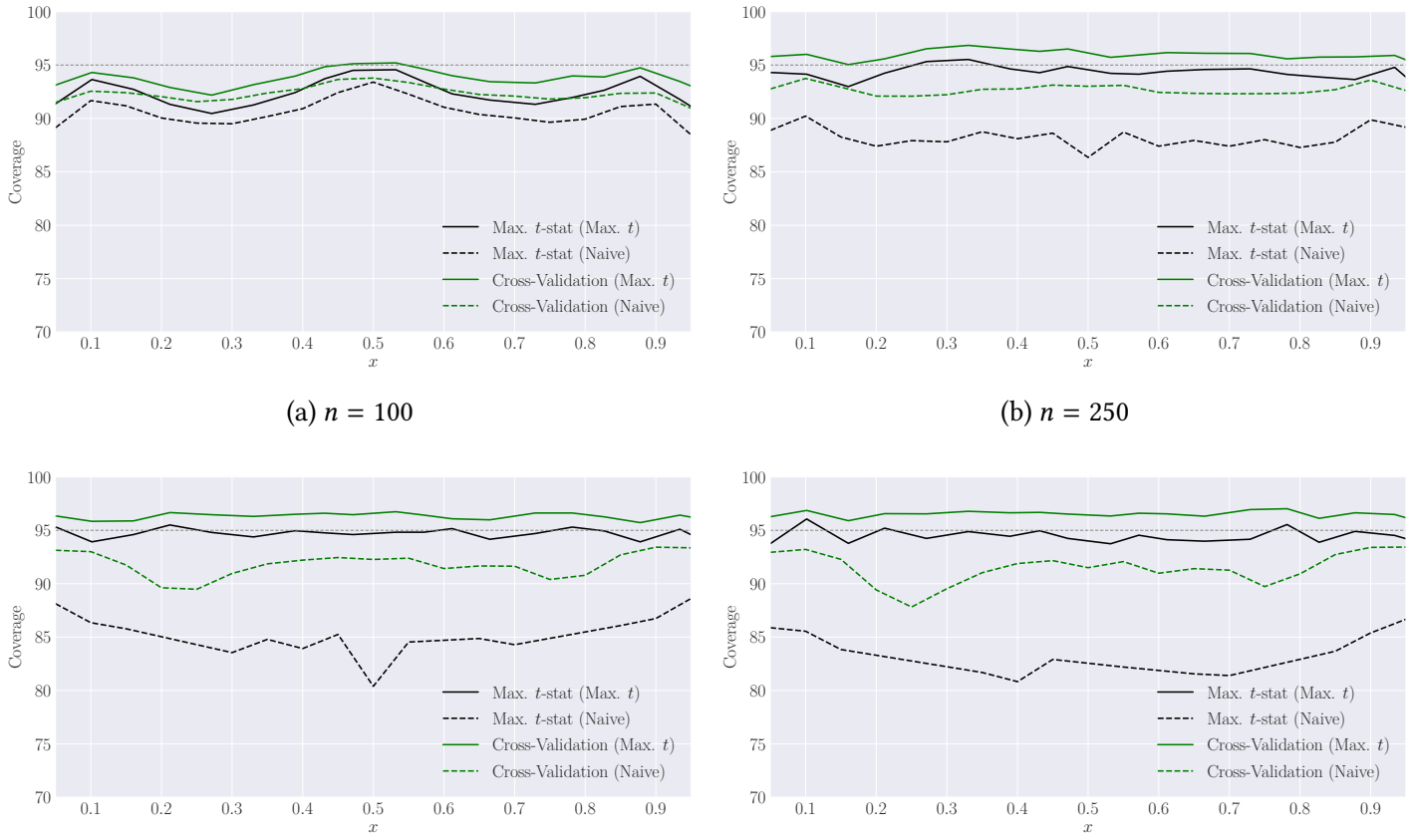
$$\left\{ \hat{g}_m^{cv}(x) - c_n^{(j)}(1-\alpha)\hat{\sigma}_m(x), \hat{g}_m^{cv}(x) + c_n^{(j)}(1-\alpha)\hat{\sigma}_m(x) \right\},$$

where $c_n^{(j)}$ is either the naive and incorrect critical value or $\tilde{c}_n^*$.

Not surprisingly, the critical values derived in this chapter perform best with the naive ones significantly undercovering the conditional mean function for any x in the grid and any of the sample sizes. The empirical coverage for both the maximal t-statistic and the cross-validated estimator improves as the sample size diverges. The cross-validated estimator results in conservative confidence bands which is also predicted by the theory as the cross-validated estimator is not necessarily the one which delivers the maximal t-statistic.

The empirical coverage level is slightly below the nominal level even for the maximal t-statistic for $n = 1000$. This is probably due to the fact that the model set is too large or that $\overline{m}$ diverged too quickly in this case. Yet, for larger sample sizes the empirical coverage is close to the nominal level as predicted by the theory showing that these critical values perform well even when the standard deviation and the critical value need

to be estimated. Similar results were obtained for the 90 and 97.5 per cent levels and are included in Appendix C.3.



(a) $n = 100$

(b) $n = 250$

(c) $n = 500$

(d) $n = 1000$

Figure 4.2: Coverage using the critical values proposed in this chapter (Max. t, —) and the incorrect critical values from the standard Gaussian tables (Naive, – –) at a 95 per cent level for $n = 100, 250, 500, 1000$ applied to both the maximal t-statistic and the t-statistics from a cross-validated estimator. Computed over an equally spaced grid of 20 points in $(0, 1)$.

4.3.2. Application to Gasoline Demand

I illustrate the inference methods developed in this chapter by applying them the nonparametric estimation of the demand function for gasoline using US household data. The data is from Blundell, Horowitz and Parey (2012) which is based on the 2001 National Household Travel Survey (NHTS).[3] The data set is comprised of household-level data from telephone surveys conducted in 2001. The main variables in the analysis are the annual gasoline consumption Q, the gasoline price P and the household income Y all denoted in US dollars. The dataset contains only the data for which P lies within the 5th and 95th percentiles of observed gasoline prices as reported in the original ORNL (2001) data. This results in 2,912 observations. As in Blundell, Horowitz and Parey (2012), I identify three income subgroups based on the midpoints \$42,500, \$57,500 and \$72,500 which will be referred to as the lower, middle and upper income groups respectively. The data is selected into a subgroup if the household income lies within 0.5 of that group's midpoint (in log terms). Figure 4.3 plots the log of gasoline prices versus the log of gasoline demand. Table C.4 reports descriptive statistics on the main variables in the data set. For a more in-depth description of the data, see Blundell, Horowitz and Parey (2012) or ORNL (2001) on the actual implementation and details on the survey.

A benchmark model for the demand for gasoline is the parametric log-linear specification. For shorthand, define $p = \log P$, $q = \log Q$ and $y = \log Y$ and X and extra set of regressors explained below which yields the log-linear model

$$q = \beta_0 + \beta_1 p + \beta_2 y + X'\delta + \varepsilon, \tag{4.20}$$

This is a workhorse model to estimate the average demand function of gasoline, see e.g. Hausman and Newey (1995), Schmalensee and Stoker (1999), Yatchew and No (2001) and

[3] The data can be downloaded from https://doi.org/10.7910/DVN/0YALNP.

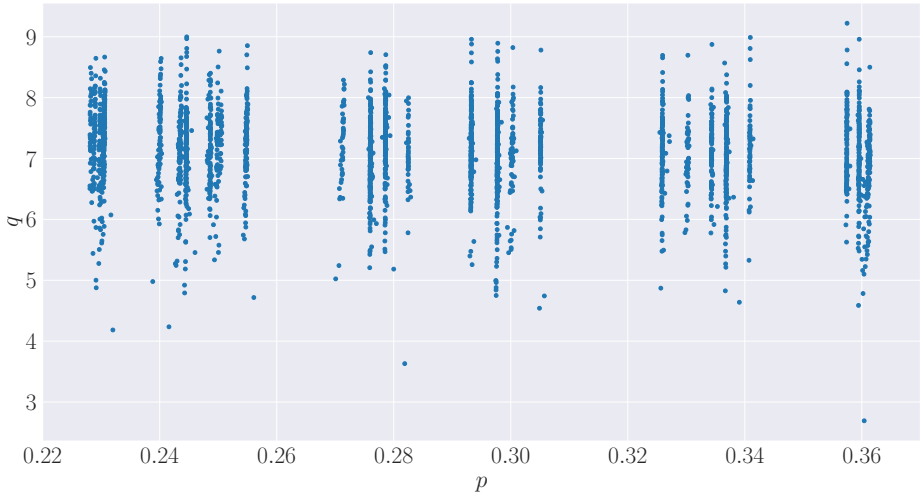


Figure 4.3: Scatter plot of the log of gasoline prices and the log of gasoline consumption.

Blundell, Horowitz and Parey (2012). Table C.5 reports the estimation results from (4.20) estimated on the full dataset and the lower, middle and upper income subgroups using the main variables described above and extra control variables to capture heterogeneity in gasoline demand at the household level. The extra covariates include the log of age of the survey respondent, log of household size, log of the number of drivers, number of workers in household, a public transport indicator and 12 variables on urbanisation and population density.

There is no good reason why, a priori, the average demand function for gasoline should be linear in prices. Therefore, I drop the log-linear specification and instead estimate the partially linear model

$$q = g(p) + \beta_2 y + X'\delta + u. \tag{4.21}$$

Blundell, Horowitz and Parey (2012) estimate a similar model to (4.21) using kernels instead of series estimators. They remark that the estimates of the demand functions for all

three subgroups are nonsensical as the price elasticities in multiple regions are positive. They propose to estimate the demand function under a Slutsky condition which stabilises the estimates and makes them conform to standard economic theory. Rather than imposing the Slutsky restriction, I directly test the monotonicity of the price elasticity of US gasoline demand using the tools developed in this chapter. The price elasticity is

$$\theta : \mathcal{P} \to \mathbf{R} : p \mapsto \frac{\partial g(p)}{\partial p},$$

and the hypothesis of interest is

$$\mathrm{H}_0 : \theta(p) \leq 0 \text{ for all } p \in \mathcal{P} \quad \text{vs} \quad \mathrm{H}_1 : \theta(p) > 0 \text{ for some } p \in \mathcal{P}.$$

The benefit of parametrising g in terms of p instead of P means that I can directly compute the price elasticity as $\partial q/\partial p \equiv \partial g(p)/\partial p$. The unknown function g is estimated with B-Splines of degrees 2 or 3 with 5 possible knots sequences: no knots, 1, 2, 3 or 4 equally spaced knots in $\mathcal{P}$. Hence, the model set $\mathcal{M}_n$ includes 10 models. I choose the model for constructing the test statistic by 5-fold cross-validation. An estimator for θ follows directly from the estimator for g as $\hat{\theta}_m(p) = \frac{\partial Z_m(p)}{\partial p}'\hat{\beta}$. The one-sided test statistic is as explained in Section 4.2.1 for $\hat{m}$ the cross-validated choice of $m \in \mathcal{M}_n$

$$T := \sup_{p \in \mathcal{P}} \frac{\hat{\theta}_{\hat{m}}(p)}{\hat{\sigma}_{\hat{m}}(p)},$$

which corresponds to the least-favourable choice of $\theta = 0$. To estimate the distribution of T_n, I use the procedure described in Algorithm 4.1 to obtain a sequence

$$\hat{T}_n^b := \sup_{(p,m) \in \mathcal{J}_n} \tilde{G}_{m,p}^{n,b},$$

for $b = 1, \ldots, B$ where $B = 100,000$ and where $(\tilde{G}_{m,p}^{n,b})_{(p,m)\in\mathcal{J}_n} \sim \mathcal{N}(0, \hat{V})$ for $\mathcal{P}^{grid}$ a fine grid on the support $\mathcal{P}$ and $\mathcal{J}_n := \mathcal{P}^{grid} \times \mathcal{M}_n$. The critical value is approximated by

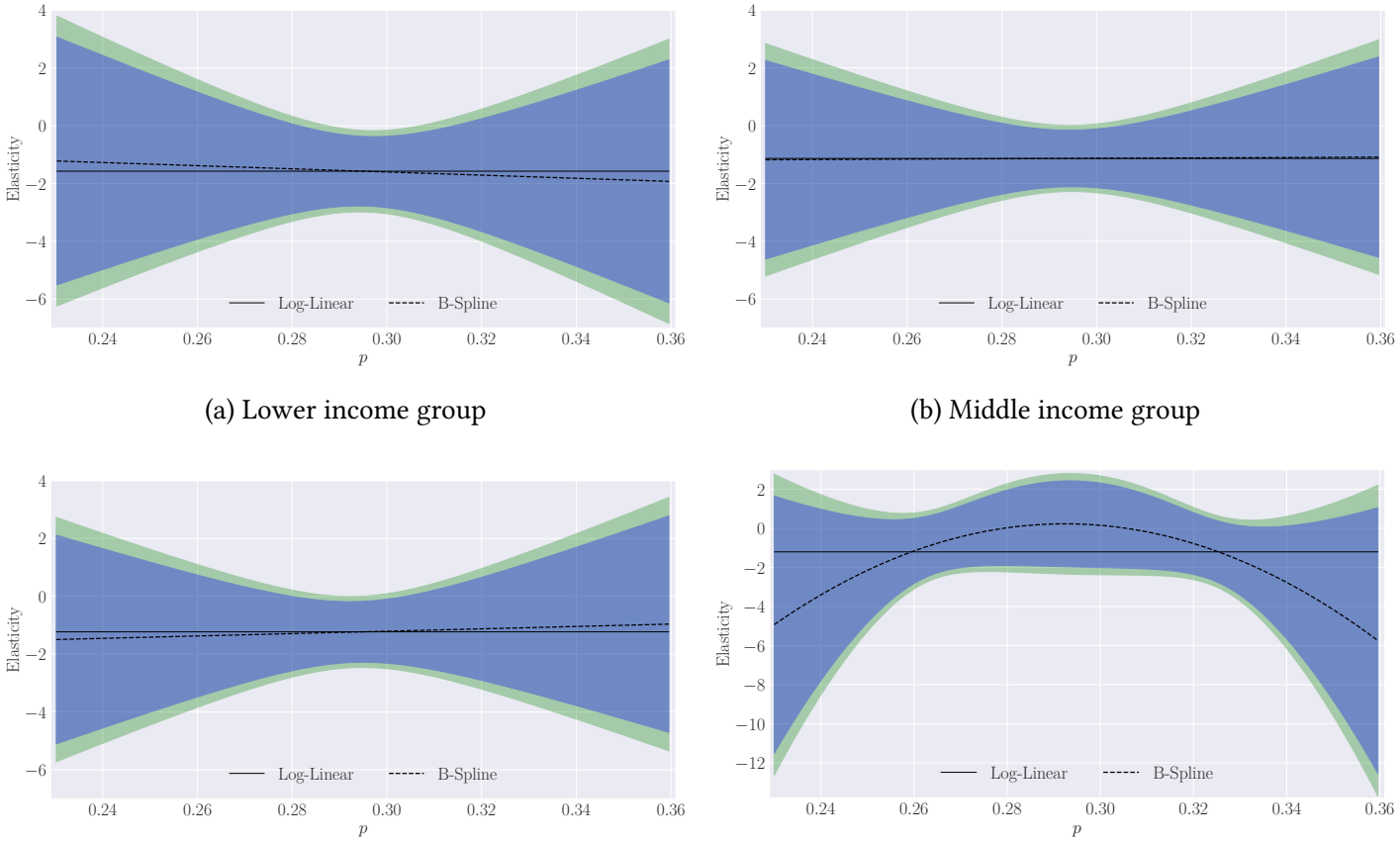
$$\hat{c}_n(\alpha) := (1-\alpha)\text{-quantile of } \left\{\hat{T}_n^{(1)}, \ldots, \hat{T}_n^{(B)}\right\}. \tag{4.22}$$

Table 4.2 contains the results of the monotonicity test for the full sample and the three income groups. The cross-validation procedure chose the most parsimonious model for the three sub-sample which implies linear elasticities. For the full sample, a more complicated model was chosen such that the price elasticity is quadratic. The naive critical value was computed by only taking the supremum over $\mathcal{P}$ and not over the model set in (4.22). The correct critical values are only slightly larger than the naive ones indicating that there is a small cost to pay for uniformity over the model set. The test does not reject the null for monotonicity in any of the specifications, supporting what one would expect a priori.

Table 4.2: Results from monotonicity test for the model chosen by 5FCV.

	degree	knots	T-statistic	$\hat{c}_n$ (naive)	$\hat{c}_n$	p-value	reject
Full sample	3	None	0.321	3.286	3.793	0.374	No
Lower sample	2	None	-0.876	3.286	3.791	0.809	No
Middle sample	2	None	-0.963	3.286	3.792	0.832	No
Upper sample	2	None	-0.789	3.286	3.793	0.785	No

The elasticities as well as uniform confidence bands using the critical values suggested in this chapter (green) and the naive critical values (blue) are plotted in (4.22). Despite the fact that the elasticities are linear in the three income sub-samples, they are in fact close to the constant elasticities estimated in the log-linear model. This casts doubt on the validity of the nonparametric specification in this application and is in stark contrast to the results found in Blundell, Horowitz and Parey (2012). Interestingly, the uniform confidence bands in the correct case are not that much wider than in the naive case. This again shows that there is a small penalty to be paid in order to get valid confidence bands after model selection.



(a) Lower income group

(b) Middle income group

(c) Upper income group

(d) Full sample

Figure 4.4: Estimates of the price elasticities of gasoline demand for the full sample and the three income sub-samples. The series estimators are B-Splines with the degree and knots chosen by 5-fold cross-validation, whereas the log-linear estimates are the constant elasticities as reported in Table C.5. The blue bands are uniform confidence bands over $\mathcal{P}$ and the green bands are uniform confidence bands over $\mathcal{P} \times \mathcal{M}_n$.

References

Allen, D. M. (1974). "The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction". *Technometrics* **16**(1), pp. 125–127.

Andrews, D. W. K. (1991a). "Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors". *Journal of Econometrics* **47**(2-3), pp. 359–377.

——— (1991b). "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models". *Econometrica* **59**(2), pp. 307–345.

Andrews, D. W. K. and Soares, G. (2010). "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection". *Econometrica* **78**(1), pp. 119–157.

Arlot, S. and Celisse, A. (2010). "A survey of cross-validation procedures for model selection". *Statistics Surveys* **4**, pp. 40–79.

Armstrong, T. B. and Kolésar, M. (2018). "A Simple Adjustment for Bandwidth Snooping". *Review of Economic Studies* **85**(2), pp. 732–765.

Belloni, A., Chernozhukov, V. and Fernandez-Val, I. (2019). "Conditional quantile processes based on series or many regressors". *Journal of Econometrics* **213**(1), pp. 4–29.

Belloni, A., Chernozhukov, V. and Hansen, C. (2010). *Lasso Methods for Gaussian Instrumental Variables Models*. https://arxiv.org/abs/1012.1297.

——— (2014). "Inference on Treatment Effects after Selection among High-Dimensional Controls". *Review of Economic Studies* **81**(2), pp. 608–650.

Belloni, A. et al. (2012). "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain". *Econometrica* **80**(6), pp. 2369–2429.

Belloni, A. et al. (2015). "Some new asymptotic theory for least squares series: Pointwise and uniform results". *Journal of Econometrics* **186**(2), pp. 345–366.

Berk, B. R. et al. (2013). "Valid post-selection inference". *Annals of Statistics* **41**(2), pp. 1–47.

Bickel, P. J. and Rosenblatt, M. (1973). "On some global measures on the deviations of density function estimates". *Annals of Statistics* **1**(6), pp. 1071–1095.

Blundell, R., Chen, X. and Kristensen, D. (2007). "Semi-nonparametric IV estimation of shape-invariant Engel curves". *Econometrica* **75**(6), pp. 1613–1669.

Blundell, R., Horowitz, J. L. and Parey, M. (2012). "Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation". *Quantitative Economics* **3**(1), pp. 29–51.

Cattaneo, M. D., Jansson, M. and Newey, W. K. (2018a). "Alternative asymptotics and the partially linear model with many regressors". *Econometric Theory* **34**(2), pp. 277–301.

——— (2018b). "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity". *Journal of the American Statistical Association* **113**(523), pp. 1350–1361.

Chen, X. (2007). "Large Sample Sieve Estimation of Semi-Nonparametric Models". *Handbook of Econometrics*. Ed. by Heckman, J. J. and Leamer, E. E. Vol. 6B. Chap. 76, pp. 5549–5632.

Chen, X. and Christensen, T. M. (2015a). *Optimal Sup-Norm Rates, Adaptivity and Inference In Nonparametric Instrumental Variables Estimation.* http://cowles.yale.edu/sites/default/files/files/pub/d19/d1923-r.pdf.

——— (2015b). "Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions". *Journal of Econometrics* **188**(2), pp. 447–465.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2014a). "Anti-concentration and honest, adaptive confidence bands". *Annals of Statistics* **42**(5), pp. 1787–1818.

——— (2014b). "Gaussian approximation of suprema of empirical processes". *Annals of Statistics* **42**(4), pp. 1564–1597.

Chernozhukov, V., Hansen, C. and Spindler, M. (2015). "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach". *Annual Review of Economics* **7**, pp. 649–688.

Chernozhukov, V., Hong, H. and Tamer, E. (2007). "Estimation and confidence regions for parameter sets in econometric models". *Econometrica* **75**(5), pp. 1243–1284.

Chetverikov, D., Liao, Z. and Chernozhukov, V. (2019). *On cross-validated Lasso*. https://arxiv.org/abs/1605.02214.

Chetverikov, D., Santos, A. and Shaikh, A. M. (2018). "The Econometrics of Shape Restrictions". *Annual Review of Economics* **10**, pp. 31–63.

Claeskens, G. and Van Keilegom, I. (2003). "Bootstrap confidence bands for regression curves and their derivatives". *Annals of Statistics* **31**(6), pp. 1852–1884.

De Boor, C. (1978). *A Practical Guide to Splines*. Berlin: Springer.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation*. Berlin: Springer-Verlag.

Dudley, R. M. (1967). "The sizes of compact subsets of Hilbert space and continuity of Gaussian processes". *Journal of Functional Analysis* **1**(3), pp. 290–330.

Einmahl, U. and Mason, D. M. (2005). "Uniform in bandwidth consistency of kernel-type function estimators". *Annals of Statistics* **33**(3), pp. 1380–1403.

Geisser, S. (1975). "The Predictive Sample Reuse Method with Applications". *Journal of the American Statistical Association* **70**(350), pp. 320–328.

Giné, E. and Nickl, R. (2010). "Confidence bands in density estimation". *Annals of Statistics* **38**(2), pp. 1122–1170.

——— (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. New York: Cambridge University Press.

Greenshtein, E. and Ritov, Y. (2004). "Persistence in high-dimensional linear predictor selection and the virtue of overparametrization". *Bernoulli* **10**(6), pp. 971–988.

Grenander, U. (1981). *Abstract Inference*. Probability and Statistics Series. New York: John Wiley & Sons.

Hansen, B. E. (2014). "Nonparametric Sieve Regression: Least Squares, Averaging Least Squares, and Cross-Validation". *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, pp. 215–248.

——— (2015). *A Unified Asymptotic Distribution Theory for Parametric and Noparametric Least Squares*. https://www.ssc.wisc.edu/~bhansen/preliminary/rnormal4.pdf.

Hausman, J. A. and Newey, W. K. (1995). "Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss". *Econometrica* **63**(6), pp. 1445–1476.

Homrighausen, D. and McDonald, D. (2013). "The lasso, persistence, and cross-validation". *International Conference on Machine Learning* **28**(1), pp. 1031–1039.

Horowitz, J. L. and Lee, S. (2012). "Uniform confidence bands for functions estimated nonparametrically with instrumental variables". *Journal of Econometrics* **168**(2), pp. 175–188.

Hsu, D., Kakade, S. M. and Zhang, T. (2014). "Random Design Analysis of Ridge Regression". *Foundations of Computational Mathematics* **14**(3), pp. 569–600.

Huang, J. Z. (2003). "Local asymptotics for polynomial spline regression". *Annals of Statistics* **31**(5), pp. 1600–1635.

Kang, B. (2018). "Inference in Nonparametric Series Estimation with Data-Dependent Number of Series Terms". Working Paper.

——— (2019). *Inference in Nonparametric Series Estimation with Data-Dependent Number of Series Terms*. https://arxiv.org/abs/1909.12162.

Leeb, H. and Pötscher, B. M. (2006). "Can one estimate the conditional distribution of post-model-selection estimators?" *Annals of Statistics* **34**(5), pp. 2554–2591.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer Texts in Statistics. New York: Springer.

Li, K.-C. (1987). "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set". *Annals of Statistics* **15**(3), pp. 958–975.

Liu, R. Y. (1988). "Bootstrap procedures under some non-i.i.d. models". *Annals of Statistics* **16**(4), pp. 1696–1708.

Loftus, J. R. (2015). *Selective inference after cross-validation*. https://arxiv.org/abs/1511.08866.

Markovic, J., Xia, L. and Taylor, J. (2017). *Comparison of prediction errors: Adaptive p-values after cross-validation*. https://arxiv.org/abs/1703.06559.

Newey, W. K. (1997). "Convergence rates and asymptotic normality for series estimators". *Journal of Econometrics* **79**(1), pp. 147–168.

Newey, W. K. and Powell, J. L. (2003). "Instrumental Variable Estimation of Nonparametric Models". *Econometrica* **71**(5), pp. 1565–1578.

ORNL (2001). *2001 National Household Travel Survey*. https://nhts.ornl.gov/2001/usersguide/UsersGuide.pdf.

Riphahn, R. T., Wambach, A. and Million, A. (2003). "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation". *Journal of Applied Econometrics* **18**(4), pp. 387–405.

Rudelson, M. (1999). "Random vectors in the isotropic position". *Journal of Functional Analysis* **72**, pp. 60–72.

Schmalensee, R. and Stoker, T. M. (1999). "Household Gasoline Demand in the United States". *Econometrica* **67**(3), pp. 645–662.

Shao, J. (1993). "Linear Model Selection by Cross-Validation". *Journal of the American Statistical Association* **88**(422), pp. 486–494.

Stone, C. J. (1982). "Optimal global rates of convergence for nonparametric regression". *Annals of Statistics* **10**(4), pp. 1040–1053.

——— (1994). "The Use of Polynomial Splines and their Tensor Products in Multivariate Function Estimation". *Annals of Statistics* **22**(1), pp. 118–184.

Stone, M. (1974). "Cross-Validatory Choice and Assessment of Statistical Predictions". *Journal of the Royal Statistical Society, Series B* **36**(2), pp. 111–147.

Tropp, J. A. (2015). *An Introduction to Matrix Concentration Inequalities*. http://arxiv.org/abs/1501.01571.

Van de Geer, S. A. (2000). *Empirical Processes in M-Estimation*. Cambridge: Cambridge University Press.

——— (2008). "High-dimensional generalized linear models and the Lasso". *Annals of Statistics* **36**(2), pp. 614–645.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer-Verlag.

Wahba, G. and Wold, S. (1975). "A Completely Automatic French Curve: Fitting Spline Functions by Cross Validation". *Communications in Statistics* **4**(1), pp. 1–17.

White, H. (2000). "A Reality Check for Data Snooping". *Econometrica* **68**(5), pp. 1097–1126.

Wu, C. F. J. (1986). "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis". *Annals of Statistics* **14**(4), pp. 1261–1295.

Yatchew, A. and No, J. A. (2001). "Household Gasoline Demand in Canada". *Econometrica* **69**(6), pp. 1697–1709.

A Appendix of Chapter 2

A.1. Proofs

Proof of Lemma 2.1. Let $\mathcal{D}_m$ be the domain over which (2.5) is minimised with m regressors. Then,

$$\begin{aligned} P\Big(\mathcal{E}(m^{CV}, m_*) < 0\Big) = P\Big(R(m^{CV}) < R(m_*)\Big) &\leq P(m^{CV} > m_*) \\ &\leq P(\overline{m} > m_*) \\ &\leq \frac{\overline{m}}{m_*} = o(1), \end{aligned}$$

where the first inequality follows because $\beta^{BLP}_{m_*}$ is a minimiser of (2.5) over $\mathcal{D}_{m_*}$ and $\mathcal{D}_{m^{CV}} \subset \mathcal{D}_{m_*}$ for any $m^{CV} < m_*$ by Condition 2.1, the second and third inequalities follow because $\overline{m} \geq m^{CV}$ by definition and Markov's inequality respectively and the last equality by the assumptions of the Lemma. The final claim follows directly if $\overline{m} = m_*$ replaces the assumption that $\overline{m} = o(m_*)$. ■

For ease of notation, let $\alpha_m = (-1, \beta_m)$, $\hat{\alpha}_m = (-1, \hat{\beta}_m)$ and let $\mathcal{S}_K = \{S_1, \ldots, S_K\}$ be the set of folds. The subscripts s and $-s$, respectively, denote computed with the data in fold s or without the data in fold s. This notation carries over to the covariance matrices Σ_n, Σ_s and Σ_{-s} where, respectively, all the data is used, the data in fold s is used and all the data without fold s is used. The actual dimension of the covariance matrices depends on the dimension of α_m in the quadratic forms in the proof Theorem 2.1 below. Quantities

with a '^' are the sample analogues of the population quantities. The proof then relies on writing $R(\hat{\beta}_{m^{CV}})$, $\hat{R}_K(m^{CV})$ and $R(\beta^{BLP}_{m_*})$ as quadratic forms. For example,

$$R(\hat{\beta}_{m^{CV}}) = \hat{\alpha}'_{m^{CV}} \Sigma_n \hat{\alpha}_{m^{CV}}.$$

Proof of Theorem 2.1. Define $\check{m} := \overline{m} \wedge m_*$ and decompose the excess risk in five terms

$$\mathcal{E}(m^{CV}, m_*) = (I) + (II) + (III) + (IV) + (V),$$

where

$$\begin{aligned}
(I) &= R(\hat{\beta}_{m^{CV}}) - \hat{R}_K(m^{CV}) \\
(II) &= \hat{R}_K(m^{CV}) - \hat{R}_K(\check{m}) \\
(III) &= \hat{R}_K(\check{m}) - \hat{R}(\hat{\beta}_{\check{m}}) \\
(IV) &= \hat{R}(\hat{\beta}_{\check{m}}) - \hat{R}(\hat{\beta}_{m_*}) \\
(V) &= \hat{R}(\hat{\beta}_{m_*}) - R(\beta^{BLP}_{m_*}).
\end{aligned}$$

Notice that both $(II) \leq 0$ and $(IV) \leq 0$. The former is true as m^{CV} is the solution to (2.3) such that $\hat{R}_K(m^{CV})$ is minimal. In the latter case, if $\check{m} = m_*$ then $\hat{R}(\hat{\beta}_{\check{m}}) = \hat{R}(\hat{\beta}_{m_*})$ by definition and if $\check{m} = \overline{m}$ then $\hat{R}(\hat{\beta}_{\check{m}}) \leq \hat{R}(\hat{\beta}_{m_*})$ as both $\hat{\beta}_{\check{m}}$ and $\hat{\beta}_{m_*}$ are least-squares solution, but $\hat{\beta}_{\check{m}}$ is a constrained version of $\hat{\beta}_{m_*}$. Therefore, only (I), (III) and (V) need to be bounded.

Step 1: Bound (I).

$$\begin{aligned}
R(\hat{\beta}_{m^{CV}}) - \hat{R}_K(m^{CV}) &= \hat{\alpha}'_{m^{CV}} \Sigma_n \hat{\alpha}_{m^{CV}} - \frac{1}{K} \sum_{s \in \mathcal{S}_K} \hat{\alpha}'_{m^{CV},-s} \hat{\Sigma}_s \hat{\alpha}_{m^{CV},-s} \\
&= \hat{\alpha}'_{m^{CV}} \Sigma_n \hat{\alpha}_{m^{CV}} + \hat{\alpha}'_{m^{CV}} \hat{\Sigma}_n \hat{\alpha}_{m^{CV}} - \hat{\alpha}'_{m^{CV}} \hat{\Sigma}_n \hat{\alpha}_{m^{CV}} - \frac{1}{K} \sum_{s \in \mathcal{S}_K} \hat{\alpha}'_{m^{CV},-s} \hat{\Sigma}_s \hat{\alpha}_{m^{CV},-s} \\
&= \hat{\alpha}'_{m^{CV}} (\Sigma_n - \hat{\Sigma}_n) \hat{\alpha}_{m^{CV}} + \hat{\alpha}'_{m^{CV}} \hat{\Sigma}_n \hat{\alpha}_{m^{CV}} - \frac{1}{K} \sum_{s \in \mathcal{S}_K} \hat{\alpha}'_{m^{CV},-s} \hat{\Sigma}_s \hat{\alpha}_{m^{CV},-s}.
\end{aligned}$$

For the first term, by Condition 2.4, Lemmas A.1 and A.2 and Theorem D.1

$$\begin{aligned}\hat{\alpha}'_{m^{CV}}(\Sigma_n - \hat{\Sigma}_n)\hat{\alpha}_{m^{CV}} &\leq (1 + \|\hat{\beta}_{m^{CV}}\|^2)\|\hat{\Sigma}_n - \Sigma_n\| \\ &\lesssim_P (1+A)\|\Sigma_n\|^{1/2}\sqrt{\frac{\xi^2_{m_*}\log(1+m_*)}{n}}.\end{aligned}$$

Rewrite the remaining two terms as follows

$$\begin{aligned}\hat{\alpha}'_{m^{CV}}\hat{\Sigma}_n\hat{\alpha}_{m^{CV}} - \frac{1}{K}\sum_{s\in\mathcal{S}_K}\hat{\alpha}'_{m^{CV},-s}\hat{\Sigma}_s\hat{\alpha}_{m^{CV},-s} &= \frac{1}{K}\sum_{s\in\mathcal{S}_K}\hat{\alpha}'_{m^{CV}}\hat{\Sigma}_n\hat{\alpha}_{m^{CV}} - \frac{1}{K}\sum_{s\in\mathcal{S}_K}\hat{\alpha}'_{m^{CV},-s}\hat{\Sigma}_s\hat{\alpha}_{m^{CV},-s} \\ &\quad + \frac{1}{K}\sum_{s\in\mathcal{S}_K}\hat{\alpha}'_{m^{CV},-s}\hat{\Sigma}_n\hat{\alpha}_{m^{CV},-s} - \frac{1}{K}\sum_{s\in\mathcal{S}_K}\hat{\alpha}'_{m^{CV},-s}\hat{\Sigma}_n\hat{\alpha}_{m^{CV},-s} \\ &= \frac{1}{K}\sum_{s\in\mathcal{S}_K}(\hat{\alpha}'_{m^{CV}}\hat{\Sigma}_n\hat{\alpha}_{m^{CV}} - \hat{\alpha}'_{m^{CV},-s}\hat{\Sigma}_n\hat{\alpha}_{m^{CV},-s}) \\ &\quad + \hat{\alpha}'_{m^{CV},-s}(\hat{\Sigma}_n - \hat{\Sigma}_s)\hat{\alpha}_{m^{CV},-s}.\end{aligned}$$

Since $\hat{\alpha}_{m^{CV}}$ minimizes $\alpha'\hat{\Sigma}_n\alpha$ over $\mathbf{R}^{m^{CV}+1}$, $\hat{\alpha}'_{m^{CV}}\hat{\Sigma}_n\hat{\alpha}_{m^{CV}} \leq \hat{\alpha}'_{m^{CV},-s}\hat{\Sigma}_n\hat{\alpha}_{m^{CV},-s}$ for any $s \in \mathcal{S}_K$ such that

$$\begin{aligned}&\leq \frac{1}{K}\sum_{s\in\mathcal{S}_K}\hat{\alpha}'_{m^{CV},-s}(\hat{\Sigma}_n - \hat{\Sigma}_s)\hat{\alpha}_{m^{CV},-s} \\ &= \frac{1}{K}\sum_{s\in\mathcal{S}_K}\hat{\alpha}'_{m^{CV},-s}(\hat{\Sigma}_n - \Sigma_n)\hat{\alpha}_{m^{CV},-s} + \hat{\alpha}'_{m^{CV},-s}(\Sigma_n - \hat{\Sigma}_s)\hat{\alpha}_{m^{CV},-s}.\end{aligned}$$

Then by Lemma A.2

$$\begin{aligned}&\leq \frac{1}{K}\sum_{s\in\mathcal{S}_K}\|\hat{\alpha}_{m^{CV},-s}\|^2\Big(\|\hat{\Sigma}_n - \Sigma_n\| + \|\Sigma_n - \hat{\Sigma}_s\|\Big) \\ &\lesssim_P \left(1 + \frac{K}{K-1}A\right)\Big(1+\sqrt{K}\Big)\|\Sigma_n\|^{1/2}\sqrt{\frac{\xi^2_{m_*}\log(1+m_*)}{n}},\end{aligned}$$

where the final inequality again holds by Condition 2.4, Lemma A.1 and Theorem D.1 and the fact that the sample size in each fold is n/K. Combining these bounds yields

$$(I) \lesssim_P \left[1 + A + \Big(1+\sqrt{K}\Big)\left(1 + \frac{K}{K-1}A\right)\right]\|\Sigma_n\|^{1/2}\sqrt{\frac{\xi^2_{m_*}\log(1+m_*)}{n}}. \tag{A.1}$$

Step 2: Bound (III). For any given fold $s \in \mathcal{S}_K$ consider

$$\begin{aligned}
\hat{\alpha}'_{\check{m},-s}\hat{\Sigma}_s\hat{\alpha}_{\check{m},-s} - \hat{\alpha}'_{\check{m}}\hat{\Sigma}_n\hat{\alpha}_{\check{m}} &= \hat{\alpha}'_{\check{m},-s}\hat{\Sigma}_s\hat{\alpha}_{\check{m},-s} - \hat{\alpha}'_{\check{m}}\hat{\Sigma}_n\hat{\alpha}_{\check{m}} + \hat{\alpha}'_{\check{m},-s}\hat{\Sigma}_{-s}\hat{\alpha}_{\check{m},-s} - \hat{\alpha}'_{\check{m},-s}\hat{\Sigma}_{-s}\hat{\alpha}_{\check{m},-s} \\
&= \hat{\alpha}'_{\check{m},-s}(\hat{\Sigma}_s - \hat{\Sigma}_{-s})\hat{\alpha}_{\check{m},-s} - \hat{\alpha}'_{\check{m}}\hat{\Sigma}_n\hat{\alpha}_{\check{m}} + \hat{\alpha}'_{\check{m},-s}\hat{\Sigma}_{-s}\hat{\alpha}_{\check{m},-s} \\
&= \hat{\alpha}'_{\check{m},-s}(\hat{\Sigma}_s - \hat{\Sigma}_{-s})\hat{\alpha}_{\check{m},-s} - \hat{\alpha}'_{\check{m}}\hat{\Sigma}_n\hat{\alpha}_{\check{m}} + \hat{\alpha}'_{\check{m},-s}\hat{\Sigma}_{-s}\hat{\alpha}_{\check{m},-s} \\
&\quad + \hat{\alpha}'_{\check{m}}\hat{\Sigma}_{-s}\hat{\alpha}_{\check{m}} - \hat{\alpha}'_{\check{m}}\hat{\Sigma}_{-s}\hat{\alpha}_{\check{m}} \\
&\leq \hat{\alpha}'_{\check{m},-s}(\hat{\Sigma}_s - \hat{\Sigma}_{-s})\hat{\alpha}_{\check{m},-s} + \hat{\alpha}'_{\check{m}}(\hat{\Sigma}_{-s} - \hat{\Sigma}_n)\hat{\alpha}_{\check{m}},
\end{aligned}$$

where the inequality follows because $\hat{\alpha}'_{\check{m},-s}$ minimizes $\alpha'\hat{\Sigma}_{-s}\alpha$ such that the final two terms together are negative. Therefore,

$$(III) \leq \frac{1}{K}\sum_{s\in\mathcal{S}_K} \hat{\alpha}'_{\check{m},-s}(\hat{\Sigma}_s - \hat{\Sigma}_{-s})\hat{\alpha}_{\check{m},-s} + \hat{\alpha}'_{\check{m}}(\hat{\Sigma}_{-s} - \hat{\Sigma}_n)\hat{\alpha}_{\check{m}}$$

Analogously to Step 1, by applying the triangle inequality and Lemma A.2 to each term individually

$$\begin{aligned}
&\leq \frac{1}{K}\sum_{s\in\mathcal{S}_K} \|\hat{\alpha}_{\check{m},-s}\|^2\Big(\|\hat{\Sigma}_s - \Sigma_n\| + \|\hat{\Sigma}_{-s} - \Sigma_n\|\Big) + \|\hat{\alpha}_{\check{m}}\|^2\Big(\|\hat{\Sigma}_{-s} - \Sigma_n\| + \|\hat{\Sigma}_n - \Sigma_n\|\Big) \\
&\lesssim_P \left(1 + \frac{K}{K-1}A\right)\|\Sigma_n\|^{1/2}\left[\sqrt{K}\sqrt{\frac{\xi^2_{m_*}\log(1+m_*)}{n}} + \sqrt{\frac{K}{K-1}}\sqrt{\frac{\xi^2_{m_*}\log(1+m_*)}{n}}\right] \\
&\quad + \Big(1 + A_n\Big)\|\Sigma_n\|^{1/2}\sqrt{\frac{K}{K-1}}\sqrt{\frac{\xi^2_{m_*}\log(1+m_*)}{n}} \\
&= \left[\left(\sqrt{K} + \sqrt{\frac{K}{K-1}}\right)\left(1 + \frac{K}{K-1}A\right) + \sqrt{\frac{K}{K-1}}(1+A)\right]\|\Sigma_n\|^{1/2}\sqrt{\frac{\xi^2_{m_*}\log(1+m_*)}{n}}.
\end{aligned}$$

Where the probabilistic bound is similar to the previous steps and the final equality is an algebraic simplification.

Step 3: Bound (V). First, consider

$$\|\beta^{BLP}_{m_*}\| = \|(\mathrm{E}[Z_{i,m_*}Z'_{i,m_*}])^{-1}\,\mathrm{E}[Z_{i,m_*}Y_i]\|$$

$$\lesssim \left(\sum_{j=1}^{m_*} \mathrm{E}\,|Z_{ij}^{m_*} Y_i|^2\right)^{1/2}$$
$$\lesssim \sqrt{m_*},$$

where the first inequality follows by Condition 2.3(iii) and Jensen's inequality, and the second by Hölder's inequality and Condition 2.3(ii). Then,

$$\begin{aligned}\hat{R}(\hat{\beta}_{m_*}) - R(\beta_{m_*}^{BLP}) &= \hat{\alpha}_{m_*}'\hat{\Sigma}_n\hat{\alpha}_{m_*} - \alpha_{m_*}^{BLP\prime}\Sigma_n\alpha_{m_*}^{BLP} \\ &= \hat{\alpha}_{m_*}'\hat{\Sigma}_n\hat{\alpha}_{m_*} - \alpha_{m_*}^{BLP\prime}\hat{\Sigma}_n\alpha_{m_*}^{BLP} + \alpha_{m_*}^{BLP\prime}\hat{\Sigma}_n\alpha_{m_*}^{BLP} - \alpha_{m_*}^{BLP\prime}\Sigma_n\alpha_{m_*}^{BLP}\end{aligned}$$

As before, $\hat{\alpha}_{m_*}'\hat{\Sigma}_n\hat{\alpha}_{m_*}$ minimises $\alpha'\hat{\Sigma}_n\alpha$ such that the first two terms together are negative

$$\begin{aligned}&\leq \alpha_{m_*}^{BLP\prime}\hat{\Sigma}_n\alpha_{m_*}^{BLP} - \alpha_{m_*}^{BLP\prime}\Sigma_n\alpha_{m_*}^{BLP} \\ &\leq \|\alpha_{m_*}^{BLP}\|^2\|\hat{\Sigma}_n - \Sigma_n\| \\ &\lesssim_P (1+\sqrt{m_*})\|\Sigma_n\|^{1/2}\sqrt{\frac{\xi_{m_*}^2\log(1+m_*)}{n}},\end{aligned}$$

where the penultimate and final inequalities are the results of Lemma A.2, and the bound on $\|\beta_{m_*}^{BLP}\|$ and Theorem D.1 respectively.

Step 4: Noticing that $K/(K-1) \leq 2$ for $K \geq 2$ such that

$$1 + A + \left(1+\sqrt{K}\right)\left(1+\frac{K}{K-1}A\right) + \left(\sqrt{K}+\sqrt{\frac{K}{K-1}}\right)\left(1+\frac{K}{K-1}A\right) + \sqrt{\frac{K}{K-1}}(1+A) + 1 + \sqrt{m_*},$$

is bounded from above by

$$1 + A + \left(1 + \sqrt{2} + 2\sqrt{K}\right)(1+2A) + \sqrt{2}(1+A) + 1 + \sqrt{m_*},$$

and combining the results from Steps 1–3 delivers the bound on $\mathcal{E}(m^{CV}, m_*)$. ■

Proof of Corollary 2.1.1. Both equalities follow directly from substituting Condition 2.5 in the non-asymptotic bound from Theorem 2.1. ■

Proof of Corollary 2.1.2. In the HDL model with bounded regressors, the assumption that $\mathrm{E}[|Y_i|^s] \lesssim 1$ implies that $\mathrm{E}[\max_{1\le i\le n} Y_i^2] + \max_{1\le i\le n} \|X_i\|^2] \lesssim n^{2/s}$, see Lemma B.2. Hence, the first result follows from $\xi_m \lesssim n^{1/s}$. For B-Splines in the NP model, it holds that $\sup_{x\in\mathcal{X}} \|Z_m(x)\| \lesssim \sqrt{m}$ for any $m \in \mathbf{N}$ (Newey 1997). This combined with

$$\mathrm{E}\left[\max_{1\le i\le n} Y_i^2\right] \lesssim n^{2/s},$$

yields $\xi_m \lesssim n^{1/s} \vee \sqrt{m}$ which establishes the second claim. The third claim for polynomials follows analogously with $\xi_m \lesssim n^{1/s} \vee m$ as $\sup_{x\in\mathcal{X}} \|Z_m(x)\| \lesssim m$. ∎

A.1.1. Additional Technical Results

Lemma A.1. *Assume that Condition 2.3 holds. Furthermore, let $\eta := \eta_n \ge 0$ be a sequence converging to zero sufficiently slowly. Then with probability at least $1 - \eta$*

$$\|\hat{\beta}_m\| \lesssim \sqrt{m} + \sqrt{\xi_m^2 \log(m/\eta)/n}, \tag{A.2}$$

for the least-squares estimator in (2.2)*. Moreover, if $\log 1/\eta \le C \log m$ then*

$$\|\hat{\beta}_m\| \lesssim \sqrt{m} + \sqrt{\xi_m^2 \log m/n}. \tag{A.3}$$

Proof. Start from

$$\begin{aligned}
\|\hat{\beta}_m\| &= \|(\mathbb{E}_n Z_i Z_i')^{-1} \mathbb{E}_n Z_i Y_i\| \\
&\le \lambda_{\min}^{-1}(\mathbb{E}_n Z_i Z_i') \|\mathbb{E}_n Z_i Y_i\| \\
&\le \sqrt{m} \lambda_{\min}^{-1}(\mathbb{E}_n Z_i Z_i') \|\sqrt{n}\mathbb{E}_n Z_i Y_i\|_\infty / \sqrt{n}.
\end{aligned}$$

By Theorem D.1, the event $\lambda_{\min}^{-1}(\mathbb{E}_n Z_i Z_i') \lesssim 1$ holds with probability at least $1 - \eta$ so long as $\eta \ge \alpha$. Next, by the triangle inequality such that

$$\|\sqrt{n}\mathbb{E}_n Z_i Y_i\|_\infty \le \|\sqrt{n}\mathbb{E}_n(Z_i Y_i - \mathrm{E}[Z_i Y_i])\|_\infty + \|\sqrt{n}\,\mathrm{E}\,\mathbb{E}_n Z_i Y_i\|_\infty \lesssim \sqrt{\log(m/\eta)} + \sqrt{n},$$

again with probability at least $1 - \eta$ which holds using Condition 2.3(ii) where the first term follows from Lemma 5 in Belloni et al. (2012) using the theory of self-normalizing sums. The second claim follows immediately from $\log 1/\eta \lesssim \log m$. ■

Lemma A.2. *Let $a \in \mathbf{R}^m$ and $A \in \mathbf{R}^{m \times m}$ be a symmetric matrix. Then*

$$a'Aa \leq \|A\| \|a\|^2.$$

Proof. The matrix A, because it is symmetric, can be diagonalised

$$A = Q \Lambda Q',$$

where Q is the orthogonal matrix of eigenvectors and Λ the diagonal matrix of eigenvalues of A. Then,

$$\begin{aligned} a'Aa &= a'(Q\Lambda Q')a \\ &= b'\Lambda b \\ &= \sum_{j=1}^{m} \lambda_j b_j^2 \\ &\leq \max_{1 \leq j \leq m} |\lambda_j| \sum_{j=1}^{m} b_j^2 \\ &= \|A\| \|a\|^2. \end{aligned}$$

The second and final equalities follows by defining $b = Q'a$, the fact that $b'b = a'QQ'a = a'a$ and A being symmetric such that its operator norm is equal to its maximal absolute eigenvalue. ■

A.2. Monte Carlo Experiment Set-Up and Extra Results

The HDL model in Design 1 uses regressors from the real-world dataset in Riphahn, Wambach and Million (2003) to construct a dataset (X, Y^*) where the true model is chosen from a subset of the regressors in X after which the parameters are estimated using the observed Y. The pseudo-response Y^* is then constructed from the fitted model by discarding the estimated residuals, but re-adding Gaussian noise. For any sample size n the design is Gaussian with fixed regressors, but as reported in Table A.1 the design changes with n. The models which the selection procedure could select from for each n are summarised in Table A.2. Figures A.1 to A.4 below contain the same results as in Section 2.3, but in these figures they show the deciles of the excess risk for each sample size.

Table A.1: Data-generating models in Design 1.

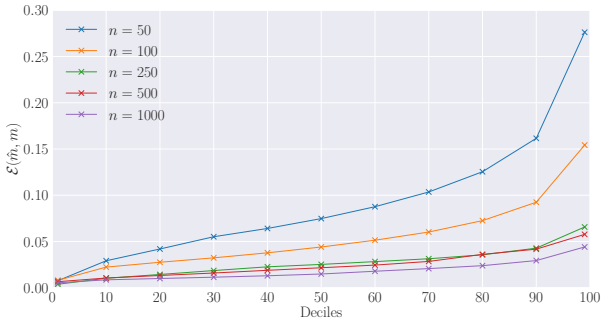
n	Model[a]
50	–
100	$\beta_3(age \times educ)$
250	$\beta_3(age \times educ) + \beta_4 age^2$
500	$\beta_3(age \times educ) + \beta_4 age^2 + \beta_5(age \times female) + \beta_6(educ \times female) + \beta_7(age \times educ \times female)$
1000	$\beta_3(age \times educ) + \beta_4 age^2 + \beta_5(age \times female) + \beta_6(educ \times female) + \beta_7(age \times educ \times female)$ $+\beta_8 hhkids + \beta_9 beamt + \beta_{10} handper$

[a] *Notes: Dependent variable is log of household income,* $\log hhninc$. *Model column reports* $X_2'\delta$ *in* $Y = X_1'\beta + X_2'\delta + \varepsilon$ *where* X_1 *contains a constant, age and educ and is common across all sample sizes.*

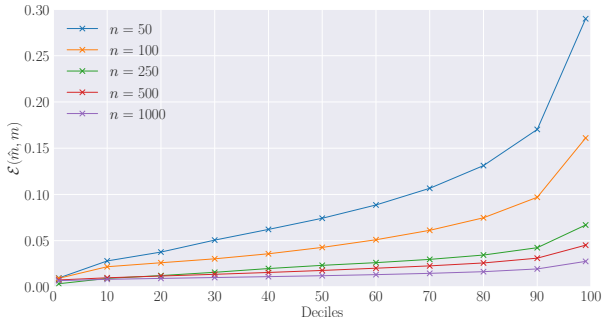
Table A.2: Regressors included in the model set for Design 1.

	Regressors[a]									
n	age	age^2	age^3	$educ$	$age \times educ$	$female$	$age \times female$	$educ \times female$	$age \times educ \times female$	$dummies$
				×						
	×									
	×	×								
50	×			×						
	×	×		×						
	×			×	×					
	×	×		×	×					
	×	×	×							
100	×	×	×	×						
	×	×	×	×	×					
				×						
	×					×				
	×	×				×				
	×			×		×				
	×	×		×		×				
500	×			×	×	×				
	×	×		×	×	×				
	×	×	×			×				
	×	×	×	×		×				
	×	×	×	×	×	×				
	×	×		×	×	×	×	×	×	
1000	×	×		×	×	×	×	×	×	×

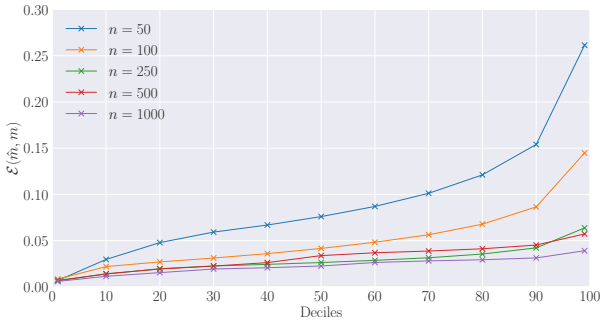
[a]*Notes*: Each sample size also contains the models for the smaller sample sizes such that $\mathcal{M}_n \subseteq \mathcal{M}_{n'}$ for $n < n'$. The sample size 250 is not included as it includes the same models as sample size 100. The *dummies* categories collects *hhkids*, *beamt* and *handper*.



(a) 5FCV

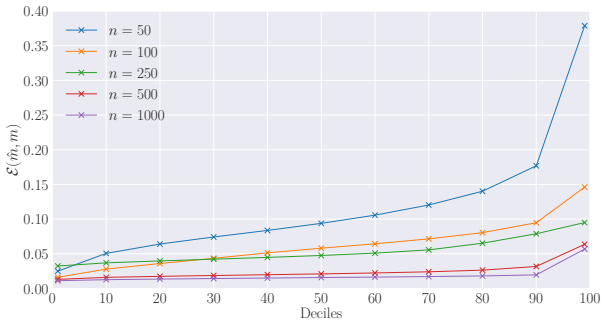


(b) AIC

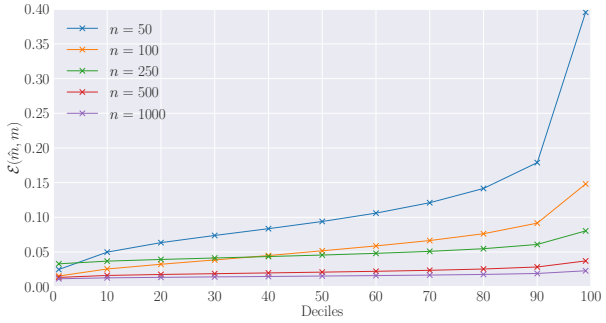


(c) LOOCV

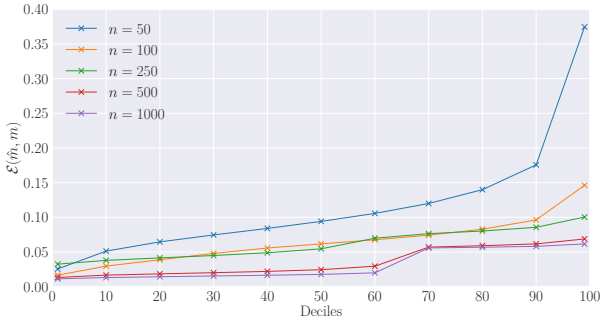
Figure A.1: Deciles of $\mathcal{E}(m^{CV}, m_*)$ in Design 1.



(a) 5FCV

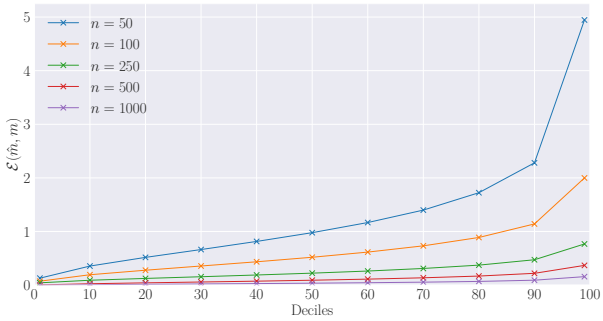


(b) AIC

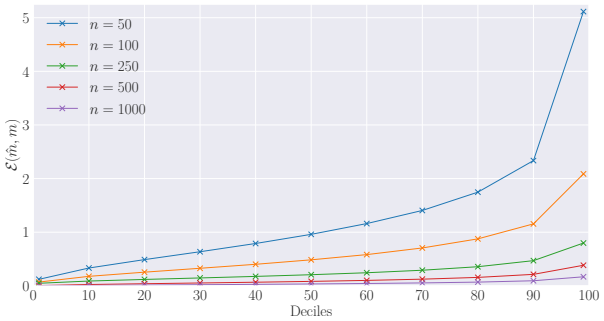


(c) LOOCV

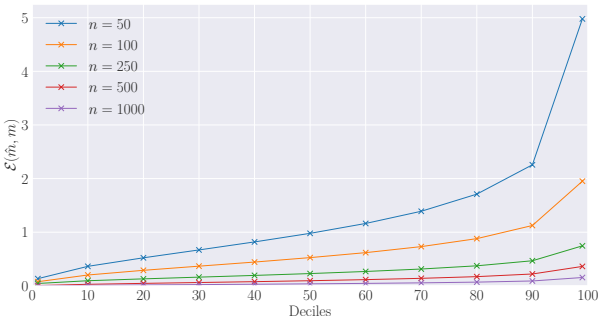
Figure A.2: Deciles of $\mathcal{E}(m^{CV}, m_*)$ in Design 2.1.



(a) 5FCV

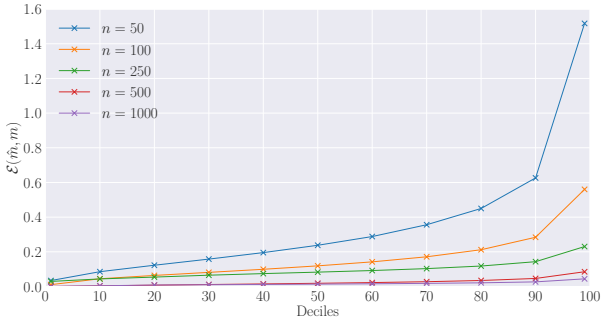


(b) AIC

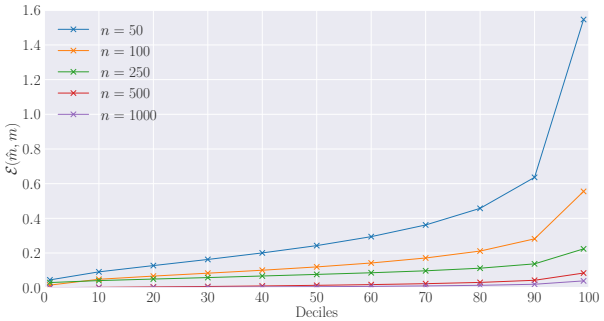


(c) LOOCV

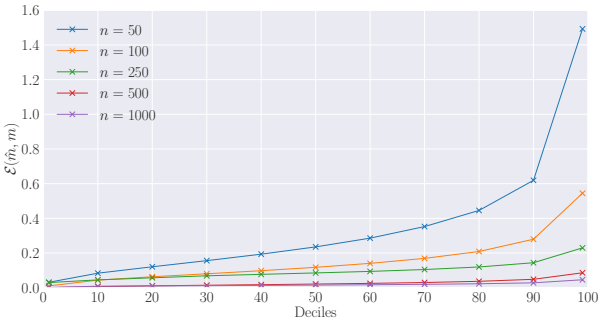
Figure A.3: Deciles of $\mathcal{E}(m^{CV}, m_*)$ in Design 2.2.



(a) 5FCV



(b) AIC



(c) LOOCV

Figure A.4: Deciles of $\mathcal{E}(m^{CV}, m_*)$ in Design 2.3.

B Appendix of Chapter 3

B.1. Proofs

The proofs of the main results in the text rely on several intermediate lemmas which are stated and derived in Section B.1.1.

Proof of Theorem 3.1. First, observe that

$$R_{1,n}(\alpha_m(x)) = \alpha_m(x)'(\hat{\Sigma}_m^{-1} - \Sigma_m)\mathbb{G}_n Z_{i,m}(\varepsilon_i + r_{i,m}).$$

By the triangle inequality

$$\sup_{(x,m)\in\mathcal{I}_n} |R_{1,n}(\alpha_m(x))| \leq \epsilon_1 + \epsilon_2,$$

where

$$\epsilon_1 := \sup_{(x,m)\in\mathcal{I}_n} |\alpha_m(x)'(\hat{\Sigma}_m^{-1} - \Sigma_m)\mathbb{G}_n Z_{i,m}\varepsilon_i|,$$

and

$$\epsilon_2 := \sup_{(x,m)\in\mathcal{I}_n} |\alpha_m(x)'(\hat{\Sigma}_m^{-1} - \Sigma_m)\mathbb{G}_n Z_{i,m} r_{i,m}|.$$

From Lemma B.1,

$$\epsilon_1 \lesssim_P \nu_n \sqrt{\frac{\xi_{\overline{m}}^2(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}}$$

$$\epsilon_2 \lesssim_P b_{\tilde{m}} \sqrt{\frac{\xi_{\overline{m}}^2 \log \overline{m}(\tilde{m}^2 \log \tilde{m})}{n}},$$

which yields the bound on $R_{1,n}(\alpha_m(x))$. Defining

$$\varepsilon_3 := \sup_{(x,m)\in\mathcal{I}_n} |R_{2,n}(\alpha_m(x))|,$$

such that by Lemma B.1

$$\varepsilon_3 \lesssim_P b_{\tilde{m}} \sqrt{\overline{m} \log \overline{m} \log \tilde{m}},$$

which immediately establishes the second claim. ■

Proof of Theorem 3.2. Both (3.3) and (3.4) are a direct consequence of Theorem 3.1, the final bound in Lemma B.1 and Condition 3.2. ■

Proof of Theorem 3.3. It follows directly from approximation theory that $\xi_{\overline{m}} \lesssim \overline{m}$ and $\xi_{\overline{m}} \lesssim \sqrt{\overline{m}}$ by Condition 3.1 for tensor products of polynomials and B-Splines respectively. Furthermore, it also holds that $b_{\tilde{m}} \lesssim \underline{m}^{1-p/d}$ for polynomials, and $b_{\tilde{m}} \lesssim \underline{m}^{-(p_0 \wedge p)/d}$ for B-Splines, see Lemma 1 from Belloni, Chernozhukov and Fernandez-Val (2019). Combined with the condition that $(v_n \vee b_{\tilde{m}})\tilde{m}\xi_{\overline{m}} \lesssim \sqrt{n}$ implies that $\overline{r}_{1n} + \overline{r}_{2n} \lesssim \sqrt{\overline{m} \log \overline{m} \log \tilde{m}}$ from Theorem 3.2 which immediately proves the claim for both polynomials and B-Splines. ■

Proof of Corollary 3.3.1. First, observe that given the choices of $\underline{m}$ and $\overline{m}$

$$\log \tilde{m} \lesssim \log n - \log \log n.$$

With the Lipschitz condition as explained in Remark 3.2, the rate for B-Splines in Theorem 3.3 simplifies to

$$\sup_{(x,m)\in\mathcal{I}_n} |\hat{\theta}_m(x) - \theta(x)| \lesssim_P \sqrt{\frac{\overline{m} \log \overline{m} \log \tilde{m}}{n}} + \underline{m}^{-(p_0 \wedge p)/d},$$

as the extra $\overline{m}$ drops out. Hence, the result follows upon realising that $(\nu_n \vee b_{\tilde{m}})\tilde{m}\xi_{\overline{m}} \lesssim \sqrt{n}$, which implies the growth rate of $\bar{r}_{1n} + \bar{r}_{2n}$, holds by Lemma B.2 if

$$\frac{2d}{2p+d} < 1 - \frac{2}{s} \quad \text{and} \quad \frac{2d-2p}{2p+d} < 1.$$

This holds if $p/d > 1/2$ and

$$s > 2 + \frac{2}{p/d - 0.5}.$$

■

Proof of Theorem 3.4. To bound the first quantity, note that by the triangle inequality

$$\max_{m \in \mathcal{M}_n} \|\hat{\Xi}_m - \Xi_m\| \leq \delta_1 + \delta_2,$$

with

$$\delta_1 := \max_{m \in \mathcal{M}_n} \|\mathbb{E}_n \hat{\epsilon}_{i,m}^2 Z_{i,m} Z_{i,m}' - (\varepsilon_i + r_{i,m})^2 Z_{i,m} Z_{i,m}'\|,$$

and

$$\delta_2 := \max_{m \in \mathcal{M}_n} \|\mathbb{E}_n (\varepsilon_i + r_{i,m})^2 Z_{i,m} Z_{i,m}' - \mathrm{E}(\varepsilon_i + r_{i,m})^2 Z_{i,m} Z_{i,m}'\|.$$

The first quantity can be bounded as follows

$$\begin{aligned}
\delta_1 &\leq \max_{m \in \mathcal{M}_n} \|\mathbb{E}_n (Z_{i,m}'(\hat{\beta}_m - \beta_m))^2 Z_{i,m} Z_{i,m}'\| + 2 \max_{m \in \mathcal{M}_n} \|\mathbb{E}_n (\varepsilon_i + r_{i,m}) Z_{i,m} Z_{i,m}'\| \\
&\leq \max_{m \in \mathcal{M}_n} \max_{1 \leq i \leq n} |Z_{i,m}'(\hat{\beta}_m - \beta_m)|^2 \|\hat{\Sigma}_m\| + 2 \max_{m \in \mathcal{M}_n} (\max_{1 \leq i \leq n} |\varepsilon_i| + \max_{1 \leq i \leq n} |r_{i,m}|) \|\hat{\Sigma}_m\| \\
&\lesssim_P \frac{\xi_{\overline{m}}^2}{n} (\sqrt{\overline{m} \log \overline{m} \log \tilde{m}} + \bar{r}_{1n} + \bar{r}_{2n})^2 + (\nu_n + b_{\tilde{m}}) \frac{\xi_{\overline{m}}}{\sqrt{n}} (\sqrt{\overline{m} \log \overline{m} \log \tilde{m}} + \bar{r}_{1n} + \bar{r}_{2n}) \\
&\lesssim (\nu_n + b_{\tilde{m}}) \sqrt{\frac{\xi_{\overline{m}}^2 \overline{m} \log \overline{m} \log \tilde{m}}{n}},
\end{aligned}$$

where the first inequality is due to the fact that $(a+b)^2 - b^2 = a^2 - 2ab$ and $\hat{\varepsilon}_{i,m} = \varepsilon_i + r_{i,m} - Z'_{i,m}(\hat{\beta}_m - \beta_m)$, the third inequality follows from Theorem 3.2 and Conditions 3.3 to 3.5 and the fact that $(\nu_n \vee b_{\tilde{m}})\tilde{m}\xi_{\overline{m}} \lesssim \sqrt{n}$ implies $\bar{r}_{1n} + \bar{r}_{2n} \lesssim \overline{m}\log\overline{m}\log\tilde{m}$. Secondly, by Theorem D.1, a union bound and Assumptions

$$\begin{aligned}\delta_2 &\lesssim_P (\mathrm{E}\max_{1\le i\le n}|\varepsilon_i|^2 + \max_{m\in\mathcal{M}_n}\max_{1\le i\le n}|r_{i,m}|^2)^{1/2}\tilde{m}\sqrt{\frac{\xi^2_{\overline{m}}\log\overline{m}}{n}} \\ &\lesssim (\nu_n + b_{\tilde{m}})\tilde{m}\sqrt{\frac{\xi^2_{\overline{m}}\log\overline{m}}{n}}.\end{aligned}$$

Hence,

$$\max_{m\in\mathcal{M}_n}\|\hat{\Xi}_m - \Xi_m\| \lesssim_P (\nu_n + b_{\tilde{m}})\sqrt{\frac{\xi^2_{\overline{m}}(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}}.$$

For the estimation of $\hat{\Omega}_m$,

$$\begin{aligned}\max_{m\in\mathcal{M}_n}\|\hat{\Omega}_m - \Omega_m\| &\le \max_{m\in\mathcal{M}_n}\left\|\hat{\Sigma}_m^{-1}(\hat{\Xi}_m - \Xi_m)\hat{\Sigma}_m^{-1}\right\| + \max_{m\in\mathcal{M}_n}\left\|(\hat{\Sigma}_m^{-1} - \Sigma_m^{-1})\Xi_m\hat{\Sigma}_m^{-1}\right\| \\ &\quad + \max_{m\in\mathcal{M}_n}\left\|\Sigma_m^{-1}\Xi_m(\hat{\Sigma}_m^{-1} - \Sigma_m^{-1})\right\| \\ &\le \max_{m\in\mathcal{M}_n}\left\|\hat{\Sigma}_m^{-1}\right\|^2\left\|\hat{\Xi}_m - \Xi_m\right\| + \max_{m\in\mathcal{M}_n}\left\|\hat{\Sigma}_m^{-1} - \Sigma_m^{-1}\right\|\left\|\Xi_m\right\|\left\|\hat{\Sigma}_m^{-1}\right\| \\ &\quad + \max_{m\in\mathcal{M}_n}\left\|\Sigma_m^{-1}\right\|\left\|\Xi_m\right\|\left\|\hat{\Sigma}_m^{-1} - \Sigma_m^{-1}\right\| \\ &\lesssim_P (\nu_n + b_{\tilde{m}})\sqrt{\frac{\xi^2_{\overline{m}}(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}},\end{aligned}$$

where the final inequality follows since the maximal eigenvalues of $\hat{\Sigma}_m^{-1}$, Σ_m^{-1} and Ξ_m are bounded from above combined with Theorem D.1, Lemma B.3 and (3.14). ■

Proof of Corollary 3.4.1. From the inequality $|x/y - 1| \leq |x^2/y^2 - 1|$ for $x, y > 0$ and the fact that the minimal eigenvalue of Ω_m is bounded from below

$$\begin{aligned}\max_{m\in\mathcal{M}_n}\left|\frac{\hat{\sigma}(\alpha_m)}{\sigma(\alpha_m)} - 1\right| \leq \max_{m\in\mathcal{M}_n}\left|\frac{\hat{\sigma}^2(\alpha_m)}{\sigma^2(\alpha_m)} - 1\right| &\leq \max_{m\in\mathcal{M}_n}\left|\frac{\alpha_m'\hat{\Omega}_m\alpha_m}{\alpha_m'\Omega_m\alpha_m} - 1\right| \\ &\leq \max_{m\in\mathcal{M}_n}\frac{\|\hat{\Omega}_m - \Omega_m\|}{\lambda_{\min}(\Omega_m)} \\ &\lesssim \max_{m\in\mathcal{M}_n}\|\hat{\Omega}_m - \Omega_m\|.\end{aligned}$$

The result then follows from Theorem 3.4. ∎

B.1.1. Additional Technical Results

Lemma B.1. *Assume that Conditions 3.1 to 3.4. For the empirical errors defined in Step 1 of the proof of Theorem 3.1 it holds that*

$$\epsilon_1 \lesssim_P \nu_n\sqrt{\frac{\xi_{\overline{m}}^2(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}}, \tag{B.1}$$

$$\epsilon_2 \lesssim_P b_{\tilde{m}}\sqrt{\frac{\xi_{\overline{m}}^2(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}}, \tag{B.2}$$

$$\epsilon_3 \lesssim_P b_{\tilde{m}}\sqrt{\overline{m}\log\overline{m}\log\tilde{m}}, \tag{B.3}$$

and

$$\epsilon_4 \lesssim_P \sqrt{\overline{m}\log\overline{m}\log\tilde{m}}. \tag{B.4}$$

Proof. **Step 1:** (Bounds on empirical errors)

First, I prove (B.1). Recall the class of functions from Lemma B.5 and

$$\mathcal{V} = \bigcup_{m\in\mathcal{M}_n}\mathcal{V}_m,$$

with envelope $V := \max_{1\le i\le n} |\varepsilon_i| \max_{m\in\mathcal{M}_n} \|\hat{\Sigma}_m^{-1} - \Sigma_m\| \|\hat{\Sigma}_m\|^{1/2}$ such that

$$\epsilon_1 \equiv \sup_{v\in\mathcal{V}} |v|.$$

Consider the symmetrised process v^o which, conditional on the data, satisfies by Theorem D.2

$$\begin{aligned}
\mathrm{E}[\|v^o\|_{\mathcal{V}}|Z,\epsilon] &\lesssim \int_0^V \sqrt{\log N(\mathcal{V}, \|\cdot\|_{2,n}, \tau)}\,\mathrm{d}\tau \\
&\le V \int_0^1 \sqrt{\log N(\mathcal{V}, \|\cdot\|_{2,n}, \tau V)}\,\mathrm{d}\tau \\
&\lesssim V \int_0^1 \sqrt{\overline{m}\log(\tilde{m}/\tau)}\,\mathrm{d}\tau,
\end{aligned}$$

where the first inequality holds as the envelope V does not depend on i such that $\|V\|_{2,n} = V$, the second inequality by a change-of-variables argument and the third inequality by Lemma B.4 and Lemma B.5. Hence,

$$\mathrm{E}[\|v^o\|_{\mathcal{V}}|Z,\epsilon] \le V\sqrt{\overline{m}\log\tilde{m}}. \tag{B.5}$$

The symmetrisation Lemma 2.3.1 in van der Vaart and Wellner (1996), the bound on V from Step 2 and (B.5) proves (B.1)

$$\|v\|_{\mathcal{V}} \le 2\,\mathrm{E}[\|v^o\|_{\mathcal{V}}|Z,\varepsilon] \lesssim \sqrt{\overline{m}\log\tilde{m}}V \lesssim_P \nu_n\sqrt{\frac{\xi_{\overline{m}}^2(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}}. \tag{B.6}$$

To prove (B.2), we have

$$\begin{aligned}
\epsilon_2 &\lesssim \max_{m\in\mathcal{M}_n} \left\|(\hat{\Sigma}_m^{-1} - \Sigma_m)\mathbb{G}_n Z_{i,m} r_{i,m}\right\| \\
&\le \max_{m\in\mathcal{M}_n} \left\|\hat{\Sigma}_m^{-1} - \Sigma_m\right\| \max_{m\in\mathcal{M}_n} \left\|\mathbb{G}_n Z_{i,m} r_{i,m}\right\|.
\end{aligned}$$

Combining this with (B.10) and (B.13) in Step 2 establishes

$$\epsilon_2 \lesssim_P b_{\tilde{m}} \sqrt{\frac{\xi_{\overline{m}}^2 \log \overline{m} (\tilde{m}^2 \log \tilde{m})}{n}}.$$

For the error ϵ_3, recall the class of functions from Lemma B.7 such that

$$\mathcal{K} = \bigcup_{m \in \mathcal{M}_n} \mathcal{K}_m,$$

yields

$$\epsilon_3 \equiv \sup_{k \in \mathcal{K}} |\mathbb{G}_n k|.$$

This class has envelope $K = b_{\tilde{m}} \xi_{\overline{m}}$ and its variance is bounded by $b_{\tilde{m}}^2$. The quantity of interest can be bounded by applying Corollary 3.5.8 from Giné and Nickl (2016) together with Lemmas B.4 and B.7 such that

$$\mathrm{E} \|\mathbb{G}_n k\|_{\mathcal{K}} \lesssim b_{\tilde{m}} \sqrt{\overline{m} \log(\xi_{\overline{m}} \tilde{m}^{1/\overline{m}})} + \frac{b_{\tilde{m}} \overline{m} \log(\xi_{\overline{m}} \tilde{m}^{1/\overline{m}})}{\sqrt{n}}. \tag{B.7}$$

Using the assumption that $\log \xi_{\overline{m}} \lesssim \log \overline{m}$ results in

$$\mathrm{E} \|\mathbb{G}_n k\|_{\mathcal{K}} \lesssim b_{\tilde{m}} \sqrt{\overline{m} \log \overline{m} + \log \tilde{m}} + \frac{b_{\tilde{m}} (\overline{m} \log \overline{m} + \log \tilde{m})}{\sqrt{n}}. \tag{B.8}$$

Similar to the calculations in Step 2

$$\begin{aligned} \frac{\overline{m} \log \overline{m} + \log \tilde{m}}{\sqrt{n}} &= \sqrt{\overline{m} \log \overline{m} + \log \tilde{m}} \sqrt{\frac{\overline{m} \log \overline{m} + \log \tilde{m}}{n}} \\ &\lesssim o\Big(\sqrt{\overline{m} \log \overline{m} + \log \tilde{m}}\Big). \end{aligned}$$

This shows that for sufficiently large n

$$\mathrm{E} \|\mathbb{G}_n k\|_{\mathcal{K}} \lesssim b_{\tilde{m}} \sqrt{\overline{m} \log \overline{m} + \log \tilde{m}}.$$

This together with $\overline{m} \log \overline{m} + \log \tilde{m} \lesssim \overline{m} \log \overline{m} \log \tilde{m}$, (B.8) and Markov's inequality establishes (B.3).

The final bound (B.4) follows completely analogously to the previous step, but now applied to the class of function in Lemma B.6 which completes the proof.

Step 2: (Auxiliary calculations)

To bound the envelope V from Step 1, first note that by Lemma B.2

$$\max_{1\leq i\leq n} |\varepsilon_i| \lesssim_P n^{1/s}, \tag{B.9}$$

and by Theorem D.1, a union bound and Markov's inequality

$$\max_{m\in\mathcal{M}_n} \|\hat{\Sigma}_m^{-1} - \Sigma_m\| \lesssim_P \tilde{m}\sqrt{\frac{\xi_{\overline{m}}^2 \log \overline{m}}{n}}. \tag{B.10}$$

And similarly,

$$\max_{m\in\mathcal{M}_n} \|\hat{\Sigma}_m\|^{1/2} \leq \sqrt{\max_{m\in\mathcal{M}_n} \|\hat{\Sigma}_m - \Sigma_m\| + \|\Sigma_m\|} \lesssim_P 1. \tag{B.11}$$

Combining (B.9), (B.10) and (B.11) yields

$$V \lesssim_P \nu_n \tilde{m}\sqrt{\frac{\xi_{\overline{m}}^2 \log \overline{m}}{n}}.$$

Next, I bound

$$\max_{m\in\mathcal{M}_n} \left\|\mathbb{G}_n Z_{i,m} r_{i,m}\right\|, \tag{B.12}$$

which is used in the bound for ϵ_2, by considering the class of functions

$$\mathcal{H} = \{(Z_m, r_m) \mapsto Z_m \cdot r_m : m \in \mathcal{M}_n\}.$$

For any $h \in \mathcal{H}$

$$\|h\|_\infty \leq b_m \xi_m,$$

and since $\|Z_{i,m}\|_\infty \leq \|Z_{i,m}\|_2$

$$\|h\|_{2,P}^2 \lesssim b_m^2 \xi_m.$$

Then, by Lemma 19.33 in van der Vaart (1998)

$$\begin{aligned} \mathrm{E}\,\|\mathbb{G}_n h\|_{\mathcal{H}} &\lesssim b_{\tilde{m}}\sqrt{\overline{m}\log\tilde{m}} + \frac{b_{\tilde{m}}\xi_{\overline{m}}\log\tilde{m}}{\sqrt{n}} \\ &\lesssim b_{\tilde{m}}\sqrt{\overline{m}\log\tilde{m}}, \end{aligned}$$

since $\frac{\xi_{\overline{m}}^2}{\overline{m}}\log\tilde{m}/\sqrt{n} \leq \sqrt{\overline{m}\log\tilde{m}}\sqrt{\frac{\xi_{\overline{m}}^2}{\overline{m}}\log\tilde{m}/n} \lesssim o(\sqrt{\overline{m}\log\tilde{m}})$. By Markov's inequality (B.12) is bounded by

$$\max_{m\in\mathcal{M}_n} \left\|\mathbb{G}_n Z_{i,m} r_{i,m}\right\| \lesssim_P b_{\tilde{m}}\sqrt{\overline{m}\log\tilde{m}}. \tag{B.13}$$

■

Lemma B.2. *Assume that Condition 3.4(iii) holds. Then,*

$$v_n := \sqrt{\mathrm{E}\left[\max_{1\leq i\leq n}\varepsilon_i^2\right]} \lesssim n^{1/s}.$$

Proof. Firstly, by Condition 3.4(iii) and Lemma 2.2.2 from van der Vaart and Wellner (1996)

$$\mathrm{E}\left[\max_{1\leq i\leq n}\varepsilon_i^s|X\right] \lesssim n\max_{1\leq i\leq n}\mathrm{E}[\varepsilon_i^s|X] \quad \text{a.s.}$$

Hence, by Hölder's inequality

$$\sqrt{\mathrm{E}\left[\max_{1\leq i\leq n}\varepsilon_i^2\right]} \leq \left(\mathrm{E}\left[\max_{1\leq i\leq n}\varepsilon_i^s\right]\right)^{1/s} \lesssim n^{1/s}.$$

■

Lemma B.3. *Let A and B be two positive definite $m \times m$ matrices with $\lambda_{\min}(A) > c_A$ and $\lambda_{\min}(B) > c_B$ for $c_A, c_B > 0$ and*

$$\left\|A - B\right\| \leq \delta. \tag{B.14}$$

Then,

$$\left\|A^{-1} - B^{-1}\right\| \leq C\delta, \tag{B.15}$$

for some constant $C > 0$.

Proof. Notice that

$$A^{-1} - B^{-1} = -A^{-1}(A - B)B^{-1}.$$

Using this,

$$\left\|A^{-1} - B^{-1}\right\| \leq \left\|A^{-1}\right\| \left\|A - B\right\| \left\|B^{-1}\right\| \leq \frac{\left\|A - B\right\|}{\lambda_{\min}(A)\lambda_{\min}(B)} \leq C\delta,$$

where the conclusion holds by setting $C = c_A^{-1} c_B^{-1}$. ∎

Lemma B.4. *Let $\mathcal{F}_1, \ldots, \mathcal{F}_m$ be a collection of classes of functions with envelopes F_i such that $\|F_i\|_{2,P} < \infty$ for each i. Let $\mathcal{F} := \cup_{i=1}^m \mathcal{F}_i$. Then,*

$$\sup_P N(\mathcal{F}, \|\cdot\|_{2,P}, \tau\|F\|_{2,P}) \leq m \max_{1 \leq i \leq m} \sup_P N(\mathcal{F}_i, \|\cdot\|_{2,P}, \tau\|F_i\|_{2,P}), \tag{B.16}$$

for $F = \max_{1 \leq i \leq m} F_i$ and $0 < \tau \leq 1$.

Proof. The covering number of $\mathcal{F}$ with balls of radius δ is bounded from above by the sum of the individual covering numbers of the classes of functions in $\mathcal{F}$. The same holds for any $\delta_i \leq \delta$ as covering numbers are non-decreasing in δ, i.e.

$$N(\mathcal{F}, \|\cdot\|_{2,P}, \delta) \leq \sum_{i=1}^m N(\mathcal{F}_i, \|\cdot\|_{2,P}, \delta_i),$$

such that

$$N(\mathcal{F}, \|\cdot\|_{2,P}, \delta) \leq m \max_{1\leq i\leq m} N(\mathcal{F}_i, \|\cdot\|_{2,P}, \delta_i).$$

The conclusion of the lemma follows by defining $\delta = \tau\|F\|_{2,P}$, $\delta_i = \tau\|F_i\|_{2,P}$ and taking the supremum over all discrete probability measures P. ∎

Lemma B.5. *For $v_{i,m}(\alpha) = \alpha'(\hat{\Sigma}_m^{-1} - \Sigma_m)Z_i\varepsilon_i$, consider the class of functions*

$$\mathcal{V}_m = \{(v_1, \ldots, v_n) \in \mathbf{R}^n : v_{i,m}(\alpha), \alpha \in \mathbb{S}^{m-1}\}.$$

Under the empirical L^2-norm, this class has envelope

$$V_m := \max_{1\leq i\leq n} |\varepsilon_i|\|\hat{\Sigma}_m^{-1} - \Sigma_m\|\|\hat{\Sigma}_m^{1/2}\|,$$

and

$$\log N(\mathcal{V}_m, \|\cdot\|_{2,n}, \tau\|V_m\|_{2,n}) \lesssim m\log(1/\tau).$$

Proof. The envelope follows immediately from the calculation below

$$\begin{aligned}\sqrt{\mathbb{E}_n[\alpha'(\hat{\Sigma}_m^{-1} - \Sigma_m)Z_i\varepsilon_i]^2} &\leq \max_{1\leq i\leq n} |\varepsilon_i|\sqrt{\mathbb{E}_n\alpha'(\hat{\Sigma}_m^{-1} - \Sigma_m)Z_iZ_i'(\hat{\Sigma}_m^{-1} - \Sigma_m)\alpha'} \\ &\leq \max_{1\leq i\leq n} |\varepsilon_i|\|\hat{\Sigma}_m^{-1} - \Sigma_m\|\|\hat{\Sigma}_m^{1/2}\| =: V_m,\end{aligned}$$

where the Cauchy-Schwarz inequality delivers the final inequality. Similarly, I can bound the diameter of $\mathcal{V}_m$ by

$$\sqrt{\mathbb{E}_n[\alpha'(\hat{\Sigma}_m^{-1} - \Sigma_m)Z_i\varepsilon_i - \check{\alpha}'(\hat{\Sigma}_m^{-1} - \Sigma_m)Z_i\varepsilon_i]^2} \leq \max_{1\leq i\leq n} |\varepsilon_i|\|\hat{\Sigma}_m^{-1} - \Sigma_m\|\|\hat{\Sigma}_m^{1/2}\|\|\alpha - \check{\alpha}\|,$$

again by the Cauchy-Schwarz inequality. The bound on the entropy follows from the fact that balls in $\mathbf{R}^m$ can be covered by $(C/\tau)^m$ balls of radius τ for some $C > 0$ and Example 19.7 in van der Vaart (1998, p. 271) ∎

Lemma B.6. *Define the class of functions*

$$\mathcal{F}_m = \{(Z, \varepsilon) \to \alpha' Z\varepsilon, \alpha \in \mathbb{S}^{m-1}\},$$

mapping $\mathbf{R}^m \times \mathbf{R}$ *into* $\mathbf{R}$ *with envelope* $F_m := F_m(Z, \varepsilon) = \xi_m|\varepsilon|$. *The uniform entropy numbers of* $\mathcal{F}_m$ *satisfy*

$$\sup_P \log N(\mathcal{F}_m, \|\cdot\|_{2,P}, \tau\|F_m\|_{2,P}) \lesssim m \log(2/\tau) \quad \text{for } 0 < \tau \leq 1.$$

Proof. For any $\alpha, \tilde{\alpha} \in \mathbb{S}^{m-1}$, we have

$$|\alpha' Z\varepsilon - \tilde{\alpha}' Z\varepsilon| \leq F_m \|\alpha - \tilde{\alpha}\|,$$

by the Cauchy-Schwarz inequality. Then, by example 19.7 from van der Vaart (1998) for $0 < \tau \leq 1$

$$N(\mathcal{F}_m, \|\cdot\|_{2,P}, \tau\|F_m\|_{2,P}) \lesssim \left(\frac{2}{\tau}\right)^m,$$

thus establishing the lemma. ∎

Lemma B.7. *Define the class of functions*

$$\mathcal{H}_m = \{(Z, r) \mapsto \alpha' Z r, \alpha \in \mathbb{S}^{m-1}\},$$

mapping $\mathbf{R}_m \times \mathbf{R}$ *into* $\mathbf{R}$ *with envelope* $H_m := H_m(Z, r) = b_m \xi_m$. *The uniform entropy numbers of* $\mathcal{H}_m$ *satisfy*

$$\sup_P \log N(\mathcal{H}_m, \|\cdot\|_{2,P}, \tau H_m) \lesssim m \log(2/\tau) \quad \text{for } 0 < \tau \leq 1.$$

Proof. For any $\alpha, \tilde{\alpha} \in \mathbb{S}^{m-1}$, we have

$$\begin{aligned} |\alpha' Z r - \tilde{\alpha}' Z r| &\leq |r| \|Z\| \|\alpha - \tilde{\alpha}\| \\ &\leq b_m \xi_m \|\alpha - \tilde{\alpha}\|, \end{aligned}$$

where the first inequality follows by the Cauchy-Schwarz inequality and the second by the assumptions on Z and r. Then, similarly to the proof of Lemma B.6

$$N(\mathcal{H}_m, \|\cdot\|_{2,P}, \tau H_m) \lesssim \left(\frac{2}{\tau}\right)^m.$$

■

C Appendix of Chapter 4

C.1. Proofs

This section contains the proofs of the results from Chapter 4. Additional technical results needed in the proofs of the main results are stated and derived in Section C.1.1. The proofs heavily rely on the results and techniques used in Section 3.3.

For the proof of Theorem 4.1 define the following processes:

$$\tilde{T}_n := \sup_{(x,m)\in\mathcal{I}_n} \left| \sqrt{n} \frac{\alpha_m(x)'(\hat{\beta}_m - \beta)}{\sigma_m} \right|, \tag{C.1}$$

$$\tilde{T}'_n := \sup_{(x,m)\in\mathcal{I}_n} \left| \frac{\alpha_m(x)'\mathbb{G}_n Z_{i,m}\varepsilon_i}{\sigma_m} \right|. \tag{C.2}$$

Proof of Theorem 4.1. By the triangle inequality,

$$\left|T_n - T_n^*\right| \leq \left|T_n - \tilde{T}_n\right| + \left|\tilde{T}_n - \tilde{T}'_n\right| + \left|\tilde{T}'_n - T_n^*\right|. \tag{C.3}$$

The first term on the RHS can be bounded by Theorem 3.2, Corollary 3.4.1, $\bar{r}_{1n} + \bar{r}_{2n} \lesssim \delta_n^{-1}$ and $\max_{m\in\mathcal{M}_n} \hat{\sigma}_m/\sigma_m \lesssim_P 1$

$$\begin{aligned}
\left|T_n - \tilde{T}_n\right| &\leq \sup_{(x,m)\in\mathcal{I}_n} \left| \sqrt{n} \frac{\alpha_m(x)'(\hat{\beta}_m - \beta)}{\hat{\sigma}_m} - \sqrt{n} \frac{\alpha_m(x)'(\hat{\beta}_m - \beta)}{\sigma_m} \right| \\
&= \sup_{(x,m)\in\mathcal{I}_n} \sqrt{n} \left|\alpha_m(x)'(\hat{\beta}_m - \beta)\right| \left| \frac{1}{\hat{\sigma}_m} - \frac{1}{\sigma_m} \right| \\
&= \sup_{(x,m)\in\mathcal{I}_n} \sqrt{n} \left|\alpha_m(x)'(\hat{\beta}_m - \beta)\right| \left| \frac{1}{\hat{\sigma}_m} - \frac{1}{\sigma_m} \right|
\end{aligned}$$

$$\lesssim_P \sup_{(x,m)\in\mathcal{I}_n} \sqrt{n}\left|\alpha_m(x)'(\hat{\beta}_m - \beta)\right|\left|\frac{\hat{\sigma}_m}{\sigma_m} - 1\right|$$

$$\lesssim_P \delta_n^{-1}(\nu_n + b_{\tilde{m}})\sqrt{\frac{\xi_{\overline{m}}^2(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}}. \tag{C.4}$$

The bound on the second quantity follows directly from Theorem 3.1 and Condition 3.6(iii)

$$\left|\tilde{T}_n - \tilde{T}_n'\right| \lesssim_P \delta_n^{-1}. \tag{C.5}$$

Then, the final term on the RHS of (C.3) can be bounded using Corollary 2.2 from Chernozhukov, Chetverikov and Kato (2014b). Recall the class of functions from Lemma B.6 with α_m replaced by $\rho_m = \alpha_m/\sigma_m$

$$\mathcal{F}_{m,n} = \left\{(Z,\varepsilon) \to \rho' T\varepsilon, \rho \in \mathbb{S}^{m-1}\right\},$$

such that for

$$\mathcal{F}_n = \bigcup_{m\in\mathcal{M}_n} \mathcal{F}_{m,n},$$

it follows that

$$\tilde{T}_n' \equiv \sup_{f\in\mathcal{F}_n} |\mathbb{G}_n f|.$$

By Lemma B.4 and Lemma B.6, this class is a VC-class of functions such that for any *fixed* n it is pre-Gaussian. Hence, for any $n \geq 1$ there exists a Gaussian random variable, $G_n \in \ell^\infty(\mathcal{F}_n)$, with mean zero and covariance function

$$\mathrm{E}[G_n(f)G_n(f')] = \mathrm{cov}[f(X_1,\varepsilon_1), f'(X_1,\varepsilon_1)] \quad \text{for } f, f' \in \mathcal{F},$$

such that $B_n = G_n f$ is the desired process. Next, let $C_4 = \sup_{x\in\mathcal{X}} \mathrm{E}[\varepsilon_1^4|X = x]$ with the auxiliary calculations

$$\mathrm{E}[(\rho_m' Z_m \varepsilon)^2] \leq \sqrt{C_4} \lesssim C_4,$$

$$\mathrm{E}[|\rho_m' Z_m \varepsilon|^3] \leq (1 + C_4)\xi_{\overline{m}}(\rho_m' \mathrm{E}[Z_m Z_m']\rho_m)$$

$$
\begin{aligned}
&= (1 + C_4)\xi_{\overline{m}} \\
\mathrm{E}[(\rho_m' Z_m \varepsilon)^4] &\leq C_4 \, \mathrm{E}[(\rho_m' Z_m)^4] \\
&\leq C_4 \xi_{\overline{m}}^2,
\end{aligned}
$$

where the third inequality follows from $|x|^3 \leq 1 + x^4$. Then, with $b_n = \xi_{\overline{m}}$ and $\sigma^2 = C_4$ with $\ell_n = \overline{m} \log \overline{m} \log \tilde{m}$ by Corollary 2.2 from Chernozhukov, Chetverikov and Kato (2014b)

$$
\left| \tilde{T}_n' - T_n^* \right| \lesssim_P n^{-1/6} \xi_{\overline{m}}^{1/3} \ell_n + n^{-1/4} \xi_{\overline{m}}^{1/2} \ell_n^{5/4} + n^{-1/2+1/q} \xi_{\overline{m}} \ell_n^{3/2}. \tag{C.6}
$$

Combining (C.4), (C.5) and (C.6) with the condition

$$
\frac{\xi_{\overline{m}}^2 (\overline{m} \log \overline{m} \log \tilde{m})^\gamma}{n^{1-1/s}} \to 0,
$$

concludes the proof. ■

Proof of Corollary 4.1.1. This follows directly from Lemma 2.3 in Chernozhukov, Chetverikov and Kato (2014b), the final bound in Lemma B.1 and Theorem 4.1. ■

Proof of Theorem 4.2. By Corollary 4.1.1 it immediately follows that

$$
P\Big(T_n \leq c_n^*(\alpha)\Big) \leq P\Big(T_n^* \leq c_n^*(\alpha)\Big) + o(1) = 1 - \alpha + o(1),
$$

where the equality is by the definition of c_n^*. The opposite direction holds by the same reasoning. Finally, (4.16) is a direct consequence of these two bounds. ■

Proof of Theorem 4.3. By Theorem 3.4 and Corollary 3.4.1 it follows by the same arguments as used in those proofs that for V and $\hat{V}$ defined in (4.6) and (4.13)

$$
\|\hat{V} - V\| \lesssim \delta_n,
$$

with

$$\delta_n = (\nu_n + b_{\tilde{m}})\sqrt{\frac{\xi^2_{\overline{m}}(\overline{m}\log\overline{m})(\tilde{m}^2\log\tilde{m})}{n}}.$$

Together with the assumption that $\overline{r}_{1n} + \overline{r}_{2n} = o_P(1/\sqrt{\overline{m}\log\overline{m}\log\tilde{m}})$, it follows that $\delta_n\sqrt{\overline{m}\log\tilde{m}} = o_P(1/\sqrt{\overline{m}\log\overline{m}\log\tilde{m}})$. This of course, immediately implies that

$$\delta_n\sqrt{\overline{m}\log\tilde{m}} = o_P\left(\frac{1}{\sqrt{\overline{m}\log\tilde{m}}}\right).$$

Using this and Lemma C.3 it holds that

$$\mathrm{E}\left|\tilde{T}^*_n - T^*_n\right| \leq o_P\left(\frac{1}{\sqrt{\overline{m}\log\tilde{m}}}\right).$$

Hence, Markov's inequality together with $\tau_n = \gamma_m$ for γ_m from Lemma C.3 which delivers the result. ■

Proof of Theorem 4.4. For shorthand, let $\kappa_n = \tau_n/\sqrt{\overline{m}\log\tilde{m}}$ be the sequence from Theorem 4.3. By the same Theorem for some sequence $\eta_n = o(1)$

$$P\left(|\tilde{T}^*_n - T^*_n| \geq \kappa_n\right) \leq \eta_n. \tag{C.7}$$

Hence,

$$P(\tilde{c}^*_n(\alpha) > c^*_n(\alpha - \eta_n) + \kappa_n) = o(1), \tag{C.8}$$

and

$$P(\tilde{c}^*_n(\alpha) < c^*_n(\alpha + \eta_n) - \kappa_n) = o(1). \tag{C.9}$$

Secondly, by Lemma C.1 and the anti-concentration theorem, Theorem D.3, due to Chernozhukov, Chetverikov and Kato (2014a, Corollary 2.1)

$$P_{\kappa_n}\left(T^*_n\right) := \sup_{t\in\mathbf{R}} P\left(\left|T^*_n - t\right| \leq \kappa_n\right) = o(1). \tag{C.10}$$

Then, as in the proof of Theorem 4.2 by Corollary 4.1.1

$$\begin{aligned} P\Big(T_n \leq \tilde{c}_n^*(\alpha)\Big) &\leq P\Big(T_n^* \leq \tilde{c}_n^*(\alpha)\Big) + o(1) \\ &\leq P\Big(T_n^* \leq c_n^*(\alpha - \eta_n) + \kappa_n\Big) + o(1) \\ &\leq P\Big(T_n^* \leq c_n^*(\alpha - \eta_n)\Big) + P\Big(\Big|T_n^* - c_n^*(\alpha - \eta_n)\Big| \leq \kappa_n\Big) + o(1) \\ &\leq P\Big(T_n^* \leq c_n^*(\alpha - \eta_n)\Big) + P_{\kappa_n}\Big(T_n^*\Big) + o(1) \\ &= 1 - \alpha + \eta_n + o(1) \\ &= 1 - \alpha + o(1), \end{aligned}$$

where the second line follows from (C.8) and the final two equalities by the definition of c_n^*, (C.10) and $\eta_n \to 0$ as $n \to \infty$. The opposite direction follows analogously instead of (C.8) using (C.9). Combining these two bounds delivers the result. ■

Proof of Theorem 4.5. The claim follows directly from the same argument in Theorem 4.4 removing the absolute values in the definitions of T_n, T_n^* and $\tilde{T}_n^*$. ■

C.1.1. Additional Technical Results

Lemma C.1. *Let $B_m(r_m)$ be an r_m-ball in $\mathbf{R}^m$ and let $\{X_m \in \mathbf{R}^m : m \in \mathcal{M}_n\}$ be a collection of i.i.d Gaussian random vectors and define the classes of functions*

$$\mathcal{L}_{n,m} = \Big\{X \mapsto \alpha' X : \alpha \in B_m(r_m)\Big\}.$$

mapping $\mathbf{R}^m$ into $\mathbf{R}$. Then, for

$$\mathcal{L}_n = \bigcup_{m \in \mathcal{M}_n} \mathcal{L}_{n,m},$$

it follows that

$$\mathrm{E}\left[\sup_{f \in \mathcal{L}_n} |f|\right] \lesssim \overline{r}\sqrt{\overline{m} \log \tilde{m}}, \tag{C.11}$$

with $\bar{r} = \max_{m \in \mathcal{M}_n} r_m$.

Proof. First note that

$$\sup_{f \in \mathcal{L}_n} |f| = \max_{m \in \mathcal{M}_n} \sup_{\alpha \in B_m(r_m)} |\alpha' X|.$$

By Lemma 2.2.2 in van der Vaart and Wellner (1996),

$$\mathrm{E}\left[\max_{m \in \mathcal{M}_n} \sup_{\alpha \in B_m(r_m)} |\alpha' X| \right] \lesssim \sqrt{\log \tilde{m}} \max_{m \in \mathcal{M}_n} \left\| \sup_{\alpha \in B_m(r_m)} |\alpha' X| \right\|_{\psi_2}. \tag{C.12}$$

For fixed $m \in \mathcal{M}_n$, by Corollary 2.2.5 in van der Vaart and Wellner (1996) given that $\alpha' X$ is a Gaussian process and thus sub-Gaussian

$$\begin{aligned}
\left\| \sup_{\alpha \in B_m(r_m)} |\alpha' X| \right\|_{\psi_2} &\lesssim \int_0^{r_m} \sqrt{\log N(B_m(r_m), \|\cdot\|_{2,P}, \tau)}\, \mathrm{d}\tau \\
&= r_m \int_0^1 \sqrt{\log N(B_m(r_m), \|\cdot\|_{2,P}, \tau r_m)}\, \mathrm{d}\tau \\
&\lesssim r_m \int_0^1 \sqrt{m \log(5/\delta)}\, \mathrm{d}\tau \\
&\lesssim r_m \sqrt{m},
\end{aligned}$$

where $\|\cdot\|_{\psi_2}$ is the ψ_2-Orlicz norm and the upper bound on the entropy of balls in $\mathbf{R}^m$ in the third line follows from Lemma 2.5 in van de Geer (2000). Plugging this bound into (C.12) delivers the result. ∎

Lemma C.2. *Let* $\{X_m \in \mathbf{R}^m : m \in \mathcal{M}_n\}$ *be a collection of Gaussian random vectors and define the processes*

$$\tilde{T}_n := \sup_{(x,m) \in \mathcal{I}_n} \left| \frac{\alpha_m(x)' \hat{\Omega}_m^{1/2}}{\sqrt{n} \hat{\sigma}_m} X_m \right|,$$

and

$$T_n^* := \sup_{(x,m)\in\mathcal{I}_n} \left| \frac{\alpha_m(x)'\Omega_m^{1/2}}{\sqrt{n}\sigma_m} X_m \right|.$$

For a sequence δ_n such that

$$\max_{m\in\mathcal{M}_n} \left\| \hat{\Omega}_m - \Omega_m \right\| + \max_{m\in\mathcal{M}_n} \left| \frac{\sigma_m(x)}{\hat{\sigma}_m(x)} - 1 \right| \lesssim_P \delta_n,$$

it holds that,

$$\mathrm{E}\left[\left| \tilde{T}_n - T_n^* \right| \right] \lesssim \delta_n \sqrt{\overline{m} \log \tilde{m}}. \tag{C.13}$$

Hence, if $\delta_n \sqrt{\overline{m} \log \tilde{m}} = o(1/\sqrt{\overline{m} \log \tilde{m}})$

$$P\left(\left| \tilde{T}_n - T_n^* \right| > \beta_n / \sqrt{\overline{m} \log \tilde{m}} \right) = o(1), \tag{C.14}$$

for some sequence β_n converging to zero slowly enough as $n \to \infty$.

Proof of Lemma C.2. Start from the observation that

$$\left| \tilde{T}_n - T_n^* \right| \leq \sup_{(x,m)\in\mathcal{I}_n} \left| \alpha_m(x)' \left(\frac{\hat{\Omega}_m^{1/2}}{\sqrt{n}\hat{\sigma}_m(x)} - \frac{\Omega_m^{1/2}}{\sqrt{n}\sigma_m(x)} \right) X_m \right|.$$

Proceeding conditional on the data such that

$$\mathcal{T}_n = \left\{ \tilde{\alpha}_{x,m}' X_m : \tilde{\alpha}_{x,m} \in \mathbf{R}^m, m \in \mathcal{M}_n \right\},$$

with

$$\tilde{\alpha}_{x,m} = \alpha_m(x)' \left(\frac{\hat{\Omega}_m^{1/2}}{\sqrt{n}\hat{\sigma}_m(x)} - \frac{\Omega_m^{1/2}}{\sqrt{n}\sigma_m(x)} \right),$$

is a collection of zero-mean Gaussian processes. Next, consider

$$\|\tilde{\alpha}_{x,m}\|_\infty = \sup_{(x,m)\in\mathcal{I}_n} \left\| \alpha_m(x)' \left(\frac{\hat{\Omega}_m^{1/2}}{\sqrt{n}\hat{\sigma}_m(x)} \pm \frac{\Omega_m^{1/2}}{\sqrt{n}\hat{\sigma}_m(x)} - \frac{\Omega_m^{1/2}}{\sqrt{n}\sigma_m(x)} \right) \right\|$$

$$
\begin{aligned}
&\leq \sup_{(x,m)\in\mathcal{I}_n} \frac{\|\alpha_m(x)\|}{\sqrt{n}\hat{\sigma}_m(x)} \left\|\hat{\Omega}_m^{1/2} - \Omega_m^{1/2}\right\| + \sup_{(x,m)\in\mathcal{I}_n} \frac{\left\|\alpha_m(x)\Omega_m^{1/2}\right\|}{\sqrt{n}\sigma_m(x)} \left|\frac{\hat{\sigma}_m(x)}{\sigma_m(x)} - 1\right| \\
&\lesssim \max_{m\in\mathcal{M}_n} \left\|\hat{\Omega}_m^{1/2} - \Omega_m^{1/2}\right\| + \max_{m\in\mathcal{M}_n} \left|\frac{\sigma_m(x)}{\hat{\sigma}_m(x)} - 1\right| \\
&\lesssim \max_{m\in\mathcal{M}_n} \left\|\hat{\Omega}_m - \Omega_m\right\| + \max_{m\in\mathcal{M}_n} \left|\frac{\sigma_m(x)}{\hat{\sigma}_m(x)} - 1\right| \\
&\lesssim_P \delta_n,
\end{aligned}
$$

where the first line is due to the triangle and Cauchy-Schwarz inequality, the second by Condition 3.3, Corollary 3.4.1, the third Lemma A.2 from Belloni et al. (2015) and the final inequality by the assumptions of the Lemma. Applying Lemma C.1 to the class of functions $\mathcal{T}_n$ with $\tilde{\alpha}_{x,m}$ restricted to lie in the ball of radius δ_n up to some constant C large enough establishes (C.13). The second claim in (C.14) is a direct consequence of Markov's inequality applied to (C.13). ■

Lemma C.3. *Let $X_m \in \mathbf{R}^m$ and $Y_m \in \mathbf{R}^m$ be two m-dimensional Gaussian random vectors. Furthermore*

$$X_m \sim \mathcal{N}_m(0, V) \quad \textit{and} \quad Y_m \sim \mathcal{N}_m(0, W),$$

such that if $\|W\| \lesssim 1$ and

$$\|V - W\| \lesssim_P \delta_m, \tag{C.15}$$

then,

$$\|X_m - Y_m\|_\infty \lesssim_P \delta_m \sqrt{m \log m}.$$

Furthermore, if $\delta_m m \log m = o(1)$, then for a γ_m converging to zero slowly enough and for some $\overline{m} \geq m$

$$P\left(\|X_m - Y_m\|_\infty \geq \frac{\gamma_m}{\sqrt{\overline{m} \log \overline{m}}} \right) = o(1).$$

Proof of Lemma C.3. First, note that $X_m = V^{1/2}\mathcal{N}_m$ and $Y_m = W^{1/2}\mathcal{N}_m$ for $\mathcal{N}_m$ a standard Gaussian random vector. Also,

$$\|V^{1/2} - W^{1/2}\|_\infty \leq \sqrt{m}\|V^{1/2} - W^{1/2}\| \lesssim \sqrt{m}\|V - W\|, \tag{C.16}$$

where the final inequality is due to Lemma A.2 from Belloni et al. (2015) and the fact that $\|W\| \lesssim 1$ by assumption. Hence,

$$\begin{aligned}
\|X_m - Y_m\|_\infty &\leq \|V^{1/2}\mathcal{N}_m - W^{1/2}\mathcal{N}_m\|_\infty \\
&\leq \|V^{1/2} - W^{1/2}\|_\infty \|\mathcal{N}_m\|_\infty \\
&\lesssim_P \delta_m \sqrt{m} \|\mathcal{N}_m\|_\infty \\
&\lesssim_P \delta_m \sqrt{m \log m},
\end{aligned}$$

where the penultimate inequality follows from (C.15) and (C.16) and the final inequality from a similar argument as used in the proof of Lemma C.2.

Similar to the final step in the proof of Lemma C.2, the final claim follows from Markov's inequality and the fact that $\delta_m m \log m = o(1)$ implies that

$$\sqrt{\overline{m} \log \overline{m}} \|X_m - Y_m\|_\infty \leq o_P(1),$$

for some $m \leq \overline{m}$ such that

$$P\left(\|X_m - Y_m\|_\infty \geq \frac{\gamma_m}{\sqrt{\overline{m} \log \overline{m}}} \right) = o(1).$$

■

C.2. Extra Monte Carlo Experiment Results

This appendix contains extra material on the Monte Carlo experiment described in Section 4.3.1. Figure C.1 plots the unknown conditional mean function $g : [0, 1] \to \mathbf{R}$

$$g(x) = \arctan\left[\left(2x + \frac{1}{2} \right) \log \left(2x + \frac{1}{2} \right) \right].$$

Table C.2 contains the data on the used to plot Figure 4.2 for the 95 per cent coverage results. Additionally, Tables C.1 and C.3 report the outcome on the coverage at the 90 and 97.5 per cent levels which are similar to the 95 per cent results.

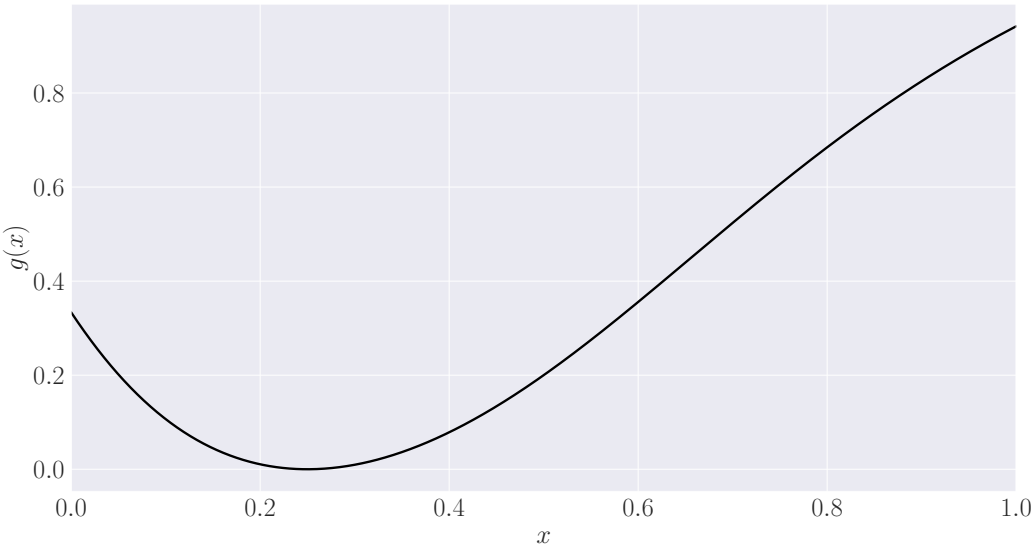


Figure C.1: Plot of $g(x)$ on $[0, 1]$ used in the Monte Carlo experiment.

Table C.1: Simulation results for a significance level α of 90%.

	Max t-stat (Naive)				Max t-stat (Max t-stat)				Cross-Validation (Naive)				Cross-Validation (Max t-stat)			
	n				n				n				n			
x	100	250	500	1000	100	250	500	1000	100	250	500	1000	100	250	500	1000
0.05	81.77	80.87	79.11	75.16	85.20	89.90	91.34	88.72	85.90	86.83	87.37	86.63	88.24	92.22	93.02	92.76
0.10	85.80	82.03	76.12	74.64	88.66	89.00	88.50	91.96	87.25	87.95	87.03	87.76	89.86	92.36	92.14	93.70
0.15	84.76	78.57	74.76	71.59	87.20	87.78	90.02	88.36	87.03	87.05	85.77	86.36	89.24	91.64	92.36	92.66
0.20	82.77	77.16	47.62	28.57	84.60	89.18	91.40	91.10	86.20	86.00	83.06	82.54	87.80	92.32	93.32	93.70
0.25	81.70	78.03	49.90	33.75	83.86	91.10	90.06	88.46	85.70	85.62	82.72	80.78	87.32	93.46	93.02	93.06
0.30	81.63	78.29	70.76	65.31	84.62	91.08	89.00	89.60	85.77	86.10	83.99	82.82	88.02	93.90	92.68	93.14
0.35	82.47	79.17	73.23	68.39	86.76	89.28	90.28	88.76	86.35	86.77	85.07	84.21	89.74	93.04	93.24	92.94
0.40	84.84	78.25	72.23	67.20	88.52	88.82	90.08	89.96	87.93	86.92	86.00	85.06	90.52	92.68	93.54	93.66
0.45	86.65	79.70	73.70	69.31	90.50	89.02	89.90	88.46	88.72	87.51	86.49	85.79	91.50	92.28	93.74	92.92
0.50	88.09	76.03	67.25	58.31	89.80	89.10	89.82	88.14	88.85	87.36	85.95	84.78	90.94	92.28	93.38	92.88
0.55	86.13	78.85	73.05	69.01	87.02	88.78	89.84	89.48	87.93	86.97	86.07	85.63	89.46	92.18	92.84	93.12
0.60	84.25	77.23	59.03	43.19	85.94	88.64	90.26	88.60	87.29	86.31	85.42	84.09	88.82	92.46	92.90	92.52
0.65	82.99	77.94	73.34	67.81	84.58	90.32	88.58	88.86	86.73	85.91	85.47	84.28	88.08	93.14	92.12	92.90
0.70	82.11	78.03	71.77	67.02	84.64	90.36	89.52	88.48	86.02	85.76	84.94	84.46	88.04	92.88	92.98	92.92
0.75	81.95	78.57	37.27	18.03	85.38	89.34	90.96	90.40	85.80	86.09	83.25	82.88	88.68	91.92	93.58	93.28
0.80	82.39	77.27	41.31	20.31	87.00	87.70	90.32	88.66	86.05	85.85	84.05	84.27	89.50	91.32	92.86	92.48
0.85	84.39	78.35	75.00	71.53	88.82	88.34	88.22	89.82	86.89	86.84	86.63	86.83	90.06	91.94	91.94	93.34
0.90	84.98	82.09	76.20	73.92	85.40	91.02	90.56	89.94	86.54	88.18	87.26	87.96	88.28	92.42	92.84	93.22
0.95	80.78	81.27	80.31	76.09	82.54	84.50	87.18	88.18	84.86	86.96	87.83	87.51	86.60	89.58	91.70	91.70

Table C.2: Simulation results for a significance level α of 95%.

	Max t-stat (Naive)				Max t-stat (Max t-stat)				Cross-Validation (Naive)				Cross-Validation (Max t-stat)			
	n				n				n				n			
x	100	250	500	1000	100	250	500	1000	100	250	500	1000	100	250	500	1000
0.05	89.13	88.88	88.13	85.88	91.40	94.30	95.30	93.78	91.50	92.76	93.13	92.94	93.14	95.80	96.34	96.28
0.10	91.67	90.22	86.35	85.55	93.64	94.14	93.92	96.06	92.55	93.74	93.00	93.21	94.30	96.00	95.84	96.86
0.15	91.18	88.26	85.78	83.84	92.72	92.98	94.60	93.78	92.41	92.93	91.75	92.28	93.80	95.04	95.88	95.90
0.20	90.04	87.40	64.44	46.54	91.32	94.24	95.50	95.20	92.04	92.09	89.62	89.42	92.90	95.58	96.66	96.56
0.25	89.56	87.92	67.64	52.09	90.46	95.32	94.80	94.24	91.57	92.08	89.48	87.82	92.18	96.52	96.46	96.54
0.30	89.50	87.81	83.55	79.41	91.26	95.52	94.38	94.88	91.78	92.22	90.96	89.53	93.16	96.84	96.30	96.78
0.35	90.18	88.75	84.79	81.69	92.42	94.64	94.96	94.44	92.38	92.73	91.86	91.03	93.96	96.50	96.50	96.64
0.40	90.91	88.10	83.93	80.83	93.72	94.28	94.76	94.96	92.79	92.76	92.21	91.89	94.84	96.28	96.60	96.68
0.45	92.42	88.62	85.25	82.91	94.50	94.86	94.60	94.24	93.65	93.12	92.45	92.16	95.12	96.50	96.46	96.52
0.50	93.40	86.35	80.40	74.08	94.56	94.22	94.82	93.74	93.78	93.01	92.27	91.50	95.20	95.72	96.74	96.34
0.55	92.27	88.72	84.54	82.20	93.42	94.14	94.82	94.54	93.36	93.10	92.39	92.07	94.62	95.94	96.42	96.60
0.60	91.05	87.40	74.27	61.81	92.30	94.42	95.16	94.12	92.75	92.44	91.42	90.98	94.00	96.16	96.08	96.54
0.65	90.37	87.94	84.87	81.56	91.72	94.58	94.16	93.98	92.24	92.35	91.66	91.41	93.44	96.10	95.98	96.32
0.70	90.05	87.40	84.29	81.40	91.32	94.64	94.70	94.16	92.09	92.31	91.64	91.27	93.32	96.08	96.62	96.94
0.75	89.64	88.01	55.35	34.26	91.96	94.12	95.30	95.54	91.81	92.32	90.40	89.73	93.98	95.58	96.62	97.02
0.80	89.93	87.28	58.30	37.24	92.64	93.88	94.94	93.88	91.94	92.37	90.79	90.92	93.88	95.74	96.26	96.12
0.85	91.11	87.78	86.08	83.68	93.94	93.64	93.92	94.90	92.35	92.69	92.70	92.73	94.74	95.76	95.72	96.64
0.90	91.34	89.87	86.74	85.39	91.78	94.78	95.10	94.52	92.38	93.59	93.42	93.40	93.46	95.90	96.42	96.48
0.95	88.46	89.17	88.62	86.67	89.12	91.16	93.06	93.30	90.95	92.61	93.36	93.43	91.74	94.24	95.68	95.26

Table C.3: Simulation results for a significance level α of 97.5%.

	Max t-stat (Naive)				Max t-stat (Max t-stat)				Cross-Validation (Naive)				Cross-Validation (Max t-stat)			
	n				n				n				n			
x	100	250	500	1000	100	250	500	1000	100	250	500	1000	100	250	500	1000
0.05	93.41	93.75	93.38	92.18	94.36	96.94	97.68	97.06	94.96	95.89	96.27	96.08	95.50	97.80	98.06	98.22
0.10	95.01	94.67	92.33	92.25	96.36	96.72	96.46	98.08	95.68	96.53	96.01	96.50	96.96	97.66	97.54	98.56
0.15	94.90	93.71	91.53	91.14	95.80	96.24	97.16	96.62	95.72	95.95	95.09	95.86	96.36	97.22	97.96	97.72
0.20	94.25	93.02	76.74	61.96	94.84	96.62	97.42	97.42	95.29	95.37	93.77	93.36	95.74	97.30	98.18	98.12
0.25	94.15	93.24	79.58	67.47	94.58	97.26	97.14	97.32	95.25	95.48	93.75	92.75	95.62	97.94	97.94	98.30
0.30	94.21	93.27	90.29	87.79	95.14	97.64	96.90	97.58	95.30	95.58	94.67	94.04	96.02	98.22	97.80	98.38
0.35	94.43	93.69	91.35	89.90	96.16	96.92	97.42	97.18	95.53	95.91	95.39	95.01	97.06	98.06	98.12	98.30
0.40	94.99	93.59	91.06	89.16	96.72	96.84	97.08	97.22	96.10	96.24	95.79	95.49	97.46	98.18	98.16	98.32
0.45	95.61	94.10	91.70	90.24	97.08	97.38	97.24	97.02	96.39	96.47	95.91	95.63	97.40	98.36	98.32	98.14
0.50	96.19	92.67	88.59	84.42	97.02	97.24	97.56	96.66	96.44	96.34	95.89	95.03	97.50	97.90	98.52	98.24
0.55	95.59	93.94	91.80	89.86	96.40	96.84	97.62	97.52	96.28	96.31	95.90	95.69	97.22	97.66	98.34	98.34
0.60	94.84	93.10	84.32	75.35	96.02	96.90	97.48	97.02	95.85	95.86	94.96	94.80	96.78	97.74	98.14	98.30
0.65	94.36	93.53	91.53	89.42	95.50	96.92	96.76	97.18	95.46	95.94	95.42	95.24	96.46	97.92	97.82	98.30
0.70	94.18	93.21	91.21	89.22	95.52	97.22	97.30	97.18	95.44	95.81	95.31	95.31	96.66	97.92	98.20	98.36
0.75	94.12	93.20	69.04	48.39	95.42	97.04	97.70	97.80	95.33	95.78	94.24	93.77	96.56	97.82	98.50	98.54
0.80	94.48	92.88	71.54	51.64	95.80	96.54	97.56	96.86	95.62	95.76	94.76	94.35	96.60	97.62	98.16	97.92
0.85	94.97	93.36	92.55	90.87	96.16	96.88	97.00	97.30	95.74	96.12	96.09	95.90	96.62	97.88	97.78	98.10
0.90	94.79	94.44	92.45	91.88	94.86	96.94	97.32	97.04	95.54	96.44	96.34	96.34	96.06	97.58	98.00	98.08
0.95	92.87	93.59	93.64	92.50	92.98	94.88	96.32	96.14	94.55	95.75	96.35	96.13	94.68	96.76	97.64	97.36

C.3. Descriptive Statistics and Estimation Results

This section reports the descriptive statistics and the results from the log-linear model in (4.20) estimated on the full sample and the three subgroups described in Section 4.3.2.

Table C.4: Descriptive statistics on Blundell, Horowitz and Parey (2012) household data.

Statistics[a]	q	p	y	$share$	$\log(hhsize)$	$\log(driver)$	$\log(hhrage)$	$total\_wrkr$
n	2912	2912	2912	2912	2912	2912	2912	2912
$\bar{x}$	7.12	0.288757	11.05	0.035	1.38	0.78	3.64	1.88
std. dev.	0.65	0.042177	0.57	0.041	0.23	0.23	0.23	0.75
max	9.22	0.361399	11.69	0.823	2.64	2.30	4.48	10.00
75%	7.53	0.326036	11.41	0.040	1.61	0.69	3.78	2.00
50%	7.16	0.282514	11.12	0.026	1.39	0.69	3.66	2.00
25%	6.77	0.248640	10.77	0.017	1.10	0.69	3.50	1.00
min	2.69	0.228076	7.82	0.001	0.00	0.00	2.89	0.00

[a] *Notes: See the text for the definition and explanation of the variables reported in the table.*

Table C.5: OLS Regressions of log-linear model.

Regressors[a]	(1)	(2)	(3)	(4)
Log price	−1.206	−1.573	−1.131	−1.228
	[0.262]**	[0.395]**	[0.324]**	[0.349]**
Log income	0.308	0.446	0.382	0.350
	[0.020]*	[0.061]**	[0.048]**	[0.065]**
Log age of household respondent	0.0303	0.036	−0.045	−0.097
	[0.048]	[0.067]	[0.061]	[0.067]
Log household size	0.111	0.070	0.129	0.131
	[0.051]*	[0.074]	[0.063]∗	[0.070]
Log number of drivers	0.380	0.332	0.294	0.286
	[0.059]**	[0.091]**	[0.079]**	[0.087]**
Number of workers in household	0.099	0.102	0.099	0.097
	[0.017]**	[0.026]**	[0.022]**	[0.024]**
Public transit indicator	−0.111	−0.081	−0.082	−0.107
	[0.026]**	[0.043]	[0.032]**	[0.034]**
Small town	0.796	0.487	0.666	0.797
	[0.059]**	[0.141]**	[0.114]**	[0.152]**
Suburban	0.675	0.378	0.561	0.682
	[0.063]**	[0.146]**	[0.120]**	[0.158]**
Second city	0.667	0.422	0.589	0.696
	[0.060]**	[0.142]**	[0.118]**	[0.156]**
Urban	0.629	0.352	0.513	0.639
	[0.071]**	[0.154]**	[0.127]**	[0.162]**
Constant	2.767	1.639	2.331	2.815
	[0.231]**	[0.563]**	[0.462]**	[0.613]**
Population density (8 categories)	Yes	Yes	Yes	Yes
Observations	2912	1351	1858	1572
R^2	0.208	0.136	0.129	0.127

[a]*Notes: The dependent variable is log of annual gasoline demand measured in gallons. The heteroskedasticity-consistent standard errors are given in brackets below the point estimates. The superscripts ∗ and ∗∗ indicate significance at the 5 and 1 per cent level respectively. Regression (1) contains the output for the whole sample, whereas (2), (3) and (4) contain the results for the lower, middle and upper income groups. See Section 4.3.2 for details on these divisions.*

D Mathematical Tools

D.1. Rudelson's Inequality

An important inequality due to Rudelson (1999) is key in establishing Theorem D.1 which is used in Chapters 2 to 4. The inequality is an ingredient for proving non-asymptotic properties of high-dimensional covariance matrices equivalent to a law of large numbers in the asymptotic framework. In particular, Lemmas D.1 and D.2 lead to Theorem D.1 below.

Lemma D.1 (Rudelson's Inequality)**.** *Let $X_1, \cdots, X_n$ be a sequence of independent random vectors in $\mathbf{R}^m$ and $\varepsilon_1, \ldots, \varepsilon_n$ be a sequence of i.i.d. Rademacher variables. Then, conditionally on $X_1, \ldots, X_n$*

$$\mathrm{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i X_i X_i' \right\| \lesssim \sqrt{\log n \wedge m} \max_{1 \le i \le n} \|X_i\| \left\| \sum_{i=1}^n X_i X_i' \right\|^{1/2}.$$

Proof. See Rudelson (1999). ∎

The following lemma is a symmetrisation lemma for random matrices.

Lemma D.2. *Let $X_1, \ldots, X_n$ be independent random elements in $\mathbf{R}^{m \times m}$ and $\varepsilon_1, \ldots, \varepsilon_n$ be a sequence of i.i.d. Rademacher variables. Then,*

$$\mathrm{E} \left\| \sum_{i=1}^n X_i - \mathrm{E} X_i \right\| \le 2 \, \mathrm{E}_X \, \mathrm{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i X_i \right\|,$$

where the expectation over ε is conditional on X_i.

Proof. Let $Y_1, \dots, Y_n$ be an independent copy of $X_1, \dots, X_n$. Then, for fixed $X_1, \dots, X_n$

$$\begin{aligned}\left\| \sum_{i=1}^n X_i - \mathrm{E} Y_i \right\| &= \left\| \sum_{i=1}^n \mathrm{E}(X_i - Y_i) \right\| \\ &\leq \mathrm{E}_Y \left\| \sum_{i=1}^n X_i - Y_i \right\|,\end{aligned}$$

by Jensen's inequality. Therefore,

$$\mathrm{E}_X \left\| \sum_{i=1}^n X_i - \mathrm{E} X_i \right\| \leq \mathrm{E}_{X,Y} \left\| \sum_{i=1}^n X_i - Y_i \right\|, \tag{D.1}$$

and by symmetry

$$\mathrm{E}_{X,Y} \left\| \sum_{i=1}^n X_i - Y_i \right\| = \mathrm{E}_{X,Y} \mathrm{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i (X_i - Y_i) \right\|, \tag{D.2}$$

where the expectation over ε is taken conditionally on X_i and Y_i. Combining (D.1), (D.2), the triangle inequality and the fact that X_i and Y_i are identically distributed by construction establishes the lemma. ∎

Below I prove Theorem D.1 which is used in Chapters 2 to 4 as the behaviour of the quantity $\mathbb{E}_n[X_i X_i']$, for some $X_i \in \mathbf{R}^m$ appropriately defined in the text, plays a central role.

Theorem D.1. *Let $a_1, \dots, a_n$ be independent random variables and $X_1, \dots, X_n$ be independent m-dimensional random vectors with $\|X_i\|_2 \leq \xi_m$ for all $i = 1, \dots, n$. Furthermore, let $\hat{\Psi}_m = \mathbb{E}_n a_i^2 X_i X_i'$, $\Psi_m = \mathrm{E}\, \mathbb{E}_n a_i^2 X_i X_i'$ and $\mathrm{E}\, \mathbb{E}_n X_i X_i' = I_m$. Then,*

$$\mathrm{E} \|\hat{\Psi}_m - \Psi_m\| \lesssim v \sqrt{\frac{\xi_m^2 \log m}{n}} \vee \frac{v^2 \xi_m^2 \log m}{n}, \tag{D.3}$$

where $v := (\mathrm{E} \max_{1 \leq i \leq n} |a_i|^2)^{1/2}$. If $v^2 \xi_m^2 \log m \lesssim n$, then with probability at least $1 - \alpha$

$$\|\hat{\Psi}_m - \Psi_m\| \lesssim v \sqrt{\frac{\xi_m^2 \log m}{n\alpha^2}}. \tag{D.4}$$

Proof of Theorem D.1. The proof is standard and based on non-commutative Khichine inequalities for matrices. First, by the triangle inequality

$$\|\hat{\Psi}_m\| \leq \|\hat{\Psi}_m - \Psi_m\| + \|\Psi_m\| \lesssim \|\hat{\Psi}_m - \Psi_m\| + 1. \tag{D.5}$$

To bound the first term on the left-hand side, consider a sequence of independent Rademacher random variables $(\varsigma_i)_{i=1}^n$

$$\begin{aligned}
\mathrm{E}\,\|\hat{\Psi}_m - \Psi_m\| &\leq 2\,\mathrm{E}_{a,X}\,\mathrm{E}_{\varsigma|a,X}\,\|\mathbb{E}_n \varsigma_i a_i X_i X_i'\| \\
&\leq C\sqrt{\frac{\log m}{n}}\,\mathrm{E}\max_{1\leq i\leq n}\|a_i X_i\|\|\mathbb{E}_n a_i^2 X_i X_i'\|^{1/2} \\
&\leq C\sqrt{\frac{\xi_m^2 \log m}{n}}\,\mathrm{E}\max_{1\leq i\leq n}|a_i|\|\hat{\Psi}_m\|^{1/2} \\
&\leq C\nu\sqrt{\frac{\xi_m^2 \log m}{n}}\sqrt{\mathrm{E}\,\|\hat{\Psi}_m\|},
\end{aligned}$$

where the first inequality follows from Lemma D.2, the second by Lemma D.1, the third by $\|X_i\|_\infty \leq \xi_m$ and the final one by the Cauchy-Schwarz inequality. For shorthand, let $E = \mathrm{E}\,\|\hat{\Psi}_m - \Psi_m\|$ and $c = C\nu\sqrt{\frac{\xi_m^2 \log m}{n}}$. Taking expectations of (D.5) and combining it with the above inequality yields

$$E \leq c(E+1)^{1/2}.$$

Notice that for $x, c \geq 0$, $x \leq c\sqrt{x+1}$ implies that $x \leq (c^2 + \sqrt{c^4 + 4c^2})/2$. Hence,

$$\mathrm{E}\,\|\hat{\Psi}_m - \Psi_m\| \lesssim c + c^2,$$

using the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$. This establishes the first claim. The second claim follows directly from an application of Markov's inequality and the fact that

$$\nu\sqrt{\frac{\xi_m^2 \log m}{n}} \vee \frac{\nu^2 \xi_m^2 \log m}{n} = \nu\sqrt{\frac{\xi_m^2 \log m}{n}},$$

since $\nu^2 \xi_m^2 \log m \lesssim n$. ∎

D.2. Additional Tools

This section outlines some of the most commonly used tools in the proofs. By no means is this meant as a textbook exposition or a review chapter, but merely as an accompaniment to make the proofs more complete and ultimately easier to follows. For a precise treatment on this subject, I refer the reader to three excellent textbooks: van der Vaart and Wellner (1996), van der Vaart (1998) and Giné and Nickl (2016). However, any graduate-level textbook on mathematical statistics, in particular on infinite-dimensional models, will suffice.

In Chapters 3 and 4 a recurring quantity is the empirical process indexed by a class of functions $\mathcal{F}$

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_i^n f(X_i) - E[f(X_i)].$$

In order to make meaningful statements about this process I need the concept of a covering number of the class of functions $\mathcal{F}$ and the entropy of $\mathcal{F}$.

Definition D.1 (Definition 2.1.5 (van der Vaart and Wellner 1996))**.** Let $(\mathcal{F}, d)$ be some metric space of real functions $f : \mathcal{X} \to \mathbf{R}$. The covering number $N(\mathcal{F}, d, \tau)$ is the minimum number of d-balls of finite radius τ needed to cover the set $\mathcal{F}$. The entropy of $\mathcal{F}$ relative to d is the logarithm of the covering number.

Intuitively, the covering numbers are a measure for the complexity of the function classes indexing the empirical processes which determine how the suprema over these classes behaves. Lemmas B.4 to B.7 control the entropy of various classes of functions which are of interest in this thesis.

Definition D.2. A centred stochastic process X_f, $f \in \mathcal{F}$ is sub-Gaussian with respect to some distance d on $\mathcal{F}$ if

$$\mathrm{E} \exp\left\{t\left(X_f - X_{\check{f}}\right)\right\} \leq \exp\left\{t^2 d^2(f, \check{f})\right\} \quad \text{for } t \in \mathbf{R} \text{ and } f, \check{f} \in \mathcal{F}.$$

Sub-Gaussian processes play an important role due to the following result, known as Dudley's inequality (Dudley 1967).

Theorem D.2. *Let $(\mathcal{F}, d)$ be a metric space where $D := \operatorname{diam}(\mathcal{F})$ and let X_f be a centred sub-Gaussian process indexed by $\mathcal{F}$. If,*

$$\int_0^\infty \sqrt{\log N(\mathcal{F}, d, \tau)}\, \mathrm{d}\tau < \infty,$$

then,

$$\mathrm{E} \sup_{f \in \mathcal{F}} \left|X_f\right| \lesssim \int_0^{D/2} \sqrt{\log 2N(\mathcal{F}, d, \tau)}\, \mathrm{d}\tau.$$

Proof. See Theorem 2.3.7 in Giné and Nickl (2016). ∎

Theorem D.2 is convenient to work with as it is possible under suitable conditions to show that

$$\mathrm{E} \sup_{f \in \mathcal{F}} |\mathbb{G}_n f| \leq \mathrm{E} \sup_{f \in \mathcal{F}} |\mathbb{G}_n^o f|, \tag{D.6}$$

where $\mathbb{G}_n^o f$ is the symmetrised process from (1.4). Proceeding conditionally on the data, the symmetrised process is a sub-Gaussian process. This follows from the fact that Rademacher random variables have a bounded support such that Definition D.2 holds. As such, the right-hand side in (D.6) can be bounded by applying Dudley's inequality

above. The final step is to take the expectation over the data and bound this final expectation which is usually a much simpler quantity to deal with. For an application of this, see Step 1 in the proof of Lemma B.1.

To conclude this section, I state the anti-concentration result for the suprema of Gaussian processes from Chernozhukov, Chetverikov and Kato (2014a).

Theorem D.3 (Corollary 2.1 (Chernozhukov, Chetverikov and Kato 2014a))**.** *Let* X_f, $f \in \mathcal{F}$ *be a separable Gaussian process indexed by a semi-normed space* $\mathcal{F}$ *such that* $\mathrm{E}[X_f] = 0$ *and* $\mathrm{E}[X_f^2] = 1$ *for all* $f \in \mathcal{F}$*. Assume that* $\sup_{f\in\mathcal{F}} X_f < \infty$ *a.s. Define* $a(|X|) := \mathrm{E}\sup_{f\in\mathcal{F}} |X_f|$*, then for all* $\kappa \geq 0$

$$\sup_{x\in\mathbf{R}} P\Big(\Big|\sup_{f\in\mathcal{F}} |X_f| - x\Big| \leq \kappa\Big) \lesssim \kappa\Big(a(|X|) + 1\Big).$$

This result can be seen as a reverse of the traditional concentration inequalities for Gaussian processes. The result is most useful when $\kappa = o(a(|X|)^{-1})$ which implies that the supremum of $|X_f|$ cannot concentrate too fast around any $x \in \mathbf{R}$. This result plays a key role in establishing the validity of the critical values in Theorem 4.4. Traditional results assume the existence of a limiting distribution of an appropriately studentised empirical process in order to establish the validity of uniform confidence bands in nonparametric regression. However, as is the case in this thesis the studentised empirical process does not have a limiting distribution. The combination of the strong approximation in Theorem 4.1 and Theorem D.3 delivers the validity of the critical values and by extension that of the uniform confidence bands in Theorem 4.4.