Dimension Adaptive Estimation

Master's Thesis presented to the
Department of Economics at the
Rheinische Friedrich-Wilhelms-Universität Bonn

In Partial Fulfillment of the Requirements for the Degree of
Master of Science (M.Sc.)

Supervisor: Prof. Dr. Joachim Freyberger

Submitted in September 2023 by:
Chencheng Fang
Matriculation Number: 3466204

Contents

1 Introduction

In Econometrics, estimating a conditional mean function of a response variable on a group of independent variables is always a pivotal task. However, more often than not, the true form of underlying model that defines the conditional mean function is unknown to researchers, thus making the estimation challenging.

To be specific, if the underlying model is in a parametric form, using a non-parametric estimator would result in a slower convergence rate than a parametric estimator. Contrarily, if the true form is non-parametric, the parametric estimator would then not converge (in probability; hereafter) to the true function, but a non-parametric estimator does. Without knowing the underlying model, it seems difficult to choose an estimator that converges at an optimal rate across all types of true models. This brings up an open question of how to construct an estimator that always converges at an optimal rate, no matter what the form of underlying model is.

To answer the question, this thesis proposes such an estimator with the use of sieve estimation and Lasso-type methods, and it is termed as *dimension adaptive estimator*. This estimator can always adapt its convergence rate to the underlying dimensionality of true models. It converges as fast as a parametric estimator in case of parametric underlying models, and it also converges in case of non-parametric models, though at a slower non-parametric rate. To put it simply, we first construct a multivariate sieve space from the group of regressors. Sieves are found to have a good property of approximating unknown functions under certain assumptions (Chen 2007). Then, we use Lasso-type methods to select significant terms in the sieve so as to achieve effective dimension reduction without losing predictability. Finally, we apply a theorem in Bunea, Tsybakov, and Wegkamp 2007 to argue that the convergence rate of our estimator is dimension adaptive.

This thesis is structured as follows. Chapter 2 explains the definition of sieve estimation and its important features we need to use. Also, several typical basis functions for the construction of sieves are introduced. Chapter 3 briefly gives the definition of two Lasso-type methods, including Lasso and adaptive Lasso. The corresponding post-selection methods are discussed as well. In Chapter 4, a key theorem from Bunea, Tsybakov, and Wegkamp 2007 is presented, which is applied in Chapter 5 to demonstrate the properties of our dimension adaptive estimator. Chapter 6 shows the results of simulations that turn out to strongly support the theoretical properties we find in Chapter 5. Chapter 7 applies our estimator in a real-life dataset. Chapter 8 closes this thesis with conclusions and potential extensions.

All codes, data and an R package we developed for this thesis can be found or tracked in the following GitHub repository: `https://github.com/ccfang2/Masters_Thesis`.

2 Sieve Estimation

It is very common that semi- or non-parametric underlying models concern unknown parameters that are in spaces of infinite dimensions, thus making it computationally hard to solve such models with finite samples. Even if one manages to solve them, the optimization problem over infinite-dimensional non-compact space[1] may be ill-posed, which leads to unattractive large sample properties including inconsistency and slow rate of convergence. The method of sieves developed by Grenander 1981 is a popular way to resolve this problem. It tries to optimize a criterion function over significantly simpler and often finite-dimensional spaces, which are also known as sieves.

The above-mentioned unknown infinite-dimensional parameters can also be seen as a member of some function space under certain regularities, and it has been widely shown that sieves can well approximate such unknown functions with computational feasibility. Sieves can be built with linear spans of power series, splines, trigonometric polynomials and many other well-known basis functions, which are very easy to implement. Despite the good approximation performance and easy implementation, the properties of sieve method cannot be justified using theories from parametric models, since any large sample theory for sieve method should consider a trade-off between approximation error and model complexity. The former originates from the replacement of original parameter space with a simpler sieve space, while the later arises from the fact that complexity of sieves needs to increase with sample size to ensure consistency of the method.

Section 2.1 discusses more deeply about how sieve estimation resolves the ill-posed problem of optimizing a criterion function over infinite-dimensional non-compact space. Section 2.2 is about series estimation, a special case of sieve estimation with concave criterion functions and finite-dimensional linear sieve spaces. Section 2.3 defines the most popular smoothness classes of functions[2] in semi- or non-parametric literature. Section 2.4 closes this chapter with an introduction of typical basis functions, which are used to construct sieves.

2.1 Sieve Extremum Estimation

Following the arguments in Chen 2007, let Θ to be an infinite-dimensional parameter space with a (pseudo[3]-) metric d. In a semi- or non-parametric model, there is usually a population criterion function $\mathcal{Q} : \Theta \to \mathbb{R}$, which is uniquely maximized at a (pseudo-)

[1] Infinite-dimensional space is often non-compact but doesn't have to, because different metrics on an infinite-dimensional space may not be equivalent to each other. Consequently, a space may be non-compact under one metric, but can be compact under another one.

[2] They are the unknown functions we hope to approximate with sieves.

[3] A pseudo-metric space keeps all properties of metric space except that it allows the distance between two distinct points to be zero.

true parameter $\theta_0 \in \Theta$. Although θ_0 is unknown, a sample of observations $\{Z_i\}_{i=1}^n$, $Z_i \in \mathbb{R}^{d_z}$, $1 \leq d_z \leq \infty$ could be drawn. An empirical criterion $\widehat{\mathcal{Q}}_n : \Theta \to \mathbb{R}$, which is a measurable function of $\{Z_i\}_{i=1}^n$ for all $\theta \in \Theta$ is then obtained. By Uniform Law of Large Numbers[4], uniform convergence of $\widehat{\mathcal{Q}}_n$ to $\mathcal{Q}$ is derived. Together with other conditions[5], this uniform convergence further ensures the maximizer $\hat{\theta} = \arg\sup_{\theta\in\Theta} \widehat{\mathcal{Q}}_n$ is consistent to the (pseudo-) true parameter θ_0, and $\hat{\theta}$ here is called an extremum estimate.

However, a sufficient[6] assumption for the uniform convergence of $\widehat{\mathcal{Q}}_n$ and consistency of $\hat{\theta}$ is the compactness of Θ. Therefore, when Θ is infinite-dimensional and potentially non-compact, the maximization of $\widehat{\mathcal{Q}}_n$ over Θ might be ill-posed. Even if $\hat{\theta}$ exists, it is often computationally infeasible, or has undesirable large sample properties.

Definition 2.1 (Ill-posed versus Well-posed Problems)
In line with the definition in Chen 2007, an optimization problem is well-posed if for all sequences $\{\theta_k\}$ in Θ such that $\mathcal{Q}(\theta_0) - \mathcal{Q}(\theta_k) \to 0$ indicates $d(\theta_0, \theta_k) \to 0$; is ill-posed if $\exists$ a sequence $\{\theta_k\}$ in Θ such that $\mathcal{Q}(\theta_0) - \mathcal{Q}(\theta_k) \to 0$ but $d(\theta_0, \theta_k) \not\to 0$

No matter whether the maximization problem is well- or ill-posed, one may think if we can approximate Θ by a sequence of a simpler compact space Θ_n such that for any $\theta \in \Theta$, there exists $\pi_n\theta \in \Theta_n$ satisfying $d(\theta, \pi_n\theta) \to 0$ as $n \to \infty$, where π_n is a projection mapping from Θ to Θ_n. The answer is yes, and the sequence of Θ_n is called sieves by Grenander 1981. Popular sieves are usually compact and non-decreasing ($\Theta_n \subseteq \Theta_{n+1} \subseteq \cdots \subseteq \Theta$)[7].

The so-called approximate sieve extremum estimate, $\hat{\theta}_n$, is an approximate maximizer of $\widehat{Q}_n(\theta)$ over the sieve space Θ_n.

$$\widehat{Q}_n(\hat{\theta}_n) \geq \sup_{\theta\in\Theta_n} \widehat{Q}_n(\theta) - O_p(\eta_n) \tag{1}$$

where $\eta_n \to 0$, as $n \to \infty$. When $\eta_n = 0$, $\hat{\theta}_n$ is called exact sieve extremum estimate. Following Theorem 2.2 in White and Wooldridge 1992, Chen 2007 shows the sufficient conditions for the existence and measurability of $\hat{\theta}_n$: (i) $\widehat{Q}_n(\theta)$ is a measurable function of observations $\{Z_i\}_{i=1}^n$ for all $\theta \in \Theta_n$; (ii) for any $\{Z_i\}_{i=1}^n$, $\widehat{Q}_n(\theta)$ is upper semicontinuous[8] on Θ_n under metric $d(\cdot,\cdot)$; and (iii) sieve space Θ_n is compact under $d(\cdot,\cdot)$.

[4]Theorem 22.2, B. Hansen 2022a

[5]Theorem 22.3, B. Hansen 2022a

[6]Compactness is a sufficient but not necessary condition for the uniform convergence and consistency. See Theorem 18.2, B. Hansen 2022b for details.

[7]In order to ensure consistency, the complexity of sieves is required to increase with sample size. So, in the limit, the sieves are dense in the original potentially non-compact parameter space Θ. Otherwise, there may be some issues of underfitting.

[8]A function f: $X \to \bar{\mathbb{R}}$ is called upper semicontinuous if and only if $\limsup_{x\to x_0} f(x) \leq f(x_0)$, where $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. A function is continuous if and only if it is both upper and lower semicontinuous. Here, as we focus on the maximizer (not minimizer) of $\widehat{Q}_n$, upper semicontinuity is sufficient.

When $\widehat{Q}_n(\theta)$ is expressed as a sample average,

$$\sup_{\theta\in\Theta_n} \widehat{Q}_n(\theta) = \sup_{\theta\in\Theta_n} \frac{1}{n}\sum_{i=1}^{n} l(\theta, Z_i) \tag{2}$$

where l: $\Theta \times \mathbb{R}^{dz} \to \mathbb{R}$ is the criterion based on a single observation, $\hat{\theta}_n$ in (1) is also called an approximate sieve maximum-likelihood-like (M-) estimate. Clearly, if $\Theta_n = \Theta$ for all n, the sieve extreme estimation method reduces to a standard extreme estimation.

2.2 Series Estimation

More often, we deal with series estimation, which is a special case of sieve M-estimation with concave criterion functions $\widehat{Q}_n(\theta)$ and finite-dimensional linear sieve space Θ_n, as defined in Chen 2007.

A criterion function $\widehat{Q}_n(\theta)$ is said to be concave if $\widehat{Q}_n(\tau\theta_1 + (1-\tau)\theta_2) \leq \tau\widehat{Q}_n(\theta_1) + (1-\tau)\widehat{Q}_n(\theta_2)$ for any $\theta_1, \theta_2 \in \Theta$ and any scalar $\tau \in (0,1)$[9]. A sieve space is said to be finite-dimensional linear if it is a linear span of finitely many known basis functions, which will be discussed in Section 2.4. These two assumptions make it much easier to resolve the optimization problem of sieve M-estimation, and simplify the descriptions of its large sample properties (See Section 3 in Chen 2007 for details).

2.2.1 Univariate LS Series Estimation

Considering the advantages of series estimation, we will use this specific type of sieve M-estimation in the construction of dimension adaptive estimator. As a consequence, a concave criterion function needs to be chosen, and in this thesis, we use Least Squares (LS) criterion. The definition of univariate LS criterion is given.

Definition 2.2 (Univariate Least Squares Criterion)
The estimation of an unknown conditional mean function $\theta_0(\cdot) = h_0(\cdot) = E(Y|X = \cdot)$ is concerned. Define $Z = (Y, X)$, where X has a bounded support $\mathcal{X}$ in $\mathbb{R}$. Suppose $h_0 \in \Theta$, where Θ is a linear subspace of the space of functions h with $E[h(X)^2] < \infty$. Then, the univariate LS criterion function is $\mathcal{Q}(\theta) = -E\{[Y - h(X)]^2\}$. $Q(\theta)$ is found to be strictly concave in $h \in \Theta$[10].

With this concave LS criterion, we can further define series LS estimator in the univariate setup. Denote $\{p_k(X)\}_{k=0}^{\infty}$ as a sequence of basis functions that can well approximate any real-valued square integrable functions of X[11]. Then, a finite-dimensional sieve space Θ_n

[9]This also implies the convexity of parameter space Θ, i.e., for any $\theta_1, \theta_2 \in \Theta$, $\tau\theta_1 + (1-\tau)\theta_2 \in \Theta$ for any scalar $\tau \in (0,1)$.

[10]This is obvious by taking the second derivative of Q on h.

[11]A function h: $\mathbb{R} \to \mathcal{C}$ is square integrable if and only if $\int_{-\infty}^{\infty} |h(x)|^2 dx < \infty$. This is a sufficient condition for $E[h(X)^2] < \infty$.

for Θ is:

$$\Theta_n = \mathcal{H}_n = \left\{ h : \mathcal{X} \to \mathbb{R}, h(x) = \sum_{k=0}^{K_n} \beta_k p_k(x) : \beta_0, \ldots, \beta_{K_n} \in \mathbb{R} \right\} \tag{3}$$

where K_n, the dimensionality of Θ_n, goes to infinity slowly as $n \to \infty$. Then, the so-called series LS estimator of the conditional mean $E(Y|X = \cdot)$ is $\hat{h} = \arg\max_{h \in \mathcal{H}} -\frac{1}{n} \sum_{i=1}^{n} [Y_i - h(X_i)]^2$. Furthermore, owing to the linearity of Θ_n, this series LS estimator $\hat{h}$ has a simple closed-form expression:

$$\hat{h}(x) = p^{K_n}(x)^T (P^T P)^+ P^T Y \tag{4}$$

where $x \in \mathcal{X}$, $p^{K_n}(X) = (p_0(X), \ldots, p_{K_n}(X))^T$ and $P = (p^{K_n}(X_1), \ldots, p^{K_n}(X_n))^T$. $(P^T P)^+$ is Moore-Penrose generalized inverse[12].

2.2.2 Multivariate LS Series Estimation

A natural extension of the univariate setup is a multivariate setting where there are more than one independent variable. Under multivariate setting, we need to include product-terms (i.e., interactions) of basis functions among different variables. Accordingly, a finite-dimensional multivariate sieve space for Θ then becomes:

$$\Theta_n = \mathcal{H}_n = \left\{ h : \mathcal{X} \to \mathbb{R}, h(x) = \sum_{k_1=0}^{K_n} \cdots \sum_{k_d=0}^{K_n} \beta_{k_1,\ldots,k_d} \prod_{l=1}^{d} p_{k_l}(x_l) : \beta_{k_1,\ldots,k_d} \in \mathbb{R} \right\} \tag{5}$$

where $\mathcal{X}$ is the support of X in $\mathbb{R}^d$, and K_n still goes to infinity slowly as $n \to \infty$. This type of multivariate space has been analyzed as *Tensor Product Space* in literature (e.g., Lin 2000; Zhang and Simon 2022). The definition of series LS estimator in univariate setup still applies in this tensor product space except for the difference of $\mathcal{H}_n$. Clearly, the closed-form expression in (4) also works as long as we add all product-terms into $p^{K_n}(X)$. Since this thesis focuses on a multivariate setup, this tensor product space will be used in construction of our dimension adaptive estimator.

2.3 Smoothness Classes

Before introducing basis functions used in the creation of sieves, we need to define the smoothness classes of functions of interest, because how well an unknown function can be approximated by basis functions depends on its smoothness. Throughout this thesis, we use Hölder class of functions, which is the most common smoothness class in semi- or non-parametric literature; see e.g., Newey 1997, Horowitz 1998. In accordance with Chen 2007, we first give the definition of Hölder condition.

[12]This is a generalization of inverse matrix. In least squares estimation, it can be used to compute the "best fit" solution to a system of linear equations that lacks a solution. See `https://en.wikipedia.org/wiki/Moore-Penrose_inverse` for details.

Definition 2.3 (Hölder Condition)
Suppose that $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ is the Cartesian product of compact intervals $\mathcal{X}_1, \ldots, \mathcal{X}_d$. A real-valued function h on $\mathcal{X}$ is considered to satisfy a Hölder condition with exponent $\gamma \in (0, 1]$ if there is a positive number c such that $|h(x) - h(y)| \leq c|x - y|_e^\gamma$ for all $x, y \in \mathcal{X}$, where $|x|_e = (\sum_{l=1}^d x_l^2)^{1/2}$ is the Euclidean norm of $x = (x_1, \ldots, x_d)$.

Prior to defining smoothness, we also need to denote the differential operator by

$$D^\alpha = \frac{\partial^{[\alpha]}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$ of nonnegative integers, and $[\alpha] = \alpha_1 + \cdots + \alpha_d$.

Definition 2.4 (The p-smooth Function)
Let m be a nonnegative integer and set $p = m + \gamma$, where γ is the exponent for Hölder condition in Definition 2.3. We say, a real-valued function h on $\mathcal{X}$ is p-smooth if it is m times continuously differentiable on $\mathcal{X}$ and $D^\alpha h$ satisfies a Hölder condition with exponent γ for all α with $[\alpha] = m$.

The Hölder (or p-smooth) classes of functions are popular because a p-smooth function can be well approximated by basis functions. In Section 5, the unknown conditional mean function we want to approximate is assumed to be p-smooth.

2.4 Typical Basis Functions

As already mentioned in Section 2.2, a sieve space is finite-dimensional linear if it is a linear span of finitely many known basis functions. Even though we are dealing with a multivariate setup, the sieve space defined in (5) is still linear if product-terms are viewed as basis functions for interactions among different independent variables. In this section, we present several typical basis functions in a univariate setup. Therefore, they correspond to $p_{k_l}(x_l)$ in (5).

2.4.1 Power Series

For $X \in \mathbb{R}$, the power series bases are $p_k(X) = X^k$ for $k = 0, 1, 2, \ldots$. In multivariate sieve space, we have an n-varying K_n rather than a fixed k. So, they could then be written in vector notation as below.

$$p^{K_n}(X) = \begin{pmatrix} 1 \\ X \\ \vdots \\ X^{K_n} \end{pmatrix} \tag{6}$$

2.4.2 Legendre Polynomials

They are orthogonal with aspect to uniform density on $[-1, 1]$. To use the advantage of orthogonality, original variables should usually be rescaled to have support in $[-1, 1]$. For $X \in [-1, 1]$, Legendre polynomial bases are:

$$p_k(X) = \frac{1}{2^k} \sum_{l=0}^{k} \binom{k}{l}^2 (X-1)^{k-l}(X+1)^l$$

Using an n-varying K_n, we can write them in vector notation.

$$p^{K_n}(X) = \begin{pmatrix} 1 \\ X \\ \frac{1}{2}(3X^2-1) \\ \frac{1}{2}(5X^3-3X) \\ \vdots \end{pmatrix} \tag{7}$$

The first six Legendre polynomials are depicted in Figure 1.

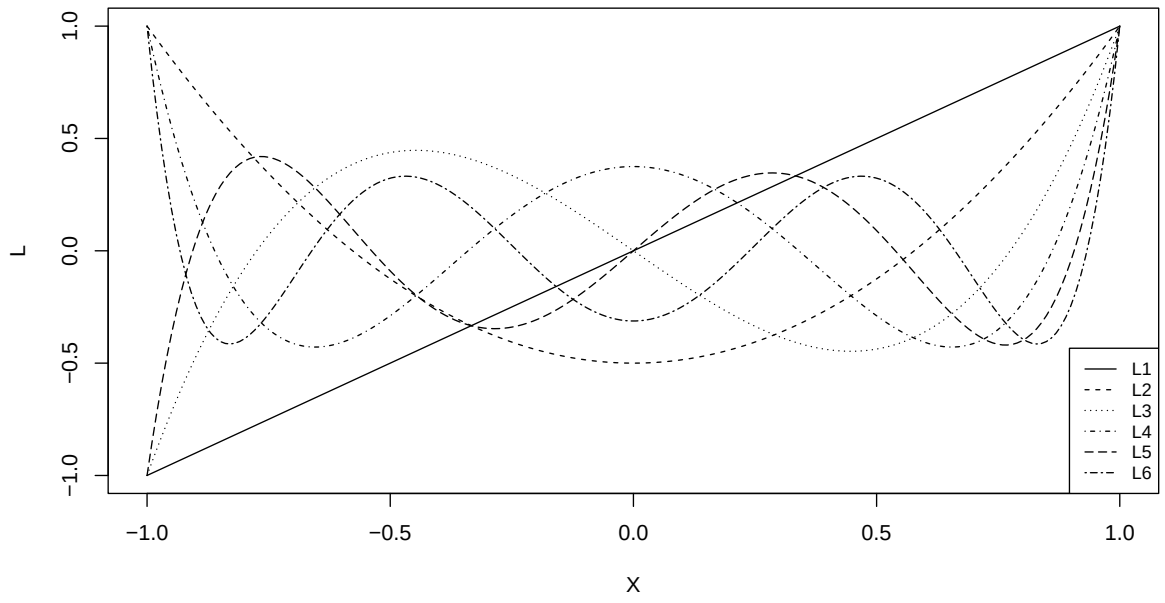

Figure 1: The first six Legendre polynomial basis

As is seen in Figure 1, range of the first six Legendre polynomials[13] are bounded in $[-1, 1]$. However, sometimes if we hope to have orthonormal Legendre polynomials, we need to further nomalize them which may then make them unbounded in range. See more discussions in Appendix 10.3.1.

[13]Actually, range of all Legendre polynomials are bounded in $[-1, 1]$.

2.4.3 Splines

In general, a spline is a piecewise polynomial, the order of which is pre-selected. The flexibility of splines lies in the number of polynomial segments. The joints between segments are called knots. To put it formally, a $(q+1)^{th}$ order[14] spline with N internal knots $\tau_1 < \tau_2 < \cdots < \tau_N$ has $N+q+1$ spline bases, and we can write them in vector notation.

$$p^{K_n}(X) = p^{N_n+q+1}(X) = \begin{pmatrix} 1 \\ X \\ \vdots \\ X^q \\ (X-\tau_1)^q \mathbb{1}\{X \geq \tau_1\} \\ \vdots \\ (X-\tau_{N_n})^q \mathbb{1}\{X \geq \tau_{N_n}\} \end{pmatrix} \tag{8}$$

where $K_n = N_n + q + 1$. Since K_n in sieve space grows slowly with n, the number of internal knots N_n should also increase slowly with n to ensure the consistency of sieve estimation. Moreover, in this thesis, once the number of internal knots N_n is given, we define the positions of knots as quantiles $j/(N_n+1)$ for $j \in (1, \ldots, N_n)$ so that the probability mass is equalized across all segments. This rule of defining knot positions also applies to the construction of following B-Splines.

2.4.4 B-Splines

B-spline is short for basis spline, and B-splines of order $(q+1)$ are basis functions of splines with the same order over the same knots. All possible splines can be built from a linear combination of B-splines, and the linear combination is unique for each spline. In consistence with the definition of splines, let N_n denote the number of internal knots, and let q be the polynomial degree of spline. Then, in the construction of B-splines, we have a total number of knots $M_n = N_n + 2 \times (q+1)$ with $q+1$ overlapping knots at each boundary. Knots are indexed as $\tau_0 \leq \tau_1 \leq \cdots \leq \tau_{M_n-1}$. The i^{th} B-spline for a $(q+1)^{th}$ order spline could then be built by Cox-de Boor recursion formula.

$$\begin{aligned} B_{i,0}(X) &= \begin{cases} 1 & \text{if} \quad \tau_i \leq X \leq \tau_{i+1} \\ 0 & \text{otherwise} \end{cases} \\ B_{i,q}(X) &= \frac{X-\tau_i}{\tau_{i+q}-\tau_i} B_{i,q-1}(X) + \frac{\tau_{i+q+1}-X}{\tau_{i+q+1}-\tau_{i+1}} B_{i+1,q-1}(X) \end{aligned} \tag{9}$$

[14] Accordingly, the degree of piecewise polynomial is q. Please be aware that degree=order-1.

where $0 \leq i \leq M_n - 1 - q - 1 = N_n + q$. Obviously, we have $M_n - q - 1 = N_n + q + 1$ B-splines for a $(q+1)^{th}$ order spline with N_n internal knots. Figure 2 depicts B-Splines with polynomial degree $q = 3$ and a total number of knots $M_n = 11$ ($q + 1 = 4$ overlapping knots at each boundary and $N_n = 3$ internal knots) on $[0, 1]$.

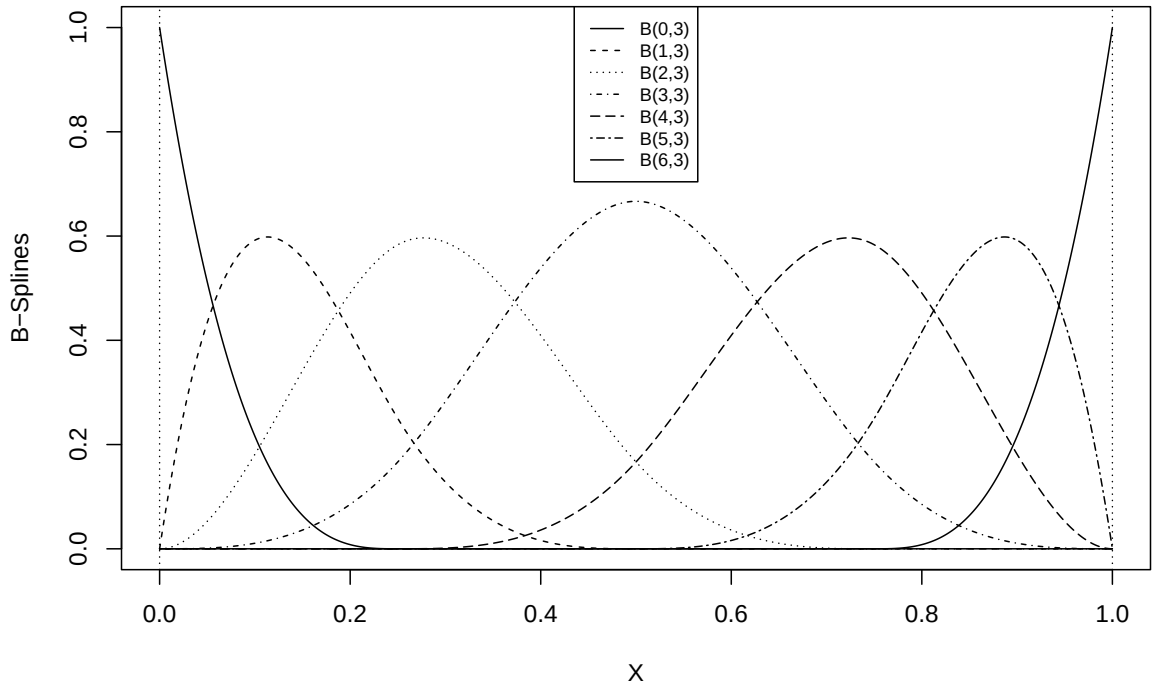

Figure 2: B-splines

In Figure 2, $B(i, q)$ correponds to $B_{i,q}$ in (9). This set of seven B-Splines can be used to create any spline function with order of 4 (i.e., polynomial degree 3) and over the same knots[15].

2.4.5 Trigonometric Polynomials

For $X \in [-1, 1]$, we can write trigonometric polynomial bases in a vector notation, using an n-varying K_n.

$$p^{K_n}(X) = \begin{pmatrix} 1 \\ \cos(\pi X) \\ \sin(\pi X) \\ \vdots \\ \cos(K_n \pi X) \\ \sin(K_n \pi X) \end{pmatrix} \tag{10}$$

[15]See `http://web.mit.edu/hyperbook/Patrikalakis-Maekawa-Cho/node17.html` for a similar example and more details about B-Splines.

Trigonometric polynomials are orthogonal in the domain of $[-1, 1]$. See Appendix 10.3.4 for the proof. They are a combination of sine and cosine polynomials, thus likely being good at approximating periodic functions.

2.4.6 Haar Wavelets

Haar sequence was proposed by Haar 1910, and later they were designed to be one of the most well-known wavelets. To construct Haar basis functions, original variables should be rescaled to $[0, 1]$, and there is an increasing level in their construction. In level 0, there is only one basis function, and in level 1, there become two basis functions. With the level going up, the number of basis functions doubles. To put it formally, we denote level as $j \in \{0, 1, 2, \ldots\}$, so the number of basis functions in each level is 2^j. Let $k \in \{0, 1, \ldots, 2^j - 1\}$ represent the index of basis functions in each level. For $X \in [0, 1]$, the normalized and shifted Haar basis functions[16] are:

$$\psi_{jk}(X) = 2^{j/2}\psi(2^j X - k) \tag{11}$$

with

$$\psi(X) = \begin{cases} 1 & \text{if} \quad 0 \leq X < 1/2 \\ -1 & \text{if} \quad 1/2 \leq X < 1 \\ 0 & \text{otherwise} \end{cases}$$

It is found that the basis function $\psi_{j,k}(X)$ is supported on a right-open interval $[\frac{k}{2^j}, \frac{k+1}{2^j})$ and it vanishes outside this interval. With some simple calculations, Haar basis functions are obviously pairwise orthogonal. Moreover, since Haar basis functions are normalized here, the L_∞-norm is $||\psi_{jk}(X)||_\infty = 2^{j/2}$, which increases with level exponentially.

[16] In practice, there is usually an additional basis function which is equal to 1 over $X \in [0, 1]$. It is used to approximate intercept.

3 Lasso-type Methods

In Section 2, we have clarified how to build up a finite-dimensional multivariate sieve space from original independent variables. Sieves are proven to have good approximation of unknown functions under certain assumptions (Chen 2007). In this section, we will introduce two Lasso-type methods, which are then applied to select significant terms in the constructed sieve, thus realizing dimension reduction without losing prediction accuracy. These methods are employed in the second step of our dimension adaptive estimation, and are crucial in adapting the convergence rate of our estimator to underlying dimensionality of different true models.

3.1 Lasso

The Lasso, proposed by Tibshirani 1996, is short for *Least Absolute Shrinkage and Selection Operator*. It is popular for high-dimensional problems due to its statistical accuracy of prediction and computationally feasible variable selection.

Consider a setting where $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ is a sample of independent random pairs (Z, Y) with $Z \in \mathbb{R}^p$ being covariates and $Y \in \mathbb{R}$ being a response variable. For a continuous Y, a simple approach is to establish a linear model.

$$Y = \sum_{j=1}^{p} Z_j \beta_j + \epsilon \tag{12}$$

where $Z_j = (Z_{1j}, Z_{2j}, \ldots, Z_{nj})^T$ for $j = 1, 2, \ldots, p$, and $E[\epsilon|Z] = 0$. A straightforward estimation method is Ordinary Least Squares (OLS). This would not be a problem if $p < n$. However, if $p > n$, OLS estimator is not unique any more and will heavily overfit the sample data. This is particularly true for our research when the number of terms in multivariate sieve space is usually much larger than the sample size. Therefore, a technique is needed to reduce complexity while maintaining approximation accuracy. Lasso is such a dimension reduction technique, and it does so with a regularization of ℓ_1 penalty. The Lasso estimator for linear model in (12) is:

$$\hat{\beta}(\text{lasso}) = \arg\min_{\beta} \left(\left|\left| Y - \sum_{j=1}^{p} Z_j \beta_j \right|\right|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{13}$$

where λ is a non-negative regularization parameter. $||\cdot||^2$ indicates squared L_2 norm. The second term in (13) is the so-call ℓ_1 penalty, which is key to dimension reduction.

Specifically, Lasso diminishes coefficients to 0 as λ increases, because it has to balance the two terms in (13) such that the summation is minimized. If λ is sufficiently large, some

coefficients are diminished to exact 0[17]. The variables corresponding to those coefficients are usually of little significance to approximating the response variable, so their being exact 0 only sacrifices a little bit in-sample approximation accuracy but gains much in dimension reduction. Owing to such a bias-variance trade-off, Lasso often helps to improve out-of-sample prediction accuracy. Also, it has been discovered that ℓ_1 regularization can find out the "right" sparse representation of underlying models with certain assumptions (Donoho and Elad 2003).

That being said, in using Lasso for prediction, there are still three more issues researchers need to be aware of. Firstly, variables need to be rescaled beforehand, because the rule of variable selection depends on the scale of coefficients, which is associated with the magnitude of variables. To remove such an unfair influence of magnitude, one needs to rescale variables before the use of Lasso[18]. Secondly, the regularization parameter λ in (13) is pre-defined, but researchers usually don't know how large it should be in practice. Thus, cross-validation is commonly applied to select the "best" λ for Lasso. Thirdly, because ℓ_1 regularization can recover the "right" sparsity of underlying models, we can always estimate the model very well simply by performing an OLS regression restricted to the selected subset (Section 11.4 of Trevor Hastie and Wainwright 2015). It is a common practice to run OLS (for a linear model) on all selected variables from Lasso, which is the so-called post Lasso Method[19]. The estimated coefficients from post Lasso can then be used for prediction on a test or new dataset.

The concept of Lasso also extends to cases of generalized linear models including Logistic model and Poisson model. Please refer to Chapter 3 of Bühlmann and Geer 2011 for such extensions. Under these models, the estimation methods for post Lasso should also be changed from OLS to other methods accordingly.

3.2 Adaptive Lasso

Despite the fact that Lasso has a good performance in dimension reduction, its consistency with respect to variable selection is not guaranteed unless some conditions are satisfied. Meinshausen and Bühlmann 2006 show a condition that such consistency of Lasso may need, and they claim that the condition can not be further relaxed. Zou 2006 also summarises and re-interprets this necessary condition, but he further asks himself if Lasso could be revised in a way such that the new estimator is still consistent even when

[17]The reason behind some coefficients being shrunk to exact 0 instead of extremely small values lies in the shape of ℓ_1 constraint. See Page 9-19 of Bühlmann and Geer 2011 for details.

[18]See Appendix 10.1 for more details of why rescaling before analysis is required. Moreover, good news is that glmnet, the most popular R package for Lasso, implicitly does standardization inside the code by default, so usually there is no need to worry.

[19]Belloni, Chernozhukov, and C. Hansen 2013 argue that one needs to be cautious about doing inference with post Lasso due to the failure of uniform inference. They develop a "post-double-selection" method which is able to perform inference uniformly.

that condition is further considerably relaxed. His answer is yes, and he terms the new estimation method as *adaptive Lasso*.

As shown in (13), with a ubiquitous penalization parameter λ, Lasso forces all coefficients to be equally penalized. To revise it, Zou 2006 allows different weights to different coefficients, and adaptive Lasso is thus actually a weighted Lasso. Suppose $\hat{\beta}$ is a root-n consistent estimator to the true parameter β. For example, it can be an OLS estimator. Pick up a tuning parameter $\gamma > 0$ (usually $\gamma = 1$), and the weight for coefficient β_j is defined as $\hat{w}_j = 1/|\hat{\beta}_j|^\gamma$. Then, the adaptive lasso estimator for the linear model in (12) is given:

$$\hat{\beta}(\text{adaptive lasso}) = \arg\min_{\beta} \left(\left|\left| Y - \sum_{j=1}^{p} Z_j \beta_j \right|\right|^2 + \lambda \sum_{j=1}^{p} \hat{w}_j \, |\beta_j| \right) \tag{14}$$

In (14), it is seen that the penalization for larger coefficients is smaller, thus making them have a higher chance of survival in variable selection. Zou 2006 has proved that adaptive Lasso is consistent in variable selection with a much more relaxed condition than Lasso. he also clarifies that adaptive Lasso could be solved by the same efficient algorithm as that for solving Lasso. Moreover, the method of adaptive Lasso can easily be extended to generalized linear models as well.

In practice, adaptive Lasso is often realized in a two-stage approach. Firstly, Lasso is performed on original dataset, so some variables are selected with non-zero coefficients. In order to obtain the estimated coefficients[20] from a root-n consistent estimator, post Lasso is usually applied on those selected variables. Secondly, estimated coefficients from post Lasso are applied to calculate weights which are then used in adaptive Lasso.

Again, to get estimated coefficients for prediction in a test or new dataset, it is a usual practice to perform post adaptive Lasso just like how we do post Lasso, i.e., running OLS (for linear models) on selected terms from adaptive Lasso.

[20]Sometimes, estimated coefficients from Lasso or ridge regression are used to calculate weights even though they are not from a root-n consistent estimator.

4 Bunea, Tsybakov, and Wegkamp 2007

In this section, we introduce a major theorem established in Bunea, Tsybakov, and Wegkamp 2007, which will be later used to argue about the properties of our dimension adaptive estimator. In their paper, they show that Lasso-type estimator in random design non-parametric regression satisfies sparsity oracle inequalities[21], i.e., bounds in terms of the number of non-zero components of the oracle vector. They prove that the finding holds even when the original dimension of data is much larger than the sample size or the regression matrix is not positive definite.

This setting of random design non-parametric regression fits well into our research, because the first step of our dimension adaptive estimation is the generation of a multivariate sieve space from original dataset as shown in (5). This is a random design non-parametric technique where the number of terms in the generated sieve grows slowly with sample size. Therefore, the oracle inequalities established in their paper is crucial in arguing about the properties of our dimension adaptive estimation.

Section 4.1 introduces the random design non-parametric setup in their paper, and Section 4.2 presents the concepts of sparsity and weak sparsity which are used in their theorem. In Section 4.3, assumptions of the theorem are listed and Section 4.4 concludes this chapter with their theorem that we will use later in Chapter 5.

4.1 Setup

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of independent random pairs with $(X, Y) \in (\mathcal{X}, \mathbb{R})$, where $\mathcal{X}$ is a Borel subset of $\mathbb{R}^d$. Denote μ as the probability measure of X. Assume $g(X) = E(Y|X)$ to be the unknown conditional mean function and $\mathcal{G}_M = \{g_1, \ldots, g_M\}$ to be a finite dictionary of real-valued functions g_j that are defined on $\mathcal{X}$. For example, $\mathcal{G}_M$ can be a collection of basis functions used to approximate g in series estimation[22]. Then, for any $\beta = (\beta_1, \ldots, \beta_M) \in \mathbb{R}^M$, the penalized least squares (i.e., Lasso-type) estimator is:

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^M} \left\{ \frac{1}{n} \sum_{i=1}^{n} \{Y_i - g_\beta(X_i)\}^2 + 2 \sum_{j=1}^{M} \omega_{n,j} |\beta_j| \right\} \tag{15}$$

where $g_\beta(x) = \sum_{j=1}^{M} \beta_j g_j(x)$, $\omega_{n,j} = r_{n,M} ||g_j||_n$ and $||g_j||_n$ is the empirical L_2 norm of

[21] In a reference to Koltchinskii 2011, "an oracle inequality is a bound on the risk of a statistical estimator that shows that the performance of the estimator is almost (often, up to numerical constants) as good as it would be if the statistician had an access to an oracle that knows what the best model for the target function is".

[22] More specifically, with a relation to our own research, $\mathcal{G}_M$ can be seen as a collection of all terms in the multivariate sieve space we have created in (5), because those product-terms can also be viewed as basis functions for interactions among different variables. So, M, the length of dictionary, would then be $(K_n + 1)^d$ in our own research.

function g_j. The choice of tuning sequence $r_{n,M} > 0$ will be discussed in Section 4.2. Consequently, $\widehat{g} = \sum_{j=1}^{M} \widehat{\lambda}_j g_j$ is an estimate of g. Despite the fact that oracle properties of such a Lasso-type estimator have been well studied in linear parametric regression and fixed design non-parametric regression where M is fixed and $n \to \infty$, its oracle properties in random design non-parametric regression have only received little attention. To bridge this gap, Bunea, Tsybakov, and Wegkamp 2007 tries to establish oracle inequalities for the approximation error $||\widehat{g} - g||^2$ and the ℓ_1-loss $|\widehat{\beta} - \beta|_1$ in a random design non-parametric setting where $M = M(n)$ and possibly much larger than n.

4.2 Sparsity and Dimension Reduction

To introduce the concept of sparsity, let us consider a linear regression at first. The OLS estimator $\widehat{\beta}_{OLS}$ for the linear regression is almost the same as $\widehat{\beta}$ defined in (15) but without a penalty term. Obviously, the ℓ_1-loss $|\widehat{\beta}_{OLS} - \beta|_1$ is of order $M/\sqrt{n}$ in probability if $\widehat{\beta}$ is based on all M regressors, then the estimation error can be very large in case of high dimensionality (large M). This gives rise to the concept of sparsity.

Definition 4.1 (Sparsity)
Let $M(\beta)$ denote the number of non-zero coefficients of β, then

$$M(\beta) = \sum_{j=1}^{M} \mathbb{I}_{\{\beta_j \neq 0\}} = \text{Card} J(\beta) \tag{16}$$

where $\mathbb{I}_{\{\cdot\}}$ is indicator function and $J(\beta) = \{j \in \{1, \ldots, M\} : \beta_j \neq 0\}$.

The smaller $M(\beta)$, the sparser β. By taking advantage of sparsity, their paper shows that the ℓ_1-loss of lasso-type estimator $\widehat{\beta}$, i.e., $|\widehat{\beta} - \beta|_1$ is bounded, up to known constants and logarithms, by $M(\beta)/\sqrt{n}$, if in the penalty term, we have $r_{n,M} = A\sqrt{(\log M)/n}$ where A is a sufficient large constant. That means, the estimation error will be much smaller if β is much sparser.

It is worthy of mentioning that in Setup 4.1, g_β is not an exact representation of g, but a series approximation instead. Thus, the concept of sparsity in Definition 4.1 is too strong for this setup. Therefore, Bunea, Tsybakov, and Wegkamp 2007 introduce weak sparsity.

Definition 4.2 (Weak Sparsity)
Let $C_g > 0$ be a constant depending on g, and denote

$$\mathcal{B} = \{\beta \in \mathbb{R}^M : ||g_\beta - g||^2 \leq C_g r_{n,M}^2 M(\beta)\} \tag{17}$$

as the oracle set. Here, $||\cdot||$ is the $L_2(\mu)$-norm with probability measure μ, and $\langle f, g\rangle$ denotes the corresponding scalar product for any $f, g \in L_2(\mu)$. Then, if $\mathcal{B}$ is non-empty, g is said to have a weak sparsity property relative to the dictionary $\mathcal{G}_M = \{g_1, \ldots, g_M\}$.

With the definition of weak sparsity, one may have the belief that, for some $\beta^* \in \mathbb{R}^M$, the squared approximation error $||g_{\beta^*} - g||^2$ is bounded, up to logarithmic factors, by $M(\beta^*)/n$. The order of this approximation error is much smaller than the one without dimension reduction. Similar to the oracle vector β in case of linear regression, β^* is the oracle vector in this case of series approximation, and it is defined as:

$$\beta^* = \arg\min\{||g_\beta - g|| : \beta \in \mathbb{R}^M, M(\beta) = k^*\} \tag{18}$$

where $k^* = \min\{M(\beta) : \beta \in \mathcal{B}\}$ is called oracle dimension.

4.3 Assumptions

The authors explicitly make three assumptions, which are necessary for their theorem to hold. Following the notation in Setup 4.1, the error terms are defined to be $W_i = Y_i - g(X_i)$. Also, recall $g(X) = E(Y|X)$ is the unknown conditional mean function.

Assumption 4.1
$X_1, \ldots, X_n$ are independent, identically distributed random variables with probability measure μ. Random variables W_i are independently distributed with $E\{W_i|X_1, \ldots, X_n\} = 0$ and $E\{\exp(|W_i|)|X_1, \ldots, X_n\} \leq b$ for some finite $b > 0$ and $i = 1, \ldots, n$.

This is a standard assumption to ensure the statistical properties of an estimator. Zero conditional mean of W_i ensures the linear independence between W_i and X, so the expectation of W_i does not change with the values of X. Meanwhile, $E\{\exp(|W_i|)|X_1, \ldots, X_n\} \leq b$ guarantees the existence of moments for error terms.

Assumption 4.2
Define $||h||_\infty = \sup_{x \in \mathcal{X}} |h(x)|$ as sup norm for any bounded function h on $\mathcal{X}$.

(a) There exists $0 < L < \infty$ such that $||g_j||_\infty \leq L$ for all $1 \leq j \leq M$.
(b) There exists $c_0 > 0$ such that $||g_j|| \geq c_0$ for all $1 \leq j \leq M$.
(c) There exists $L_0 < \infty$ such that $E[g_i^2(X)g_j^2(X)] \leq L_0$ for all $1 \leq i, j \leq M$.
(d) There exists $L_* < \infty$ such that $||g||_\infty \leq L_* < \infty$.

This assumption imposes mild conditions on g and basis functions g_j, which are in the dictionary $\mathcal{G}_M$. It is noted that (a) implies (c) trivially. But, as L in (a) might be too large, they state (c) separately. The implication of (a) and (d) is that for any fixed $\beta \in \mathbb{R}^M$, there exists a positive constant $L(\beta)$, such that $||g - g_\beta||_\infty = L(\beta)$.

Assumption 4.3
For any $M \geq 2$, there exist constants $\kappa_M > 0$ such that $\Psi_M - \kappa_M \text{diag}(\Psi_M)$ is positive semi-definite, where Ψ_M is a $M \times M$ matrix given by

$$\Psi_M = (Eg_j(X)g_{j'}(X))_{1 \leq j,j' \leq M} = \left(\int g_j(x) g_{j'(x)} \mu(dx) \right)_{1 \leq j,j' \leq M} \tag{19}$$

Due to the positive semi-definiteness of $\Psi_M - \kappa_M \text{diag}(\Psi_M)$, they note that $0 < \kappa_M \leq 1$. Along with part (b) of Assumption (4.2), Assumption (4.3) implies that Ψ_M is positive definite, with the minimal eigenvalue τ bounded below by $c_0\kappa_M$.

4.4 Theorem of Sparsity Oracle Inequalities

With assumptions (4.1) - (4.3), the authors have proved the sparsity oracle inequalities for the lasso-type estimator $\widehat{\beta}$, as defined in (15).

Theorem 4.1
Assume that Assumptions (4.1) - (4.3) hold. Then, for all $\beta \in \mathcal{B}$ we have

$$\mathbb{P}\left\{||\widehat{g} - g||^2 \leq B_1 \kappa_M^{-1} r_{n,M}^2 M(\beta)\right\} \geq 1 - \pi_{n,M}(\beta) \tag{20}$$

$$\mathbb{P}\left\{|\widehat{\beta} - \beta|_1 \leq B_2 \kappa_M^{-1} r_{n,M} M(\beta)\right\} \geq 1 - \pi_{n,M}(\beta) \tag{21}$$

where $B_1 > 0$ and $B_2 > 0$ are constants depending on c_0 and C_g only and

$$\begin{aligned}\pi_{n,M}(\beta) \leq & 10M^2 \exp\left(-c_1 n \min\left\{r_{n,M}^2, \frac{r_{n,M}}{L}, \frac{1}{L^2}, \frac{\kappa_M^2}{L_0 M^2(\beta)}, \frac{\kappa_M}{L^2 M(\beta)}\right\}\right) \\ & + \exp\left(-c_2 \frac{M(\beta)}{L^2(\beta)} n r_{n,M}^2\right)\end{aligned} \tag{22}$$

for some positive constants c_1, c_2 depending on c_0, C_g and b only and $L(\beta) = ||g - g_\beta||_\infty$.

This theorem is very useful in the case of weak sparsity. By replacing β and $M(\beta)$ with β^* and $M(\beta^*)$ which are defined at the end of Section 4.2, Theorem 4.1 indicates that, under some assumptions, the squared approximation error of lasso-type estimator is controlled, up to a logarithmic factor, by a bound that only relates to the number of non-zero components of oracle vector, instead of all components. In other words, the sparser the oracle vector is, the smaller the squared error and the faster the convergence rate will be. This is conducive to the dimension adaptive properties of our estimator in Section 5.

5 Dimension Adaptive Estimation

In this section, we will apply Theorem 4.1 to show that a combination of sieve estimation and Lasso-type methods leads to an estimator, the convergence rate of which can adapt to the dimensionality of underlying true model, up to a logarithmic factor.

Section 5.1 restates and summarizes the steps of our dimension adaptive estimation, and in Section 5.2, we list three different types of underlying true models, on which we hope to examine the properties of our dimension adaptive estimation. Section 5.3 tailors assumptions (4.1) - (4.3) to fit the underlying models we consider. Section 5.4 closes this chapter with a theorem on the convergence rate of our dimension adaptive estimation.

5.1 Estimation Steps

There are generally two steps in dimension adaptive estimation. The first step is to construct a multivariate sieve space from the original group of independent variables by using a specific type of basis functions, with reference to (5). The total number of terms in the sieve should grow slowly with sample size to ensure consistency of sieve estimation, and in Section 5.2, we will discuss how it changes with sample size.

The second step is to apply Lasso-type methods as discussed in Chapter 3 to select significant terms in the constructed sieve, thus achieving dimension reduction without losing prediction accuracy. The Lasso-type methods are essential in finding the oracle dimension in underlying models. If a researcher hopes to make prediction on a new dataset, it is recommended to get estimated coefficients from running OLS (for linear models) on selected terms of Lasso-type methods, and use those coefficients for prediction.

5.2 Three Underlying Models

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of independent random pairs with $(X, Y) \in (\mathcal{X}, \mathbb{R})$, where $\mathcal{X}$ is a Borel subset of $\mathbb{R}^d$. Here, X is normalized such that $X_d \in [0, 1]$ for all $d \in \{1, 2, \ldots\}$. Consider the following model:

$$Y = g(X) + W, \quad E(W|X) = 0 \tag{23}$$

where W is the error term with zero conditional mean, then $E(Y|X) = g(X)$ is the unknown conditional mean function. Suppose $g(X)$ to be a p-smooth function with s-times continuous derivatives as defined in 2.4.

Let $\{p_k(\cdot)\}_{k=0}^{\infty}$ be basis functions for each dimension of X, i.e., each regressor. Then, we consider the product-terms $\{\prod_{l=1}^{d} p_{k_l}(x_l)\}_{k_l=0}^{\infty}$ to be basis functions for interactions among different regressors. In this thesis, we consider three different underlying models.

In model (a) and (b), we are dealing with a non-parametric model with infinite dimension, while in model (c), parametric model with finite dimension. Hopefully, our estimator can successfully adapt its convergence rate to all three types of underlying true models.

5.2.1 Model (a): Unrestricted Underlying Model

In this model, unknown parameters are assumed to lie in infinite-dimensional parameter spaces, so it would be computationally difficult to estimate such a model by using finite samples. Instead, we use sieve estimation, specifically series estimation. But to ensure consistency, the complexity of sieves should increase with the sample size so the number of basis functions for each regressor, K_n, grows slowly with sample size n. To put it formally, suppose there is a positive constant s and $\beta \in \mathbb{R}^{(K_n+1)^d}$ such that

$$\sup_{x\in\mathcal{X}} \left| g(x) - \sum_{k_1=0}^{K_n} \cdots \sum_{k_d=0}^{K_n} \beta_{k_1,\ldots,k_d} \prod_{l=1}^{d} p_{k_l}(x_l) \right| = \mathcal{O}(K_n^{-s}) \tag{24}$$

Sufficient conditions for this supremum consistency has been briefly discussed in Section 2.1, or one can refer to Chapter 3 of Chen 2007. From the above supremum norm of approximation bias, it is easy to show that the squared bias of series estimator is of order K_n^{-2s}. The variance of series estimator is of order K_n^d/n, as shown in Newey 1997.

The orders of both squared bias and variance involve K_n. It implies that the growth of K_n will reduce the squared bias but also increase variance. It is important to find a trade-off between them so that the optimal rate of K_n is achieved. In other words, we can choose K_n such that

$$K_n^{-2s} \asymp \frac{K_n^d}{n} \Leftrightarrow K_n \asymp n^{\frac{1}{2s+d}}$$

Therefore, $(K_n+1)^d$, the number of non-zero coefficients in model (a), is of order $n^{\frac{d}{2s+d}}$.

5.2.2 Model (b): Additive Underlying Model

In this model, unknown parameters are still assumed to be in infinite-dimensional spaces but with an additional assumption of additivity[23]. To ensure consistency, K_n still grows slowly with n. To put it formally [24], suppose there is a positive constant s and $\beta_1, \ldots, \beta_d \in \mathbb{R}^{K_n+1}$ such that

$$\sup_{x\in\mathcal{X}} \left| g(x) - \sum_{l=1}^{d} \sum_{k=0}^{K_n} \beta_{lk} p_k(x_l) \right| = \mathcal{O}(K_n^{-s}) \tag{25}$$

[23] It implies no interactions between different regressors

[24] Or, we can also use the formal representation from model (a), but with an additional restriction that $\sum_{l=1}^{d} \mathbb{I}_{\{}k_l \neq 0\} = 1$ or 0, where $\mathbb{I}_{\{\cdot\}}$ is indicator function.

Sufficient conditions for the supremum consistency can be found in Chapter 3 of Chen 2007. It is easy to find that the squared bias is of order K_n^{-2s}, while the variance is of order dK_n/n, as proven in Newey 1997. By following similar arguments as in model (a), we should choose K_n such that

$$K_n^{-2s} \asymp \frac{dK_n}{n} \Leftrightarrow K_n \asymp n^{\frac{1}{2s+1}}$$

Therefore, $d(K_n + 1)$, the number of non-zero coefficient in model (b), is of order $n^{\frac{1}{2s+1}}$.

5.2.3 Model (c): Parametric Underlying Model

In this model, unknown parameters are assumed to lie in a finite-dimensional space. By Stone-Weierstrass Theorem (Stone 1948), we know that the unknown function $g(X)$ can be uniformly approximated by basis functions. To put it formally, suppose there is $\beta \in \mathbb{R}^{(K+1)^d}$ for some fixed K[25] such that

$$\sup_{x \in \mathcal{X}} \left| g(x) - \sum_{k_1=0}^{K} \cdots \sum_{k_d=0}^{K} \beta_{k_1,\ldots,k_d} \prod_{l=1}^{d} p_{k_l}(x_l) \right| = 0 \tag{26}$$

That is, we are dealing with a flexible parametric model. K is finite and irrelevant to the sample size n. Apparently, the most usual parametric linear model falls into model (c).

5.3 Assumptions

In order to apply Theorem 4.1, we need to tailor assumptions (4.1) - (4.3) proposed in Section 4.3 to fit into the setting of our own underlying models.

Please be aware that the use of different basis functions may lead to different valuation of parameters (e.g., L, c_0, L_0 and Ψ_M) in assumptions. To make our statements more consolidated over different basis functions, we orthonormalize them before the theoretical analysis, no matter which basis function we use[26]. How those parameters may change if we don't orthonormalize them beforehand is briefly discussed at the end of Appendix 10.3.

Assumption 5.1
In accordance to Assumption (4.1), we assume:

(a) $\{(X_i, Y_i)\}_{i=1}^n$ is a sample of independent random pairs with zero conditional mean of error terms, i.e., $E(W|X) = 0$.
(b) $E(\exp(|W|)|X) \leq C_w$ for some constant $C_w > 0$.
(c) g is a Hölder (p-smooth) continuous function and its smoothness, s, is bounded below by $\underline{s}$, that is, $s \geq \underline{s}$.

[25]The fixed K for different regressors can be varied. It doesn't make a change to our analysis.
[26]If we use Legendre and trigonometric polynomials as basis functions, we rescale the domain of X_d to be $[-1, 1]$ for its good property of orthogonality within this domain.

(d) g is a bounded function on $\mathcal{X}$.

(e) There exists a constant $C > 0$ such that $||\prod_{l=1}^{d} p_{k_l}|| \leq C$.

Notice that, in the additive underlying model, K_n is of order $n^{\frac{1}{2s+1}}$. Thus, for the unknown function g with smoothness $s \geq \underline{s}$, the largest sieve space we use can be a linear span of basis functions up to an index sum of $C_s n^{\frac{1}{2\underline{s}+1}}$ with C_s being sufficiently large (e.g., $C_s = \log(n)$). This sieve space would contain the best approximation in additive underlying model. That is,

$$\prod_{l=1}^{d} p_{k_l}(x_l) : \quad \sum_{l=1}^{d} k_l \leq C_s n^{\frac{1}{2\underline{s}+1}} \tag{27}$$

Analogously, in unrestricted underlying model, the largest sieve space is a linear span of basis functions up to an index sum of $C_s n^{\frac{1}{2\underline{s}+d}}$ for C_s large enough. This upper bound is smaller than that of additive model with $d \geq 2$, therefore $C_s n^{\frac{1}{2\underline{s}+1}}$ could also be used as an upper bound for the unrestricted model. Likewise, this sieve space also contains the best approximation for parametric model because K is fixed. As a consequence, in the language of Bunea, Tsybakov, and Wegkamp 2007, we then have $M \lesssim n^{\frac{1}{2\underline{s}+1}}$.

Assumption 5.2

In accordance to Assumption (4.2), we need to impose mild conditions on g and on basis functions.

(a) There exists $0 < L < \infty$ such that $||\prod_{l=1}^{d} p_{k_l}||_\infty \leq L$ for all basis functions.

(b) There exists $c_0 > 0$ such that $||\prod_{l=1}^{d} p_{k_l}|| \geq c_0$ for all basis functions. Since our basis functions are orthonormal, $c_0 = 1/C$ with C from (e) of Assumption (5.1).

(c) From (a) and (b), we have

$$E\left[\prod_{l=1}^{d} p_{k_l}^2(x_l) \prod_{l=1}^{d} p_{j_l}^2(x_l)\right] \leq L^2 E\left[\prod_{l=1}^{d} p_{k_l}^2(x_l)\right] \leq L^2 C_l \left|\left|\prod_{l=1}^{d} p_{k_l}(x_l)\right|\right|^2$$

where C_l is a positive constant. The first inequality comes from $||\prod_{l=1}^{d} p_{k_l}||_\infty \leq L$. Due to the orthonormality of basis functions, we have $E\left[\prod_{l=1}^{d} p_{k_l}^2(x_l) \prod_{l=1}^{d} p_{j_l}^2(x_l)\right] \leq L_0$ with $L_0 = L^2$, ignoring the constant C_l.

(d) By assumption, g is a bounded function on $\mathcal{X}$, so L^* is obviously bounded above.

Assumption 5.3

To save notation, let $g_k(X) = \prod_{l=1}^{d} p_{k_l}(x_l)$ denote basis functions with $1 \leq k \leq M$, and $g^M(X) = (g_1(X), \ldots, g_M(X))^T$ is then the vector of all basis functions. Accordingly, $\Psi_M = E[g^M(X) g^M(X)^T]$ in our setting. In accordance to Assumption (4.3), for any $M \geq 2$, there exist constants $\kappa_M > 0$ such that $\Psi_M - \kappa_M \text{diag}(\Psi_M)$ is positive semi-definite. In fact, due to the orthonormality of basis functions we use, such a κ_M exists and equal to $1/C^2$, which is bounded away from 0. See Appendix 10.2 for proof.

5.4 Oracle Properties

5.4.1 Convergence Rates

With assumptions (5.1)-(5.3), Theorem 4.1 is then applied to obtain the optimal convergence rate of our estimator across all three underlying models we discuss in Section 5.2.

In Theorem 4.1, ignoring constants, the squared L_2 norm of Lasso-type estimator, $||\hat{g} - g||^2$ is of order $r_{m,M}^2 M(\beta)$, as long as $\pi_{n,M}(\beta)$ is asymptotically zero.

(a) In unrestricted underlying model, the number of non-zero coefficients $(K_n + 1)^d$ is of order $n^{\frac{d}{2s+d}}$. In the language of Bunea, Tsybakov, and Wegkamp 2007, $M(\beta) = C_s n^{\frac{d}{2s+d}}$. Also, as suggested in Definition 4.1, $r_{n,M} = A\sqrt{(\log M)/n}$. Then,

$$r_{n,M}^2 M(\beta) = A^2 \frac{\log M}{n} C_s n^{\frac{d}{2s+d}} = A^2 C_s \log M n^{-\frac{2s}{2s+d}}$$

Hence, the optimal convergence rate of our estimator is $n^{-\frac{2s}{2s+d}}$, up to a logarithmic factor, in unrestricted underlying model.

(b) In additive underlying model, the number of non-zero coefficients $d(K_n + 1)$ is of order $dn^{\frac{1}{2s+1}}$ with d being a constant. That means, $M(\beta) = C_s n^{\frac{1}{2s+1}}$. Then,

$$r_{n,M}^2 M(\beta) = A^2 \frac{\log M}{n} C_s n^{\frac{1}{2s+1}} = A^2 C_s \log M n^{-\frac{2s}{2s+1}}$$

Hence, the optimal convergence rate of our estimator is $n^{-\frac{2s}{2s+1}}$, up to a logarithmic factor, in additive underlying model.

(c) In parametric underlying model, the number of non-zero coefficients is $(K + 1)^d$, which is finite. That means, $M(\lambda) = (K + 1)^d$. Then,

$$r_{n,M}^2 M(\beta) = A^2 \frac{\log M}{n} (K+1)^d = A^2 (K+1)^d \log M n^{-1}$$

Hence, the optimal convergence rate of our estimator is n^{-1}, up to a logarithmic factor, in parametric underlying model.

To summarize, the theorem on convergence rate of our estimator is given as follows.

Theorem 5.1
Suppose that Assumptions (5.1) - (5.3) hold, and $\pi_{n,M}(\beta)$ in (22) is asymptotically zero. Then, the convergence rate of our estimator in squared L_2 norm is dimension adaptive with respect to underlying true models. Specifically, in unrestricted underlying model (5.2.1), it converges at an optimal rate of $n^{-\frac{2s}{2s+d}}$; in additive model (5.2.2), $n^{-\frac{2s}{2s+1}}$; and in parametric model (5.2.3), n^{-1}. All convergence rates are up to a logarithmic factor.

It is seen that our dimension adaptive estimator can achieve an optimal convergence rate, regardless of the underlying true models, as if it were an "oracle" estimator. It

converges in all three underlying models, even though with a different rate. Specifically, It converges as fast as a parametric estimator in parametric underlying model. In additive underlying model, the convergence rate is slower, and in unrestricted underlying model, it has the slowest convergence rate. In the latter two models, a parametric estimator will not convergence in probability to the true function.

5.4.2 Lower Bound of Smoothness

Theorem 5.1 holds true only if $\pi_{n,M}(\beta)$ defined in (22) is asymptotically zero. It is noted that $\pi_{n,M}(\beta)$ is related to parameters L and L_0 which may change with the basis functions we use. For example, for unbounded basis functions, L and L_0 might increase with sample size n at some rate related to smoothness s, while for bounded basis functions, L and L_0 are constants, irrelevant to n. Moreover, $M(\beta)$ in the inequality of $\pi_{n,M}(\beta)$ also relates to sample size n and smoothness s.

Therefore, the constraint on $\pi_{n,M}(\beta)$ will finally lead to a constraint on the lower bound of smoothness, and this lower bound is different for different basis functions. We thus have the following corollary, and see Appendix 10.3 for proof.

Corollary 5.1
Theorem 5.1 holds true only if the smoothness s of the unknown regression function is bounded below by $\underline{s}$. This lower bound varies among different basis functions that we use to generate sieve for our dimension adaptive estimation.

(a) For normalized Legendre polynomials, $\underline{s} > \frac{(2d-1)+\sqrt{8d^2+1}}{4}$, or a sufficient condition $\underline{s} \geq \frac{3d-1}{2}$.
(b) For orthonormalized B-splines, $\underline{s} > \frac{d}{2}$, or a sufficient condition $\underline{s} \geq \frac{d+1}{2}$.
(c) For normalized Haar wavelets, the same as (a).
(d) For normalized trigonometric polynomials, same as (b).

As is mentioned in Section 5.3, to make our statement more consolidated, all basis functions are orthonormalized before analysis. Otherwise, their boundedness may change. Legendre polynomials, Haar wavelets and trigonometric polynomials on a specific domain are orthogonal by design, so a further normalization is enough. From Corollary 5.1, it is seen that the lower bound of smoothness is more relaxed for orthonormalized B-splines and normalized trigonometric polynomials than the others.

It is worthy of mentioning that the result in Corollary 5.1 is asymptotic. Even though the lower bound of smoothness for orthonormalized B-splines and normalized trigonometric polynomials are much more relaxed, it doesn't mean that they always outperform normalized Legendre polynomials and normalized Haar wavelets in finite samples.

6 Simulation Study

6.1 Simulation Setup

To illustrate how our dimension adaptive estimator behaves in finite samples, we apply it to simulated data and compare the results with "oracle" estimators that already know the dimensionality of underlying true models. According to the three types of underlying models discussed in Section 5.2, the following three functions are used in the illustrative simulated settings.

$$m_1(x) = 3x_1 + 1.8x_2 + x_3 + 2.5x_4 + x_5 \qquad (x \in [0,1]^5)$$

$$m_2(x) = \sin(4x_1) + 1.5\log(x_2) + \frac{1}{\cos(x_3)} + \sin(\sqrt{x_4}) + \sin(x_5^2) \qquad (x \in [0,1]^5)$$

$$m_3(x) = 3\sqrt[4]{x_1 + 4x_2 + x_3x_4x_5} + 2\sin(x_4 + x_5^2 + x_1x_2x_3) + 3\log(x_3^2 + x_4 + 2x_5) \qquad (x \in [0,1]^5)$$

Among these functions, m_1 represents a parametric underlying model; m_2, an additive non-parametric model and m_3, an unrestricted non-parametric model. As shown in Theorem 5.1, the dimension adaptive estimator can adapt its convergence rate to the unknown dimension of underlying model. That is, in model m_1, this estimator should converge as fast as a parametric estimator; in m_2, it should then be as good as a non-parametric estimator with additivity restriction; finally in m_3, it should outperform the other two estimators.

In our simulation, the dimension adaptive estimator (*dimada*) is the sieve estimator with Lasso-type regularization as defined in Section 5.1, and the non-parametric estimator with additivity restriction (*addt*) is an additive dimension adaptive estimator. The parametric estimator is just an ordinary least squares (*ols*) linear estimator.

The n observations of type (X, Y) are generated with the following data generation process:

$$Y = m_i(X) + \sigma_j \cdot \epsilon \qquad (i \in \{1,2,3\}, j \in \{1,2\})$$

where X is uniformly distributed on $[0,1]^5$ and ϵ is standard normally distributed and independent of X. The parameters scaling the noise are $\sigma_1 = 5\%$ and $\sigma_2 = 20\%$.

Empirical squared L_2 error is applied in our simulation to examine the performance of estimators, so it is bounded below by the variance of error term (i.e., σ_j^2). Specifically, we use out-of-sample empirical mean squared error (MSE) to measure the approximation error. In view of the fact that simulation results rely on randomly drawn data points, we compute out-of-sample empirical MSE for 500 repeatedly generated realizations of X and examine the average and median of all MSEs. In the main results of our simulation, the

numbers of observation in train and test datasets are 400 and 1000, respectively. Four types of basis functions are used, including power series, Legendre polynomials, B-Splines and trigonometric polynomials. Two Lasso-type methods are used, including Lasso and adaptive Lasso, as discussed in Section 3.

To compute out-of-sample empirical MSE for *dimada*, we firstly construct sieves of regressors from original train and test dataset, using the same set of basis functions. Afterwards, we use Lasso and adaptive Lasso to select significant terms of sieve in train dataset. Notice that Lasso and adaptive Lasso can correctly recover the "right" sparsity of underlying models, so it is a common practice to simply perform post Lasso and post adaptive Lasso, which literally run OLS on the selected terms. Then, by using the estimated coefficients from post-selection methods, we compute out-of-sample empirical MSEs in the separate test dataset. All procedures also apply to the estimator *addt* except that the sieves are constructed with an additionally additive restriction, i.e., no interaction among regressors.

6.2 Out-of-sample Empirical MSE

The results of average out-of-sample empirical MSEs for all three underlying models are presented in Tables 1, 2 and 3, respectively. The smallest MSEs for each type of basis functions are in bold.

Table 1: Average of out-of-sample empirical MSEs for m_1

		$\sigma_1 = 5\%$				$\sigma_2 = 20\%$			
		Post LASSO		Post Adaptive LASSO		Post LASSO		Post Adaptive LASSO	
Basis	Estimator	MSE	Terms[c]	MSE	Terms[c]	MSE	Terms[c]	MSE	Terms[c]
Power Series	*dimada*	0.002631	17.610	0.002544	5.000	0.04185	16.854	0.04084	5.228
	addt	0.002562	6.944	0.002544	5.000	0.04100	7.922	0.04071	5.004
	ols[b]	**0.002544**	5.000	**0.002544**	5.000	**0.04070**	5.000	**0.04070**	5.000
Legendre	*dimada*	0.002627	16.838	0.002544	5.000	0.04194	16.122	0.04085	5.276
	addt	0.002559	6.770	0.002544	5.000	0.04113	7.790	0.04070	5.000
	ols[b]	**0.002544**	5.000	**0.002544**	5.000	**0.04070**	5.000	**0.04070**	5.000
B-Splines	*dimada*	0.029481[a]	188.392	0.010295	106.384	0.37211[a]	129.458	0.09397	94.110
	addt	0.002680	27.792	0.002707	25.020	0.04288	27.786	0.04318	24.740
	ols[b]	**0.002544**	5.000	**0.002544**	5.000	**0.04070**	5.000	**0.04070**	5.000
Trigonometric	*dimada*	0.006355	146.906	0.005549	21.278	0.05643	44.192	0.05526	33.322
	addt	0.006144	13.130	0.006280	9.896	0.04410	19.804	0.04401	11.616
	ols[b]	**0.002544**	5.000	**0.002544**	5.000	**0.04070**	5.000	**0.04070**	5.000

[a] The performance of *dimada* with B-Splines is not satisfactory due to some extreme values in estimated coefficients from Post LASSO. The influence is reduced in Table 7 when we use median MSEs.
[b] The *ols* estimator doesn't involve any basis function or Lasso-type methods.
[c] Average number of selected terms with non-zero coefficients.

Table 2: Average of out-of-sample empirical MSEs for m_2

		$\sigma_1 = 5\%$				$\sigma_2 = 20\%$			
		Post LASSO		Post Adaptive LASSO		Post LASSO		Post Adaptive LASSO	
Basis	Estimator	MSE	Terms[c]	MSE	Terms[c]	MSE	Terms[c]	MSE	Terms[c]
Power Series	*dimada*	0.17800	37.96	0.17098	18.364	0.2211	38.13	0.2127	19.384
	addt	**0.15299**	13.05	**0.15748**	8.818	**0.1917**	13.16	**0.1957**	9.016
	ols[b]	0.75747	5.00	0.75747	5.000	0.79552	5.00	0.79552	5.000
Legendre	*dimada*	0.17885	40.95	0.17243	22.350	0.2224	40.99	0.2153	23.158
	addt	**0.15436**	12.06	**0.15742**	9.928	**0.1929**	12.18	**0.1960**	9.960
	ols[b]	0.75747	5.00	0.75747	5.000	0.7955	5.00	0.7955	5.000
B-Splines	*dimada*	1.936*e*07[a]	74.66	0.19182	45.806	357.5585[a]	71.01	0.2617	47.866
	addt	**0.09209**	26.82	**0.09298**	22.004	**0.1321**	26.61	**0.1329**	21.670
	ols[b]	0.75747	5.00	0.75747	5.000	0.7955	5.00	0.7955	5.000
Trigonometric	*dimada*	1.31567[a]	164.72	0.19406	101.924	0.2993	124.62	0.2751	78.680
	addt	**0.07343**	23.74	**0.07592**	18.786	**0.1139**	23.96	**0.1167**	19.036
	ols[b]	0.75747	5.00	0.75747	5.000	0.7955	5.00	0.7955	5.000

[a] See Table 8 for median MSEs which are not influenced by extreme values.
[b] The *ols* estimator doesn't involve any basis function or Lasso-type methods.
[c] Average number of selected terms with non-zero coefficients.

Table 3: Average of out-of-sample empirical MSEs for m_3

		$\sigma_1 = 5\%$				$\sigma_2 = 20\%$			
		Post LASSO		Post Adaptive LASSO		Post LASSO		Post Adaptive LASSO	
Basis	Estimator	MSE	Terms[c]	MSE	Terms[c]	MSE	Terms[c]	MSE	Terms[c]
Power Series	*dimada*	**0.04031**	33.796	**0.04575**	23.850	**0.08123**	34.40	**0.08670**	24.180
	addt	0.26063	10.960	0.26206	9.082	0.29934	10.96	0.30077	9.076
	ols[b]	0.31930	5.000	0.31930	5.000	0.35748	5.00	0.35748	5.000
Legendre	*dimada*	**0.04064**	34.482	**0.04563**	24.304	**0.08199**	35.31	**0.08635**	24.756
	addt	0.26164	9.946	0.26215	8.996	0.30030	10.01	0.30100	8.974
	ols[b]	0.31930	5.000	0.31930	5.000	0.35748	5.00	0.35748	5.000
B-Splines	*dimada*	15.73708[a]	174.894	**0.10349**	115.158	204.11388[a]	131.12	**0.20356**	97.996
	addt	**0.26655**	27.340	0.26760	23.534	**0.30671**	27.32	0.30805	23.404
	ols[b]	0.31930	5.000	0.31930	5.000	0.35748	5.00	0.35748	5.000
Trigonometric	*dimada*	**0.03586**	206.714	**0.03573**	117.976	**0.11705**	128.23	**0.11137**	90.098
	addt	0.27013	21.462	0.26944	15.798	0.31003	21.57	0.30915	15.780
	ols[b]	0.31930	5.000	0.31930	5.000	0.35748	5.00	0.35748	5.000

[a] See Table 9 for median MSEs which are not influenced by extreme values.
[b] The *ols* estimator doesn't involve any basis function or Lasso-type methods.
[c] Average number of selected terms with non-zero coefficients.

In Table 1, all average MSEs are above the variance (σ_j^2) of error term, since they are empirical MSEs. Also, it is observed that *ols* estimator always achieve the best approximation. It is because m_1 is a parametric linear model. In this sense, *ols* is an estimator that already knows the dimension of underlying model, so it should have the best estimation. It worths noting that estimators *dimada* and *addt* have a performance almost as good as *ols*, particularly when basis functions are power series or Legendre polynomials. The good performance of *dimada* indicates that our dimension adaptive estimator can achieve a convergence rate as fast as *ols* in a parametric underlying model.

In Table 2, we find that *ols* estimator has the worst approximation while *addt* approximates the best. m_2 is an additive non-parametric model with infinite underlying dimension and hence the parametric estimator *ols* is unable to work well. In this occasion, *addt* is the estimator that knows the true dimension of underlying model. It is again noted that *dimada* obtains very close approximation errors to *addt*, which further substantiate our theoretical finding that *dimada* converges as fast as an "oracle" estimator.

In Table 3, when the underlying model is non-parametric with no other restrictions, we discover that *dimada* outperforms the other two estimators by a lot, because *ols* is a parametric estimator and *addt* fails to consider the interactions among regressors. What is noteworthy is that for B-Splines, the average MSEs of *dimada* seem to be heavily influenced by extreme values in estimated coefficients from Post Lasso. Such an influence is gone in Table 9 when we compute median MSEs instead.

To sum up, the simulation results support the theoretical finding in Section 5. Our dimension adaptive estimator is weakly dominant in terms of squared L_2 error across all underlying models we consider. The estimator *dimada* always converges as fast as an "oracle" estimator that knows the underlying dimensionality.

6.3 Convergence Rates

Section 6.2 compares the performance of our *dimada* estimator and two other estimators in three distinct underlying models m_1, m_2 and m_3. In this section, we focus on our *dimada* estimator and hope to further compare its specific convergence rates across these three underlying models. As given in Theorem 5.1, our dimension adaptive estimator has a fast parametric rate in a parametric model (m_1); a slower rate in an additive non-parametric model (m_2) and the slowest rate in an unrestricted non-parametric model (m_3).

To compare the convergence rates of *dimada* in all three models, we firstly calculate empirical out-of-sample MSEs by using increasing sample sizes of train dataset from 100 to 450 with a step of 50. The size of test dataset remains the same, i.e., 1000. Legendre polynomials are used as basis functions in this section. To account for the randomness of

a single realization of dataset X, we again calculate empirical out-of-sample MSEs for 500 independent repetitions of X, and analyze the average and median of all MSEs instead. Most importantly, in Theorem 4.1, true squared L_2 error ($||\widehat{g} - g||^2$) is used. Thus, we need to deduct the variance of error term (σ_j^2) from the average or median empirical MSEs we have obtained. The remaining portion is the true MSEs we target for. Finally, with the true MSEs at sample size of 100 as a base, we compute the ratios of true MSEs at other sample sizes to that base, and plot them out for different σ_j and Lasso-type methods. Figure 3 depicts the ratios from using average MSEs for each pair of σ_j and Lasso-type methods. See Appendix 10.6 for plots of ratios from using median MSEs.

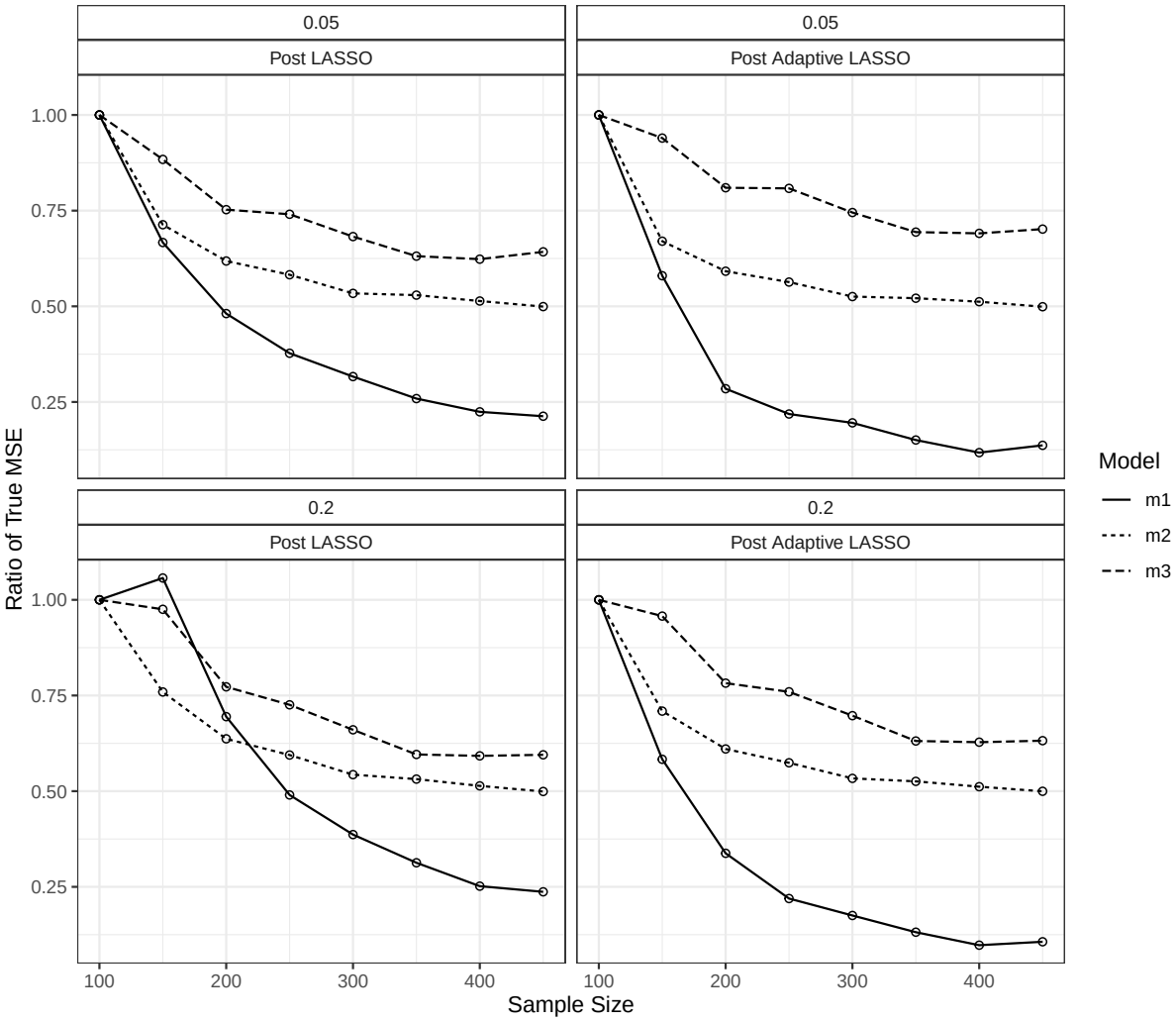

Figure 3: Comparison of convergence rates (using average MSEs)

In all panels of Figure 3, it is observed that the ratios start at 1 and go downward with the increase of sample size, except for the case of m_1 and sample size 150 in the panel of $\sigma_2 = 20\%$ and post Lasso. This exception is assumed to originate from the influence of some extreme values. Such an influence is reduced in Figure 5 when we use median MSEs instead. The downward ratios indicate that true MSE decreases with sample size, which is undoubtedly expected.

Furthermore, it is obvious that in all panels, the ratios decrease the fastest for m_1, slower for m_2 and the slowest for m_3. This finding strongly corroborates Theorem 5.1 that our dimension adaptive estimator can always have the optimal convergence rate across different underlying models. In parametric underlying model (m_1), it has an optimal convergence rate of n^{-1}; in additive model (m_2), $n^{-\frac{2s}{2s+1}}$; and in unrestricted model (m_3), $n^{-\frac{2s}{2s+d}}$. All rates are up to an logarithmic factor. Actually, not only can we compare convergence rates across different underlying models, we are also able to verify the convergence rate for each model with the use of some data points in Figure 3.

Firstly, it is already known that in parametric underlying model (m_1), if the sample size grows from n to $2n$, true MSE will be halved[27]. That is, $\text{MSE}(2n)/\text{MSE}(n) = (2n)^{-1}/n^{-1} = 1/2$. In the top-left panel of Figure 3, it is clearly seen that when sample size increases from 100 to 200, the ratio decreases from 1 to 0.5; and when the sample size further increases to 400, the ratio is further halved, to around 0.25. This finding still holds in other three panels even with a little bit deviation. Since the evidence of convergence rate in parametric model at the top-left panel is very strong, we will focus on that panel for verification of convergence rates in other two models.

It is further observed that for additive model (m_2), the ratio at sample size of 200 is around 0.625. That means, $\text{MSE}(200)/\text{MSE}(100) = (200)^{-\frac{2s}{2s+1}}/(100)^{-\frac{2s}{2s+1}} = 2^{-\frac{2s}{2s+1}} \approx 0.625$. Then, we have $s \approx 1.05314$. With this approximation of s, we can verify any other point for m_2. For example, the observed ratio for sample size of 150 is around 0.75 on the plot. With the $s \approx 1.05314$, we have approximated ratio $\text{MSE}(150)/\text{MSE}(100) = 1.5^{-\frac{2s}{2s+1}} \approx 1.5^{-0.67807} \approx 0.75962$. The approximated ratio coincides with the observed one, thus supporting the convergence rate in additive model, as found in our theorem.

Finally, for unrestricted model (m_3), the ratio at sample size of 200 is observed to be about 0.75. It means, $\text{MSE}(200)/\text{MSE}(100) = (200)^{-\frac{2s}{2s+5}}/(100)^{-\frac{2s}{2s+5}} = 2^{-\frac{2s}{2s+5}} \approx 0.75$. Then, we have $s \approx 1.77377$. Likewise, with this approximation of s, we can verify any other point for m_3. For instance, the observed ratio for sample size of 150 is about 0.875 on the plot, and with $s \approx 1.77377$, we have approximated ratio $\text{MSE}(150)/\text{MSE}(100) = 1.5^{-\frac{2s}{2s+5}} \approx 1.5^{-0.41504} \approx 0.84512$. Again, observed and approximated ratios coincide roughly. This substantiates our theoretical finding of convergence rate in unrestricted model.

[27]We don't need to consider the constant in the rate, because it is generally the same for different sample size.

7 Application

To make our dimension adaptive estimator much easier for empirical researchers to implement, we developed an R package dimada which can be conveniently installed locally. Not only does it contain functions to generate sieves from original dataset and implement our estimator, but also it has functions to plot out and summarize the estimation results. In this section, with the R package, we perform our estimator on a real-life dataset and compare the out-of-sample empirical MSE with that of an OLS linear estimator (ols) and an additive estimator ($addt$).

7.1 Data

The dataset consists of selling price and multiple attributes of cars. Our aim is to make prediction of car price based on the available attributes[28]. The dataset is sourced from a data science online comminity, Kaggle. Table 4 presents the descriptive statistics of our dataset.

Table 4: Descriptive statistics

	Description	Type	Mean	St. Dev.	Min	Median	Max
price	selling price (in £)	cont.[a]	23,469.940	16,406.720	450	18,999	145,000
year	production year	cont.	2,016.738	2.884	1,997	2,017	2,020
mileage	mileage	cont.	24,956.290	24,443.330	1	19,000	259,000
tax	sale tax (in £)	cont.	152.333	82.404	0	145	580
mpg	miles per gallon	cont.	50.371	35.747	2.800	47.100	470.800
engine	size of engine (in litres)	cont.	2.124	0.789	0.000	2.000	6.600

[a] "cont." stands for "continuous".

The total number of observations in this dataset is 4960, and there is no missing data. To compute out-of-sample empirical MSE, we split the original dataset into a train set (60%, 2976 observations) and a test set (40%, 1984 observations). As shown in Table 4, *price* is the dependent variable and all other variables are regressors.

7.2 Procedures

To apply our dimension adaptive estimator, it is worthy of mentioning that rescaling original regressors is required before sieve generation. The rule of variable selection in Lasso-type methods pertains to the magnitude of variables. In order to remove such an

[28] We excluded four binary variables *diesel*, *electricity*, *hybrid* and *petrol* from original dataset. These variables include the information whether a car uses fuel or the specific resource. They are excluded because B-Spline sieves of these variables lead to some extreme values in estimated coefficients, which then influence out-of-sample MSEs. For sieves of other basis functions, there is no such influence. We computed out-of-sample MSEs with power series sieves and Legendre sieves of all variables from original dataset, and the results are consistent with those in Table 5.

unfair influence, regressors have to be rescaled before analysis of Lasso[29]. See Appendix 10.1 for detailed discussions. Obviously, original regressors are not on the same scale, as seen in Table 4. In this section, we use z-score standardization to rescale original variables.

With rescaled variables, we first generate sieves of regressors from a specific type of basis functions. Four types of basis functions are taken into account: power series, Legendre polynomials, B-splines and trigonometric polynomials. We consider interactions among all regressors when constructing sieves, so the maximum number of interacting regressors in a single term of sieves is 5. Usually, the number of basis functions generated for each single regressor is calculated directly from formula (27), where the lower bound of smoothness for different basis functions can be obtained from Corollary 5.1. It then means 5 basis functions for each regressor when using power series or Legendre polynomials and 10 when using B-splines or trigonometric polynomials in this specific example. Nonetheless, in order to achieve better series approximation, we arbitrarily increase the number of basis functions for each regressor to 20[30], instead of using the relatively low values computed directly from formula 27.

Then, Lasso-type methods are performed on the generated sieves, and significant terms are selected. Finally, to follow a common practice, post-selection OLS linear models are estimated on those selected terms, from which estimated coefficients can is used to compute out-of-sample MSEs on the separate test dataset.

7.3 Results

Figure 4 portrays how cross-validated MSE changes with different number of non-zero coefficients for Lasso and adaptive Lasso within the train dataset. This is an example from using basis of trigonometric polynomials. The U-shape of curves signifies a trade-off between bias and variance of series approximation, and the shape of curves from using other basis functions is similar. As mentioned at the end of Section 3.1, cross validation is commonly used to choose the "best" regularization parameter λ, because it is pre-defined and researchers usually don't know how large it should be. As is indicated in Figure 4, the point with smallest cross-validated MSE is chosen, and the corresponding λ and terms are selected. The number of selected terms is relatively large for basis trigonometric polynomials, which are bounded basis functions.

[29]In our package dimada, we use glmnet package to implement Lasso-type methods, which can actually be configurated to standardize regressors inside the command itself, so users usually do not need to standardize regressors beforehand. However, in our case, we have a separate test dataset which does not go through package glmnet, so it would be safer to rescale all regressors beforehand. It is unnecessary to standardize the dependent variable. But, the scale of dependent variable in this example is too large, we rescale it to have a small scale of errors for comparison.

[30]The number of basis functions for each regressor when using B-splines still remains to be 10, because in finite samples, higher number of B-splines doesn't necessarily mean better series approximation. If the number is too large, there will be some B-splines with only a few observations, leading to high approximation bias.

Actually, the cross-validated empirical MSEs in Figure 4 are also out-of-sample MSEs which can be used for model evaluation, but the number of folds in cross-validation is 10, so the out-of-bag sample is of relatively small size ($= 2976 \times 0.1 \approx 298$). Therefore, a separate test dataset of larger size (=1984) is employed to examine the out-of-sample performance of our dimension adaptive estimator.

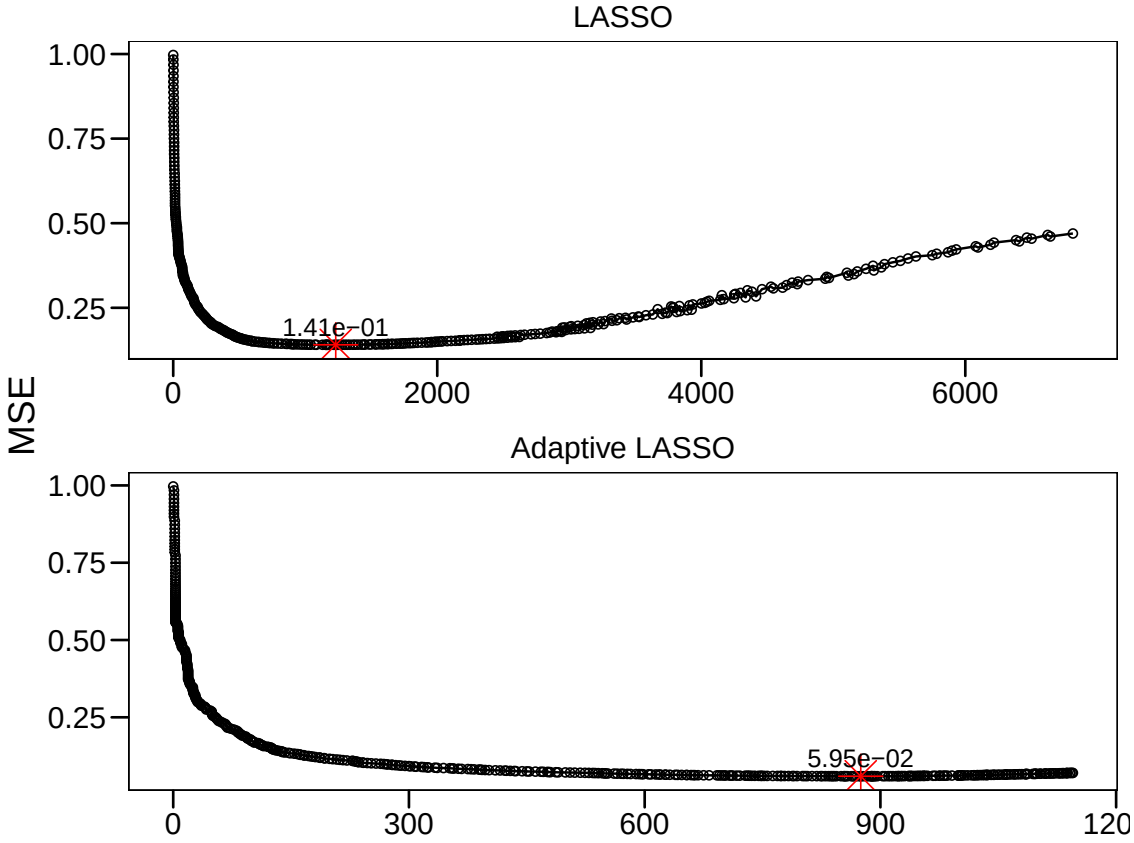

Figure 4: Cross-validated empirical MSE

In Table 5, the out-of-sample MSEs of our dimension adaptive estimator (*dimada*) on the separate test dataset are displayed, together with that of a simple OLS linear estimator (*ols*) and an additive estimator (*addt*). The definition and abbreviation of these estimators are consistent with those in Section 6. In the last column of Table 5, we compute the ratio of each MSE to the MSE of OLS estimator.

It is observed that the out-of-sample MSEs of estimator *addt* are smaller than that of an OLS linear estimator, and the range of ratio is between 72.40% to 85.57%, ignoring the outlier. This is because OLS estimator only considers linear relationship between regressors and the dependent variable while *addt* estimator further considers non-linearity, even though both of them fail to include interactions among regressors.

Moreover, the out-of-sample MSEs of our estimator *dimada* are much smaller than that of *ols*, and the ratio ranges from 56.39% to 72.75%. The performance of estimator *dimada* varies greatly with the choice of basis functions. In this example, B-splines

and trigonometric polynomials seem to perform better than power series and Legendre polynomials. For the same basis function, performance of *dimada* is always better than *addt*. This indicates that the underlying model of this dataset is very likely not linearly parametric or additive, but non-parametric with interactions among regressors.

Table 5: Out-of-sample MSEs

Estimator	Basis	Method	MSE	MSE/MSE(*ols*)
ols		Linear OLS	0.3181	100%
addt	Power Series	Post LASSO	0.2719	85.48%
		Post Adaptive LASSO	0.2722	85.57%
	Legendre Polynomials	Post LASSO	0.2719	85.47%
		Post Adaptive LASSO	0.3232[a]	101.61%[a]
	B-Splines	Post LASSO	0.2303	72.40%
		Post Adaptive LASSO	0.2307	72.52%
	Trigonometric Polynomials	Post LASSO	0.2638	82.92%
		Post Adaptive LASSO	0.2636	82.87%
dimada	Power Series	Post LASSO	0.2299	72.27%
		Post Adaptive LASSO	0.2296	72.18%
	Legendre Polynomials	Post LASSO	0.2314	72.75%
		Post Adaptive LASSO	0.2296	72.18%
	B-Splines	Post LASSO	0.1964	61.74%
		Post Adaptive LASSO	0.2011	63.22%
	Trigonometric Polynomials	Post LASSO	0.1897	59.65%
		Post Adaptive LASSO	0.1794	56.39%

[a] This MSE may be affected by some extreme values in the estimated coefficients of basis functions.

7.4 Practical Suggestions

From the results of simulation and application, we summarize some suggestions for researchers who wish to apply our dimension adaptive estimator in their own studies.

In the first step of generating a multivariate sieve space,

(a) Try different types of basis functions.
This can be configurated through the argument *basis* in our R package dimada, where there are 8 types of basis functions to choose from. There is no specific type that always works better than others, particularly in finite samples.

(b) Try different number of basis functions generated for each single regressor.
This corresponds to the argument *n.basis* in our R package. By default, the formula (27) is used to compute the number of basis functions generated for each regressor. Usually, larger the number, better the series approximation. However, in finite

samples, this may be violated, particularly for B-splines. If the number of knots for B-splines is too large, there might be some intervals with only a few observations, which result in poor approximation.

(c) **Consider interactions among regressors, unless it is certain that the model is additive**.

This can be set with the argument *max.interaction* in the R package. Economic variables are always entangled with other, so it is a good idea to consider interactions among them. When the number of regressors is not large, one can consider interactions among all variables. Nevertheless, if there are too many regressors, it is adviced to control the maximal number of interacting regressors in a single term under a certain value, such as 5 in order to expedite computation.

(d) **Rescale all regressors with z-score standardization beforehand**.

This is required if the original dataset is split into train and test datasets or there is a new dataset on which prediction needs to be made. Dependent variable doesn't need to be rescaled but if it is, remember to reverse it back to original scale after prediction.

In the second step of applying Lasso-type methods,

(a) **Use cross validation to select the "best" regularization parameter λ**

In Lasso-type methods, instead of defining the regularization parameter λ discretionarily, cross validation is usually used to choose the "best" λ that gives the smallest out-of-bag MSE. This is also by default in our R package.

(b) **Use post-selection methods to obtain estimated coefficients for prediction, if necessary**.

Since Lasso and adaptive Lasso are found to correctly recover the "right" sparsity of underlying model, we can estimate the model well in ℓ_2-norm simply by doing an OLS (for linear model) restricted to the selected subset. It is a common practice to run the post-selection methods on chosen terms from Lasso-type methods. However, one needs to be cautious about doing inference with post-selection methods because they usually fail to provide uniform inference. Belloni, Chernozhukov, and C. Hansen 2013 develop a "post-double-selection" method which has a property of uniform inference, but it is out of the scope of this thesis.

8 Conclusion

This thesis proposes a dimension adaptive estimator which can achieve an optimal convergence rate across all types of underlying models, inclusive of parametric, additive and unrestricted non-parametric models. Specifically, in parametric underlying model, it converges as fast as a parametric estimator, usually at a rate of n^{-1}, up to logarithmic factor. In additive and unrestricted non-parametric models, its convergence rate is slower but it still converges while parametric estimators do not converge to the true function. Hence, this estimator can perform as well as an "oracle" estimator in all underlying true models.

This estimator is constructed with a two-step approach, as describe in Section 5.1. To put it simply, the first step is to generate a multivariate sieve space from original regressors; the second step is to select significant terms in the sieve by Lasso-type methods. By applying a theorem in Bunea, Tsybakov, and Wegkamp 2007, the oracle properties of the dimension adaptive estimator is established in Theorem 5.1. Also, the required lower bound of smoothness for the unknown conditional mean function is different across various basis functions. This is summarised in Corollary 5.1.

The oracle properties of dimension adaptive estimator is strongly supported in the results of simulation. Furthermore, this estimator is applied in a real-life dataset about the prediction of car price. It is found that its performance is much better than an OLS linear estimator or an additive estimator in terms of out-of-sample MSE. Finally, practical suggestions are listed for researchers who are interested in using this estimator.

This thesis has several limitations which may be considered in future research. First, the convergence rate usually has two portions multiplying together. The central portion is something related to the sample size n, such as n^{-1}. The other portion is a constant which is relative to smoothness s and dimension d of original data X, but not associated with n. However, in high dimensional analysis, d sometimes grows with n, thus relating the "constant" portion to n. This may change the convergence rates discovered in Theorem 5.1. It can be an interesting extension of the thesis. Second, cross validation is suggested in this thesis to select the "best" λ for Lasso-type methods. However, the process of cross validation involves randomness in splitting data, which may be unwanted. Also, if the sample size is large, it may then be time-consuming to run cross validation. So, an alternative is to try Bayesian Information Criterion (BIC) instead. The performance of BIC in this research awaits to be tested. Last but not least, other types of basis functions needs to be discussed. In particularly, it is unknown how the lower bound of smoothness is defined for other basis functions except those in Corollary 5.1.

9 Bibliography

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, et al. (2015). "Some new asymptotic theory for least squares series: Pointwise and uniform results". In: *Journal of Econometrics* 186.2. High Dimensional Problems in Econometrics, pp. 345–366. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2015.02.014.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (Nov. 2013). "Inference on Treatment Effects after Selection among High-Dimensional Controls†". In: *The Review of Economic Studies* 81.2, pp. 608–650. ISSN: 0034-6527. DOI: 10.1093/restud/rdt044. URL: https://doi.org/10.1093/restud/rdt044.

Bühlmann, Peter and Sara van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin, Heidelberg. ISBN: 978-3-642-20191-2. DOI: https://doi.org/10.1007/978-3-642-20192-9.

Bunea, Florentina, Alexandre Tsybakov, and Marten Wegkamp (2007). "Sparsity oracle inequalities for the Lasso". In: *Electronic Journal of Statistics* 1.none, pp. 169–194. DOI: 10.1214/07-EJS008.

Chen, Xiaohong (2007). "Large Sample Sieve Estimation of Semi-Nonparametric Models". In: ed. by James J. Heckman and Edward E. Leamer. Vol. 6. Handbook of Econometrics. Elsevier, pp. 5549–5632. DOI: https://doi.org/10.1016/S1573-4412(07)06076-X. URL: https://www.sciencedirect.com/science/article/pii/S157344120706076X.

Donoho, David L. and Michael Elad (2003). "Optimally sparse representation in general (nonorthogonal) dictionaries via L_1 minimization". In: *Proceedings of the National Academy of Sciences* 100.5, pp. 2197–2202. DOI: 10.1073/pnas.0437847100. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.0437847100. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0437847100.

Grenander, U. (1981). *Abstract Inference*. Probability and Statistics Series. Wiley. ISBN: 9780471082675. URL: https://books.google.de/books?id=ng2oAAAAIAAJ.

Haar, Alfred (1910). "Zur Theorie der orthogonalen Funktionensysteme". In: *Mathematische Annalen* 69.3, pp. 331–371. DOI: 10.1007/BF01456326. URL: https://doi.org/10.1007/BF01456326.

Hansen, Bruce (2022a). *Econometrics*. Economics & Finance. Princeton University Press. ISBN: 9780691235899. URL: https://press.princeton.edu/books/hardcover/9780691235899/econometrics.

— (2022b). *Probability and Statistics for Economists*. Economics & Finance. Princeton University Press. ISBN: 9780691235943. URL: https://press.princeton.edu/books/hardcover/9780691235943/probability-and-statistics-for-economists.

Horowitz, Joel L. (1998). *Semiparametric Methods in Econometrics*. Lecture Notes in Statistics. Springer New York, NY. ISBN: 9780387984773. DOI: https://doi.org/10.1007/978-1-4612-0621-7.

Koltchinskii, Vladimir (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Lecture Notes in Mathematics. Springer Berlin, Heidelberg. ISBN: 978-3-642-22146-0. DOI: https://doi.org/10.1007/978-3-642-22147-7.

Lin, Yi (2000). "Tensor product space ANOVA models". In: *The Annals of Statistics* 28.3, pp. 734–755. DOI: 10.1214/aos/1015951996. URL: https://doi.org/10.1214/aos/1015951996.

Mason, J. C., G. Rodriguez, and S. Seatzu (1993). "Orthogonal splines based on B-splines — with applications to least squares, smoothing and regularisation problems". In: *Numerical Algorithms* 5, pp. 25–40. ISSN: 1572-9265. DOI: https://doi.org/10.1007/BF02109281.

Meinshausen, Nicolai and Peter Bühlmann (2006). "High-dimensional graphs and variable selection with the Lasso". In: *The Annals of Statistics* 34.3, pp. 1436–1462. DOI: 10.1214/009053606000000281. URL: https://doi.org/10.1214/009053606000000281.

Newey, Whitney K. (1997). "Convergence rates and asymptotic normality for series estimators". In: *Journal of Econometrics* 79.1, pp. 147–168. ISSN: 0304-4076. DOI: https://doi.org/10.1016/S0304-4076(97)00011-0.

Stone, M. H. (1948). "The Generalized Weierstrass Approximation Theorem". In: *Mathematics Magazine* 21.4, pp. 167–184.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x.

Trevor Hastie, Robert Tibshirani and Martin Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Monographs on Statistics and Applied Probability. CRC Press, Taylor & Francis Group. ISBN: 978-1-498-71216-3. URL: https://searchworks.stanford.edu/view/11165952.

White, H. and J. Wooldridge (1992). "Some results on sieve estimation with dependent observations". In: ed. by W.A. Barnett, J. Powell, and G. Tauchen. Non-parametric and Semi-parametric Methods in Econometrics and Statistics. Cambridge University Press, pp. 459–493.

Zhang, Tianyu and Noah Simon (2022). *Regression in Tensor Product Spaces by the Method of Sieves*. arXiv: 2206.02994 [stat.ME].

Zou, Hui (2006). "The Adaptive Lasso and Its Oracle Properties". In: *Journal of the American Statistical Association* 101.476, pp. 1418–1429. DOI: 10.1198/016214506000000735. eprint: https://doi.org/10.1198/016214506000000735. URL: https://doi.org/10.1198/016214506000000735.

10 Appendix

10.1 Standardization of Lasso

In this section, we attempt to show that a standardization of regressors is required for an effective variable selection in Lasso. We use the same setting as shown in Section 3.1. $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ is iid data with $Z_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$, following a linear model.

$$Y = \sum_{j=1}^{p} Z_j \beta_j + \epsilon$$

where β_j are unknown coefficients. $Y = (Y_1, Y_2, \ldots, Y_n)^T$, $Z_j = (Z_{1j}, Z_{2j}, \ldots, Z_{nj})^T$ and $E[\epsilon | Z_1, Z_2, \ldots, Z_n] = 0$. To save notation in the proof, we further assume that Z_i does not contain a constant. Also, we have an orthonormality design, i.e., regressors are mutually orthonormal. To put it formally,

$$\frac{1}{n} Z^T Z = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

where $Z = (Z_1, Z_2, \ldots, Z_j)$. Variances of all regressors Z_j are normalized to 1. In this orthonormal design, it can be easily shown[31] that a lasso estimate $\hat{\beta}_{\lambda,j}$ is

$$\hat{\beta}_{\lambda,j} = \begin{cases} \hat{\beta}_{OLS,j} + \lambda & , \text{if } \hat{\beta}_{OLS,j} < -\lambda \\ 0 & , \text{if } |\hat{\beta}_{OLS,j}| \leq \lambda \\ \hat{\beta}_{OLS,j} - \lambda & , \text{if } \hat{\beta}_{OLS,j} > \lambda \end{cases}$$

where $\hat{\beta}_{OLS,j} = \frac{1}{n} \sum_{i=1}^{n} Z_{ij} Y_i$ is the coefficients estimated by OLS in this orthonormal design. It is seen that Lasso only selects regressor Z_j if $|\hat{\beta}_{OLS,j}| = |\frac{1}{n} \sum_{i=1}^{n} Z_{ij} Y_i| > \lambda$. It means that the selection rule depends on the scale of coefficients. For example, if Z_1 is centered at 10000, while Z_2 is centered at 0, Lasso would unfairly prefer to selecting Z_1 since $|\frac{1}{n} \sum_{i=1}^{n} Z_{i1} Y_i| > \lambda$ even for a large λ. To remove such an unfair influence, it is required to standardize regressors beforehand.

$$\tilde{Z}_{ij} = \frac{Z_{ij} - \bar{Z}_j}{\hat{\sigma}_j}$$

where $\hat{\sigma}_j$ is the standard deviation of Z_j. By using standardized $\tilde{Z}_{ij}$, the selection rule is not affected by the magnitude of variables, thus achieving an effective variable selection.

[31] Please refer to Section 2.4.1, Trevor Hastie and Wainwright 2015 for details.

10.2 Derivation of κ_M in Assumption 5.3

Let $g_k(X) = \prod_{l=1}^{d} p_{k_l}(x_l)$ denote basis functions with $1 \leq k \leq M$, and $g^M(X) = (g_1(X), \ldots, g_M(X))^T$ is the vector of all basis functions. Then, $\Psi_M = E[g^M(X)g^M(X)^T]$ in our setting. Let $v \in \mathbb{R}^M$, then

$$\begin{aligned}
v^T\Psi_M v =& v^T E[g^M(X)g^M(X)^T]v \\
=& E[(v^T g^M(X))^2] \\
\geq& \frac{1}{C}\int_{\mathcal{X}} (v^T g^M(x))^2 dx \\
=& \frac{1}{C}\int_{\mathcal{X}} \sum_{k=1}^{M}\sum_{j=1}^{M} v_k v_j g_k(x) g_j(x) dx \\
=& \frac{1}{C}\sum_{k=1}^{M} v_k^2 \\
=& \frac{1}{C} v^T v
\end{aligned}$$

To move from the second to the third line, we use the assumption that $||g_k(X)|| = ||\prod_{l=1}^{d} p_{k_l}(x_l)|| \geq c_0 = \frac{1}{C}$. To move from the fourth to fifth line, we use the orthonormality property of our basis functions. Moreover,

$$\begin{aligned}
v^T \mathrm{diag}(\Psi_M) v \leq& C\int_{\mathcal{X}} \sum_{k=1}^{M} v_k^2 g_k(x)^2 dx \\
=& C\sum_{k=1}^{M} v_k^2 \int_{\mathcal{X}} g_k(x)^2 dx \\
=& C v^T v
\end{aligned}$$

In the first line, we use the assumption that $||g_k(X)|| = ||\prod_{l=1}^{d} p_{k_l}(x_l)|| \leq C$. It follows that for any non-zero real vector v, we have

$$\begin{aligned}
v^T\left(\Psi - \frac{1}{C^2}\mathrm{diag}(\Psi_M)\right) v =& v^T\Psi_M v - \frac{1}{C^2} v^T \mathrm{diag}(\Psi_M) v \\
\geq& \frac{1}{C} v^T v - \frac{1}{C} v^T v \\
=& 0
\end{aligned}$$

Therefore, there exists a constant $\kappa_M = \frac{1}{C^2} > 0$ such that $\Psi_M - \kappa_M \mathrm{diag}(\Psi_M)$ is positive semi-definite. Assumption 4.3 is thus guaranteed. It is noted that orthonormality of basis functions simplifies the expression of κ_M, but even if basis functions are not orthonormal, there can exist such a $\kappa_M > 0$ that makes $\Psi_M - \kappa_M \mathrm{diag}(\Psi_M)$ positive semi-definite, though probably with a more complicated derivation process and expression.

10.3 Proof of Corollary 5.1

10.3.1 Normalized Legendre Polynomials

Step 1: Deriving L and L_0 in Assumption 5.2

By definition, Legendre polynomials are orthogonal with respect to a uniform density on $[-1, 1]$ (B. Hansen 2022a). Suppose $p_k(X)$ $(k = 0, 1, 2, \ldots)$ are Legendre polynomials defined on interval $[-1, 1]$. By fixing $p_k(1) = 1$, we have[32]:

$$\int_{-1}^{1} p_j(x)p_k(x)\,\mathrm{d}x = \frac{2}{2k+1}\delta_{jk}$$

where δ_{jk} denotes Kronecker delta[33], which equals to 1 if $j = k$, and equals to 0 otherwise. That means, the squared L_2 norm of $p_k(X)$ on $[-1, 1]$ is:

$$||p_k(X)||^2 = \int_{-1}^{1} p_k(x)^2\,\mathrm{d}x = \frac{2}{2k+1}$$

where $||\cdot||$ denotes L_2 norm. Then, if we further normalize the Legendre polynomials $p_k(X)$, we have $\tilde{p}_k(X) = \sqrt{\frac{2k+1}{2}}p_k(X)$ such that:

$$||\tilde{p}_k(X)||^2 = \int_{-1}^{1} \tilde{p}_k(X)^2\,\mathrm{d}x = \int_{-1}^{1} \left(\sqrt{\frac{2k+1}{2}}p_k(x)\right)^2 \mathrm{d}x = \frac{2k+1}{2}||p_k(X)||^2 = 1$$

Therefore, $\tilde{p}_k(X)$ $(k = 0, 1, 2, \ldots)$ are normalized Legendre polynomials that are discussed in Corollary 5.1. They are orthonormal. In Figure 1, it is seen that Legendre polynomials $p_k(X)$ are bounded above by 1, so for normalized Legendre polynomials, we have:

$$||\tilde{p}_k(X)||_\infty = \left|\left|\sqrt{\frac{2k+1}{2}}p_k(X)\right|\right|_\infty \leq \sqrt{\frac{2k+1}{2}}$$

In other words, $||\tilde{p}_k||_\infty \lesssim \sqrt{k}$. As a consequence, the L_∞ norm of basis functions in our multivariate sieve space is $||\prod_{l=1}^{d}\tilde{p}_{k_l}||_\infty \lesssim \sqrt{\prod_{l=1}^{d} k_l}$. Because $\sum_{l=1}^{d} k_l \lesssim n^{\frac{1}{2\underline{s}+1}}$ with reference to (27), we can maximize the product $\prod_{l=1}^{d} k_l$ by having $k_l = \frac{1}{d}n^{\frac{1}{2\underline{s}+1}}$. Then,

$$\left|\left|\prod_{l=1}^{d}\tilde{p}_{k_l}\right|\right|_\infty \lesssim n^{\frac{d}{4\underline{s}+2}}$$

As a result, according to Assumption 5.2, we have $L = n^{\frac{d}{4\underline{s}+2}}$ and $L_0 = L^2 = n^{\frac{d}{2\underline{s}+1}}$.

[32] `https://en.wikipedia.org/wiki/Legendre_polynomials`

[33] Kronecker delta is a function of two variables, which are usually non-negative integers. The function is 1 if the two variables equal, and is 0 otherwise. For example, $\delta_{12} = 0$, because $1 \neq 2$; while $\delta_{11} = 1$, because $1 = 1$.

Step 2: Discussing the first part of inequality about $\pi_{n,M}(\beta)$ in (22)

To make $\pi_{n,M}(\beta)$ asymptotically zero, it is sufficient to make its upper bound to be asymptotically zero. There are two parts in its upper bound, then it is sufficient to have both parts going to zero in the limit. Ignoring constants, the first part of upper bound of $\pi_{n,M}(\beta)$ is

$$\begin{aligned}
&M^2 \exp\left(-\min\left\{nr_{n,M}^2, \frac{nr_{n,M}}{L}, \frac{n}{L^2}, \frac{n}{L_0 M^2(\beta)}, \frac{n}{L^2 M(\beta)}\right\}\right) \\
=&M^2 \exp\left(-\min\left\{A^2 \log M, A\sqrt{\log M} n^{\frac{1}{2}-\frac{d}{4\underline{s}+2}}, n^{1-\frac{d}{2\underline{s}+1}}, n^{1-\frac{d}{2\underline{s}+1}} M^{-2}(\beta), n^{1-\frac{d}{2\underline{s}+1}} M^{-1}(\beta)\right\}\right) \\
=&M^2 \exp\left(-\min\left\{A^2 \log M, A\sqrt{\log M} n^{\frac{1}{2}-\frac{d}{4\underline{s}+2}}, n^{1-\frac{d}{2\underline{s}+1}}, n^{1-\frac{d}{2\underline{s}+1}-\frac{2d}{2s+d}}, n^{1-\frac{d}{2\underline{s}+1}-\frac{d}{2s+d}}\right\}\right)
\end{aligned}$$

To move from the first to second line, we use $L = n^{\frac{d}{4\underline{s}+2}}$ and $L_0 = n^{\frac{d}{2\underline{s}+1}}$ for normalized Legendre polynomials. To move from the second to third line, we use $M(\beta) = n^{\frac{d}{2s+d}}$ because it is the largest among all three types of underlying models. A sufficient condition for this upper bound to be asymptotically zero is that all terms within the brace should go to infinity, otherwise the upper bound is not asymptotically zero. The terms $A^2 \log M$ is naturally non-zero. Then the sufficient condition is reduced to

$$\begin{cases}
\frac{1}{2} - \frac{d}{4\underline{s}+2} & \geq 0 \\
1 - \frac{d}{2\underline{s}+1} & > 0 \\
1 - \frac{d}{2\underline{s}+1} - \frac{2d}{2s+d} & > 0 \\
1 - \frac{d}{2\underline{s}+1} - \frac{d}{2s+d} & > 0 \\
\underline{s} & \geq 0
\end{cases}$$

If this set of inequalities works for $\underline{s}$, then it should work for all s. So, we can replace s with $\underline{s}$. Then, by solving this set of inequalities, we have

$$\underline{s} > \frac{2d - 1 + \sqrt{8d^2 + 1}}{4}$$

For example, when $d = 2$, we need $\underline{s} > 2.187$ to maintain the optimal convergence rate in Theorem 5.1. The formula is a little bit complicated, but we can find a simpler sufficient condition.

$$\underline{s} \geq \frac{3d - 1}{2}$$

This is a sufficient condition because $\frac{3d-1}{3} > \frac{2d-1+\sqrt{8d^2+1}}{4}$ for all $d \in \{1, 2, \ldots\}$. It is worthy of mentioning that the above condition is derived from using $M(\beta) = n^{\frac{d}{2s+d}}$, which is the largest $M(\beta)$ in all three underlying models.

Hence, even if $\underline{s} \leq \frac{2d-1+\sqrt{8d^2+1}}{4}$, it can still help to maintain the optimal convergence rate in additive and parametric model. By following a similar argument as above but with $M(\beta) = n^{\frac{1}{2\underline{s}+1}}$, we have a sufficient condition $\underline{s} \geq \frac{d+1}{2}$, which is smaller than the lower bound for maintaining convergence rate in all three models. Likewise, if we only need the optimal convergence rate to hold in parametric model, then we can use $M(\beta) = K^d$ with K being a positive constant. The resulting condition is $\underline{s} > \frac{d-1}{2}$.

Step 3: Discussing the second part of inequality about $\pi_{n,M}(\beta)$ in (22)
Regarding the second part of upper bound of $\pi_{n,M}(\beta)$, we have

$$\exp\left(-\frac{M(\beta)}{L^2(\beta)} n r_{n,M}^2\right) = \exp\left(-\frac{M(\beta)}{L^2(\beta)} A^2 \log M\right)$$

where $L(\beta) = ||f - f_\beta||_\infty$. With reference to the formula (3.7) in Belloni, Chernozhukov, Chetverikov, et al. 2015, we have $L(\beta) \lesssim M(\beta)^{-\frac{s}{d}}$. Then,

$$\frac{M(\beta)}{L^2(\beta)} \gtrsim M(\beta)^{\frac{2s+d}{d}} = n^{\frac{d}{2s+d}\frac{2s+d}{d}} = n$$

where we use $M(\beta) = n^{\frac{d}{2s+d}}$. Obviously, the second part is asymptotically zero, too. The result also holds when $M(\beta) = n^{\frac{1}{2s+1}}$ or K^d. The intuitive explanation is that when $n \to \infty$, the L_∞ norm of approximation error $L(\beta)$ goes to zero, and the number of coefficients $M(\beta)$ goes to infinity, thus making $\frac{M(\beta)}{L^2(\beta)}$ goes to infinity.

10.3.2 Orthonormalized B-Splines

Step 1: Deriving L and L_0 in Assumption 5.2
In Appendix 10.4, it is shown that on $[a, b]$, the squared L_2 norm of orthogonal B-Splines $||P_k||^2$ are bounded above. From the example in Table 6, it is further observed that $||P_k||^2$ grows with k. Hence, if P_k is normalized, the resulting orthonomalized B-Splines $\tilde{P}_k$ would have an L_∞ norm that goes downward with k. Then, it is bounded above, supposedly by a constant C. As a consequence, the L_∞ norm of basis functions in multivariate sieve space is $||\prod_{l=1}^d \tilde{P}_{k_l}||_\infty \lesssim C^d$. According to Assumption 5.2, we have $L = C^d$ and $L_0 = L^2 = C^{2d}$.

Step 2: Discussing the first part of inequality about $\pi_{n,M}(\beta)$ in (22)
Again, it is sufficient to argue that both parts of the upper bound of $\pi_{n,M}(\beta)$ goes to zero in the limit. Ignoring constants, the first part is

$$\begin{aligned}
&M^2 \exp\left(-\min\left\{n r_{n,M}^2, \frac{n r_{n,M}}{L}, \frac{n}{L^2}, \frac{n}{L_0 M^2(\beta)}, \frac{n}{L^2 M(\beta)}\right\}\right) \\
=&M^2 \exp\left(-\min\left\{A^2 \log M, A\sqrt{\log M} C^{-d} n, C^{-2d} n, C^{-2d} n M^{-2}(\beta), C^{-2d} n M^{-1}(\beta)\right\}\right) \\
=&M^2 \exp\left(-\min\left\{A^2 \log M, A\sqrt{\log M} C^{-d} n, C^{-2d} n, C^{-2d} n^{1-\frac{2d}{2s+d}}, C^{-2d} n^{1-\frac{d}{2s+d}}\right\}\right)
\end{aligned}$$

To move from the first to second line, we use $L = C^d$ and $L_0 = C^{2d}$ for orthonormalized B-splines. To move from the second to third line, we use $M(\beta) = n^{\frac{d}{2s+d}}$ because it is the largest among all three underlying models. Again, a sufficient condition is that all terms should go to infinity. The terms $A^2 \log M$, $A\sqrt{\log M}C^{-d}n$ and $C^{-2d}n$ are nonzero, so the condition is reduced to

$$\begin{cases} 1 - \frac{2d}{2s+d} & > 0 \\ 1 - \frac{d}{2s+d} & > 0 \\ \underline{s} & \geq 0 \end{cases}$$

We need to replace s with the lower bound of smoothness, i.e., $\underline{s}$, so this set of inequalities should work for all s. By solving this set of inequalities, we have

$$\underline{s} > \frac{d}{2}$$

A sufficient condition is $\underline{s} \geq \frac{d+1}{2}$. For example, when $d = 2$, we need $\underline{s} \geq 1.5$ to maintain the convergence rate in Theorem 5.1; when $d = 10$, we need $\underline{s} \geq 5.5$. This restriction is less stricter than that for normalized Legendre polynomials. If we only need optimal convergence rate for additive and parametric model, we can use $M(\beta) = n^{\frac{1}{2s+1}}$ instead. The resulting condition is $\underline{s} > \frac{1}{2}$. By following a similar argument, we find out we don't need any further restriction on $\underline{s}$ except that $\underline{s} \geq 0$ if we only need an optimal rate in parametric model.

Step 3: Discussing the second part of inequality about $\pi_{n,M}(\beta)$ in (22)
Regarding the second part of upper bound of $\pi_{n,M}(\beta)$, we can verify that it goes to zero with $n \to \infty$ in the same way as we argue in Section 10.3.1.

10.3.3 Normalized Haar wavelets

As shown in Section 2.4.6, normalized Haar basis functions are denoted as $\psi_{jk}(X)$. They are orthogonal with respect to $[0, 1]$, by design. With normalization, the L_∞ norm of Haar basis functions becomes $||\psi_{jk}(X)||_\infty = 2^{j/2}$. Let k' be the total number of basis functions in all levels of Haar wavelets up to the level of j, then $k' = 2^{j+1}$, which means $j = \log_2 k' - 1$. The L_∞ norm becomes:

$$||\psi_{jk}(X)||_\infty = 2^{\frac{\log_2 k' - 1}{2}} = 2^{-1/2} 2^{\log_2 \sqrt{k'}} = 2^{-1/2}\sqrt{k'} \lesssim \sqrt{k'}$$

It is found that the derived condition of L_∞ norm for normalized Haar wavelets coincides with that for normalized Legendre polynomials. As a consequence, the same L and L_0 are obtained, i.e., $L = n^{\frac{d}{4\underline{s}+2}}$ and $L_0 = L^2 = n^{\frac{d}{2\underline{s}+1}}$. All other proof for normalized Legendre polynomials in Section 10.3.1 also apply in this section.

10.3.4 Normalized Trigonometric Polynomials

First, we can show that the system $\mathcal{T} := \{1, \cos(\pi X), \sin(\pi X), \cos(2\pi X), \sin(2\pi X), \ldots\}$ is a complete orthogonal system for $X \in [-1, 1]$. In order to prove the orthogonality of this system, we have to show[34]

$$\begin{cases} \int_{-1}^{1} \cos(m\pi x)\cos(n\pi x)\,\mathrm{d}x = 0, & m, n \in \mathbb{Z}_+, m \neq n \\ \int_{-1}^{1} \sin(m\pi x)\sin(n\pi x)\,\mathrm{d}x = 0, & m, n \in \mathbb{Z}_+, m \neq n \\ \int_{-1}^{1} \cos(m\pi x)\sin(n\pi x)\,\mathrm{d}x = 0, & m, n \in \mathbb{Z}_+, m \neq n \end{cases}$$

To show the second integral, we use the following product-to-sum trigonometric identity.

$$\sin A \sin B = \frac{\cos(A-B) - \cos(A+B)}{2}$$

So, if $m, n \geq 1$ and $m \neq n$, we have

$$\begin{aligned} \int_{-1}^{1} \sin(m\pi x)\sin(n\pi x)\,\mathrm{d}x =& \int_{-1}^{1} \frac{\cos\left((m-n)\pi x\right) - \cos\left((m+n)\pi x\right)}{2}\,\mathrm{d}x \\ =& \frac{1}{2}\left[\frac{\sin\left((m-n)\pi x\right)}{m-n} - \frac{\sin\left((m+n)\pi x\right)}{m+n}\right]_{-1}^{1} \\ =& 0 \end{aligned}$$

Likewise, the other two integrals can be solved by using the following identities.

$$\begin{cases} \cos A \cos B = & \frac{\cos(A-B)+\cos(A+B)}{2} \\ \cos A \sin B = & \frac{\sin(A+B)-\sin(A-B)}{2} \end{cases}$$

Also, for $m, n \geq 1$ and $x \in [-1, 1]$, we have $||1||^2 = 2$, $||\cos(m\pi x)||^2 = \pi$ and $||\sin(n\pi x)||^2 = \pi$. Therefore, to normalize the system $\mathcal{T}$ of trigonometric polymials, we would have a new system $\tilde{\mathcal{T}} = \{1/\sqrt{2}, \cos(\pi X)/\sqrt{\pi}, \sin(\pi X)/\sqrt{\pi}, \cos(2\pi X)/\sqrt{\pi}, \sin(2\pi X)/\sqrt{\pi}, \ldots\}$. This is then an orthonormal system. Moreover, as $\cos(m\pi x)$ and $\cos(n\pi x)$ is bounded above by 1, the L_∞ norm of basis functions in the new system $\tilde{\mathcal{T}}$ is then bounded above by $1/\sqrt{2}$, which is a constant. Consequently, L and L_0 in Assumption 5.2 are also constant, thus making the rest of proof the same as that in Section 10.3.2 for orthonormalized B-splines.

To conclude, all basis functions we use in this section are orthonormalized only for the consolidation and simplicity of proof. If basis functions are not orthonormal, Theorem 5.1 can still hold but with some changes to parameters. For example, if we don't normalize Legendre polynomials, L and L_0 are then constants, and κ_M is not $1/C^2$ any more. This may change the lower bound of smoothness in Corollary 5.1. In practical implementation of our dimension adaptive estimation, it is not a must to orthonormalize basis functions.

[34] The basis 1 is naturally orthogonal to other bases on the domain $[-1, 1]$.

10.4 Orthogonal B-splines

In this section, we are going to show that the squared L_2 norm of orthogonal B-splines are bounded below and above. To simplify our proof, we use linear B-splines, but the result also applies to B-splines with higher orders with more notation. The main idea of this proof is from Mason, Rodriguez, and Seatzu 1993.

Suppose a family of linear B-splines $\{L_k\}(k = 0, \ldots, n)$ is defined on $[a, b]$, and the set of ordered knots $\{x_k\}$ with $x_{-1} < x_0 = a$ and $b = x_n < x_{n+1}$. Then, L_k is continuous in $[x_{k-1}, x_{k+1}]$, and by normalization we assume that $L_k(x_k) = 1$. Now, we hope to convert the linear B-splines $\{L_k\}$ to a basis of orthogonal splines. Typically, it can be done by the following recurrence:

$$\begin{aligned} P_0 =& L_0 \\ P_k =& L_k - a_{k-1}P_{k-1} \quad (k = 1, \ldots, n) \end{aligned}$$

where P_k is orthogonal linear B-splines with support $[a, x_{k+1}]$ and a_{k-1} are undetermined parameters. Then, we need to determine the values of a_{k-1}. The inner product $\langle P_r, L_k\rangle = 0$ $(r \leq k-2)$ because P_r and L_k have disjoint supports. Hence, it is sufficient to have

$$\langle P_k, P_{k-1}\rangle = \langle L_k - a_{k-1}P_{k-1}, P_{k-1}\rangle = \langle L_k, P_{k-1}\rangle - a_{k-1}||P_{k-1}||^2$$

Then, we have $\langle L_k, P_{k-1}\rangle - a_{k-1}||P_{k-1}||^2$. Denote $n_k = ||P_k||^2 = \langle P_k, P_k\rangle$ and $v_k = \langle L_{k+1}, P_k\rangle$. So, $v_k = a_k n_k$. It is noted that $\langle L_{k+1}, P_k\rangle = \langle L_{k+1}, L_k - a_{k-1}P_{k-1}\rangle = \langle L_{k+1}, L_k\rangle$ due to $\langle L_{k+1}, P_{k-1}\rangle = 0$, then $v_k = \langle L_k, L_{k+1}\rangle$ too. Thus, the squared L_2 norm of P_k is

$$\begin{aligned} n_k =& \langle P_k, P_k\rangle \\ =& \langle L_k - a_{k-1}P_{k-1}, L_k - a_{k-1}P_{k-1}\rangle \\ =& \langle L_k, L_k\rangle - 2a_{k-1}\langle L_k, P_{k-1}\rangle + a_{k-1}^2\langle P_{k-1}, P_{k-1}\rangle \\ =& u_k - 2a_{k-1}v_{k-1} + a_{k-1}^2 n_{k-1} \\ =& u_k - 2a_{k-1}v_{k-1} + a_{k-1}v_{k-1} \\ =& v_k - a_{k-1}v_{k-1} \end{aligned}$$

where v_k and u_k are computable constants and should be bounded due to the property of B-splines. By using $v_k = a_k n_k$ and $n_k = u_k - a_{k-1}v_{k-1}$, we can solve for n_k and a_k $(k = 0, 1, \ldots, n-1)$:

$$\begin{aligned} n_0 =& u_0 \\ a_k =& \frac{v_k}{n_k} \\ n_{k+1} =& u_{k+1} - a_k v_k \end{aligned}$$

Note that $\{a_k\}$ can be alternatively eliminated. Then, we have

$$\begin{aligned} n_0 =& u_0 \\ n_k =& u_k - \frac{v_{k-1}^2}{n_{k-1}} \quad (k = 1, \ldots, n-1) \end{aligned}$$

Obviously, n_k is bounded above since u_k is bounded above and $\frac{v_{k-1}^2}{n_{k-1}} \geq 0$; n_k is also bounded below since it is a squared L_2 norm. To put it formally, there exists constants C_1 and C_2 such that $C_1 \leq n_k \leq C_2$ for $k = 0, 1, \ldots, n-1$.

Mason, Rodriguez, and Seatzu 1993 also presents an example by giving explicit values to constants u_k, v_k and formulae for $\{L_k\}$. Denote $h_k = x_k - x_{k-1}$, and they show

$$L_k(x) = \begin{cases} (h_k)^{-1}(x - x_{k-1}), & x \in [x_{k-1}, x_k] \\ (h_{k+1})^{-1}(x_{k+1} - x), & x \in [x_k, x_{k+1}] \end{cases}$$

and

$$\begin{aligned} u_0 =& \frac{1}{3} h_1 \\ u_k =& \frac{1}{3}(h_k + h_{k+1}) \quad (k = 1, \ldots, n-1) \\ u_n =& \frac{1}{3} h_n \\ v_k =& \frac{1}{6} h_{k+1} \quad (k = 0, \ldots, n-1) \end{aligned}$$

Hence, it gives rise to the following equations: $n_0 = \frac{1}{3} h_1$ and $n_k = \frac{1}{3}(h_k + h_{k+1}) - \frac{h_k^2}{36 n_{k-1}}$ for all $k = 1, \ldots, n-1$. Without loss of generality, assume $h_k = 1$ for all k, then the explicit values of n_k are as follows.

Table 6: Example Values of n_k

k	1	2	3	4	5	6	7	8
n_k	0.33333	0.58333	0.61905	0.62180	0.62199	0.622010	0.622010	0.622010

In Table 6, values of n_k are observed to converge to 0.622010 in this specific example, testifying the existence of an upper bound.

10.5 Additional Tables

Table 7: Median of out-of-sample empirical MSE for m_1

		$\sigma_1 = 5\%$				$\sigma_2 = 20\%$			
		Post LASSO		Post Adaptive LASSO		Post LASSO		Post Adaptive LASSO	
Basis	Estimator	MSE	Terms[b]	MSE	Terms[b]	MSE	Terms[b]	MSE	Terms[b]
Power Series	*dimada*	0.002622	17	0.002537	5	0.04180	17	0.04067	5
	addt	0.002561	7	0.002537	5	0.04091	8	0.04060	5
	ols[a]	**0.002537**	5	**0.002537**	5	**0.04060**	5	**0.04060**	5
Legendre	*dimada*	0.002616	17	0.002537	5	0.04194	16	0.04068	5
	addt	0.002554	7	0.002537	5	0.04105	8	0.04060	5
	ols[a]	**0.002537**	5	**0.002537**	5	**0.04060**	5	**0.04060**	5
B-Splines	*dimada*	0.010528	191	0.009168	107	0.10267	130	0.08736	94
	addt	0.002676	28	0.002678	25	0.04282	28	0.04308	25
	ols[a]	**0.002537**	5	**0.002537**	5	**0.04060**	5	**0.04060**	5
Trigonometric	*dimada*	0.006139	159	0.004363	14	0.05596	42	0.05487	33
	addt	0.006612	12	0.006563	10	0.04397	20	0.04387	11
	ols[a]	**0.002537**	5	**0.002537**	5	**0.04060**	5	**0.04060**	5

[a] The *ols* estimator doesn't involve any basis function or Lasso-type methods.
[b] Average number of selected terms with non-zero coefficients.

Table 8: Median of out-of-sample empirical MSE for m_2

		$\sigma_1 = 5\%$				$\sigma_2 = 20\%$			
		Post LASSO		Post Adaptive LASSO		Post LASSO		Post Adaptive LASSO	
Basis	Estimator	MSE	Terms[b]	MSE	Terms[b]	MSE	Terms[b]	MSE	Terms[b]
Power Series	*dimada*	0.17097	37	0.16437	18	0.2133	38	0.2016	19
	addt	**0.14298**	13	**0.14767**	9	**0.1820**	13	**0.1844**	9
	ols[a]	0.74993	5	0.74993	5	0.7876	5	0.7876	5
Legendre	*dimada*	0.17357	41	0.16556	22	0.2154	41	0.2066	23
	addt	**0.14379**	12	**0.14612**	10	**0.1824**	12	**0.1846**	10
	ols[a]	0.74993	5	0.74993	5	0.7876	5	0.7876	5
B-Splines	*dimada*	0.18213	72	0.16437	44	0.2499	72	0.2284	47
	addt	**0.08363**	27	**0.08407**	22	**0.1228**	27	**0.1237**	22
	ols[a]	0.74993	5	0.74993	5	0.7876	5	0.7876	5
Trigonometric	*dimada*	0.20162	172	0.18461	105	0.2893	110	0.2665	73
	addt	**0.06523**	24	**0.06703**	19	**0.1047**	24	**0.1066**	19
	ols[a]	0.74993	5	0.74993	5	0.7876	5	0.7876	5

[a] The *ols* estimator doesn't involve any basis function or Lasso-type methods.
[b] Average number of selected terms with non-zero coefficients.

Table 9: Median of out-of-sample empirical MSE for m_3

		$\sigma_1 = 5\%$				$\sigma_2 = 20\%$			
		Post LASSO		Post Adaptive LASSO		Post LASSO		Post Adaptive LASSO	
Basis	Estimator	MSE	Terms[b]	MSE	Terms[b]	MSE	Terms[b]	MSE	Terms[b]
Power Series	*dimada*	**0.03686**	32.0	**0.04233**	23	**0.07812**	33	**0.08208**	23.5
	addt	0.25622	11.0	0.25778	9	0.29562	11	0.29708	9.0
	ols[a]	0.31443	5.0	0.31443	5	0.35318	5	0.35318	5.0
Legendre	*dimada*	**0.03691**	32.0	**0.04128**	24	**0.07851**	34	**0.08277**	24.0
	addt	0.25669	10.0	0.25677	9	0.29639	10	0.29720	9.0
	ols[a]	0.31443	5.0	0.31443	5	0.35318	5	0.35318	5.0
B-Splines	*dimada*	**0.09503**	177.0	**0.08396**	121	**0.21478**	134	**0.17898**	100.0
	addt	0.26228	27.0	0.26367	24	0.30361	27	0.30527	24.0
	ols[a]	0.31443	5.0	0.31443	5	0.35318	5	0.35318	5.0
Trigonometric	*dimada*	**0.03187**	204.5	**0.03228**	117	**0.11270**	124	**0.10746**	87.0
	addt	0.26477	21.0	0.26485	15	0.30635	21	0.30560	15.0
	ols[a]	0.31443	5.0	0.31443	5	0.35318	5	0.35318	5.0

[a] The *ols* estimator doesn't involve any basis function or Lasso-type methods.
[b] Average number of selected terms with non-zero coefficients.

10.6 Additional Figures

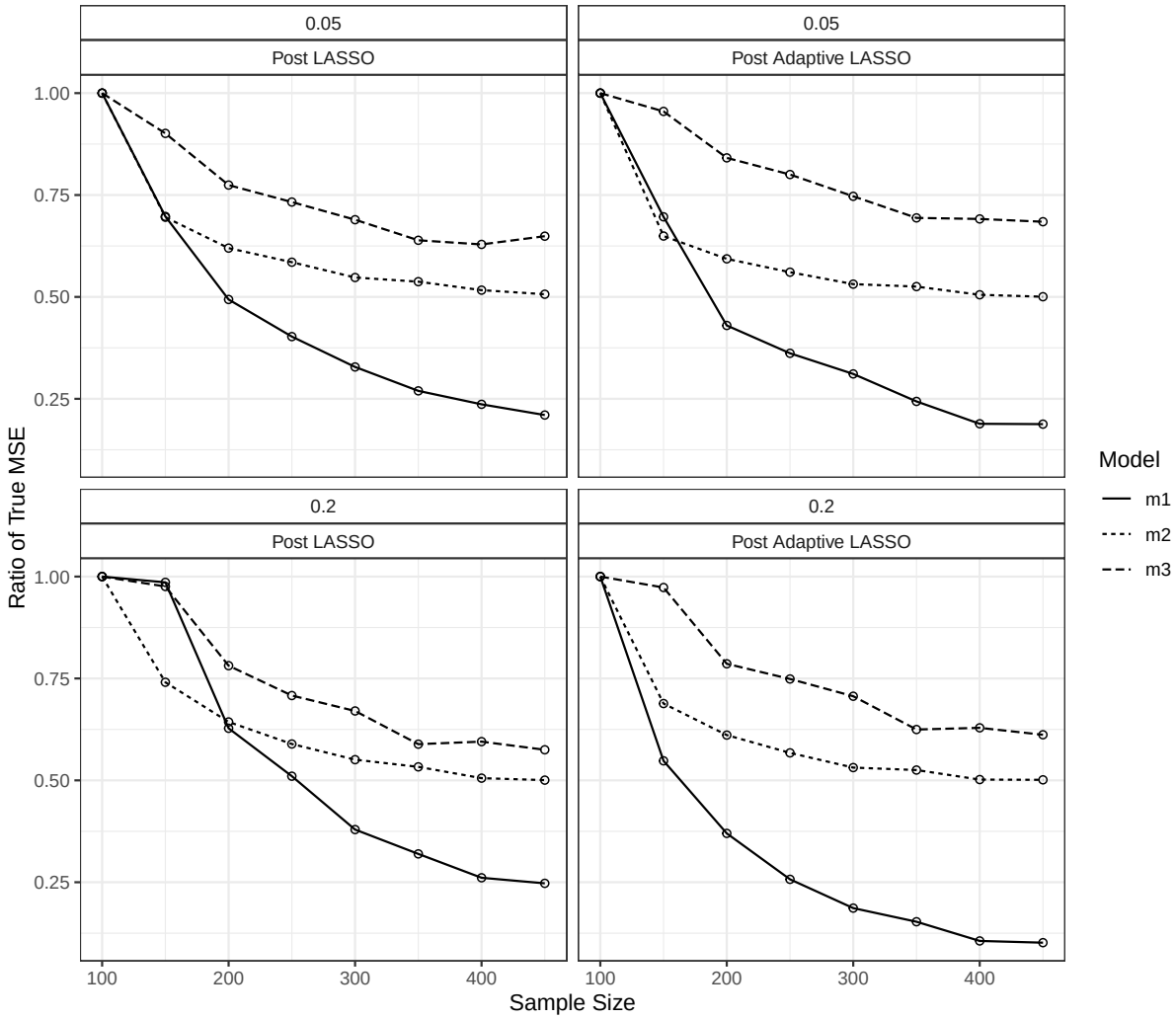

Figure 5: Comparison of convergence rates (using median MSE)

Versicherung an Eides statt

Ich versichere hiermit, dass ich die vorstehende Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass die vorgelegte Arbeit noch an keiner anderen Hochschule zur Prüfung vorgelegt wurde und dass sie weder ganz noch in Teilen bereits veröffentlicht wurde. Wörtliche Zitate und Stellen, die anderen Werken dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall kenntlich gemacht.

Bonn, 25.09.2023 _____

Chencheng Fang