Journal of Multivariate Analysis 128 (2014) 186–202

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Minimax adaptive dimension reduction for regression

Quentin Paris

IRMAR, ENS Cachan Bretagne, CNRS, UEB, Campus de Ker Lann, Avenue Robert Schuman, 35170 Bruz, France

ARTICLE INFO

Article history:
Received 26 December 2012
Available online 4 April 2014

AMS 2000 subject classifications:
62H12
62G08

Keywords:
Regression estimation
Dimension reduction
Minimax rates of convergence
Empirical risk minimization
Metric entropy

ABSTRACT

In this paper, we address the problem of regression estimation in the context of a p-dimensional predictor when p is large. We propose a general model in which the regression function is a composite function. Our model consists in a nonlinear extension of the usual sufficient dimension reduction setting. The strategy followed for estimating the regression function is based on the estimation of a new parameter, called the reduced dimension. We adopt a minimax point of view and provide both lower and upper bounds for the optimal rates of convergence for the estimation of the regression function in the context of our model. We prove that our estimate adapts, in the minimax sense, to the unknown value d of the reduced dimension and achieves therefore fast rates of convergence when $d \ll p$.

Contents

E-mail address: quentin.paris@bretagne.ens-cachan.fr.

http://dx.doi.org/10.1016/j.jmva.2014.03.008
0047-259X/

1. Introduction

1.1. The curse of dimensionality in regression

From a general point of view, the goal of regression is to infer about the conditional distribution of a real-valued response variable Y given an $\mathcal{X}$-valued predictor variable X where $\mathcal{X} \subset \mathbb{R}^p$. In the statistical framework, one usually focuses on the estimation of the regression function

$$r(x) = \mathbb{E}(Y|X = x), \tag{1.1}$$

based on a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n independent and identically distributed random variables with same distribution P as the generic random couple (X, Y).

A major issue in regression, known as the curse of dimensionality, is basically that the rates of convergence of estimates of the regression function are slow when the dimension p of the predictor variable X is high. For instance, if r is assumed to be β-Hölder and if $\hat{r}$ refers to any classical estimate (say a kernel, a nearest-neighbors or a least-squares estimate), the mean squared error $\mathbb{E}\left(\hat{r}(X) - r(X)\right)^2$ of $\hat{r}$ converges to 0 at the rate $n^{-2\beta/(2\beta+p)}$, which gets slower as p increases. To get a deeper understanding of the problem, one may refer to the minimax point of view. First, we recall the definition of optimal rates of convergence in the minimax sense. Given a set $\mathcal{D}$ of distributions P of the random couple (X, Y), υ_n is said to be an optimal rate of convergence in the minimax sense for $\mathcal{D}$ if it is a lower minimax rate, i.e.,

$$\liminf_{n\to+\infty} \upsilon_n^{-2} \inf_{\hat{r}} \sup_{P\in\mathcal{D}} \mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 > 0,$$

where the infimum is taken over all estimates based on our sample, and if there exists an estimate $\hat{r}$ such that

$$\limsup_{n\to+\infty} \upsilon_n^{-2} \sup_{P\in\mathcal{D}} \mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 < +\infty.$$

Then, in a word, when $\mathcal{D}$ is taken as the set of all distributions P of the random couple (X, Y) for which r is β-Hölder, the optimal rate of convergence for $\mathcal{D}$ is $\upsilon_n = n^{-\beta/(2\beta+p)}$ (for more details on optimal rates of convergence, we refer the reader to [22,12,14,23]). Accordingly, there is no hope of constructing an estimate which converges at a faster rate under the only general assumption that r is regular. Hence, the only alternative to obtain faster rates is to exploit additional information on the regression function.

1.2. A general model for dimension reduction in regression

In practice, when such additional information is available, it is often encoded in regression models as so called structural assumptions on the regression function. Statistical procedures based on such models are usually referred to as dimension reduction techniques. In the recent years, much attention has been paid to dimension reduction techniques due to the increasing complexity of the data considered in applications. Among popular models for dimension reduction in regression, one can mention for example the single index model (see, e.g., [1], and the references therein), the additive regression model or the projection pursuit model (see, e.g., Chapter 22 in [12]). Another important dimension reduction framework is called sufficient dimension reduction. In this framework, one assumes that

$$\mathbb{E}\left(Y|X\right) = \mathbb{E}\left(Y|\Lambda X\right) \quad \text{and} \quad \mathbb{E}\left(Y|\Lambda X = \cdot\right) \in \mathcal{G}, \tag{1.2}$$

are satisfied for a matrix $\Lambda \in \mathcal{M}_p(\mathbb{R})$ of rank smaller than p, and a class $\mathcal{G}$ of regular functions (see, e.g., [13,18,3], and the references therein). The motivation for studying such a model is that, provided the matrix Λ may be estimated, the predictor variable X may be replaced by ΛX which takes its values in a lower dimensional space. Many methods have been introduced in the literature to estimate Λ among which we mention average derivative estimation (ADE) [13], sliced inverse regression (SIR) [18], principal Hessian directions (PHD) [19], sliced average variance estimation (SAVE) [7], kernel dimension reduction (KSIR) [10] and, more recently, the optimal transformation procedure [9]. Discussions, improvements and other relevant papers on that topic can be found in [5,11,26,6,29], and in the references therein. In the last years, little attention has been paid to measuring the impact of the estimation of Λ in terms of the estimation of r. Recently, Cadre and Dong [2] have used these methods to show that, in the context of model (1.2), one could indeed construct an estimate $\hat{r}$ of the regression function such that

$$\mathbb{E}\left(\hat{r}(X) - r(X)\right)^2 = O\left(n^{-2/(2+\mathrm{rank}(\Lambda))}\right),$$

when $\mathcal{G}$ is taken as a class of Lipschitz functions.

In the present article, we tackle the problem of dimension reduction for regression by studying a model which consists in a nonlinear extension of (1.2) and which is described as follows.

Our model—For a given class $\mathcal{H}$ of functions $h : \mathcal{X} \to \mathbb{R}^p$ and a given class $\mathcal{G}$ of regular functions $g : \mathbb{R}^p \to \mathbb{R}$, we assume that the two conditions

$$\text{(i) } \mathbb{E}\left(Y|X\right) = \mathbb{E}\left(Y|h(X)\right) \quad \text{and} \quad \text{(ii) } \mathbb{E}(Y|h(X) = \cdot) \in \mathcal{G}, \tag{1.3}$$