Computational Statistics and Data Analysis 62 (2013) 122–135

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Sparse regularized local regression

Diego Vidaurre [a,*], Concha Bielza [b], Pedro Larrañaga [b]

[a] Oxford Centre for Human Brain Activity, Warneford Hospital, Department of Psychiatry, University of Oxford, UK

[b] Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Spain

ARTICLE INFO

Article history:
Received 26 January 2012
Received in revised form 17 December 2012
Accepted 15 January 2013
Available online 21 January 2013

Keywords:
Bandwidth selection
Kernel smoothing
Local linear regression
Multiple regression
Non-parametric regression
Variance reduction
Sparsity
Sparse estimation

ABSTRACT

The intention is to provide a Bayesian formulation of regularized local linear regression, combined with techniques for optimal bandwidth selection. This approach arises from the idea that only those covariates that are found to be relevant for the regression function should be considered by the kernel function used to define the neighborhood of the point of interest. However, the regression function itself depends on the kernel function. A maximum posterior joint estimation of the regression parameters is given. Also, an alternative algorithm based on sampling techniques is developed for finding both the regression parameter distribution and the predictive distribution.

1. Introduction

Consider p independent covariates $\{X_1, \ldots, X_p\}$ and a response variable Y. Let $\boldsymbol{X}$ and $\boldsymbol{y} = (y_1, \ldots, y_N)^t$ be, respectively, an $N \times p$ data matrix and a continuous-valued vector, so that each row $\boldsymbol{x}_i$ is i.i.d. related to a continuous response y_i by means of some unknown (nonlinear) function $m(\cdot)$:

$$y_i = m(\boldsymbol{x}_i) + e_i,$$

where $m(\cdot)$ is assumed to be sparse and have continuous second-order derivatives and e_i is the irreducible error term, with $\mathrm{E}[e_i|\boldsymbol{x}_i] = 0$. Therefore, $\mathrm{E}[y_i|\boldsymbol{x}_i] = m(\boldsymbol{x}_i)$. We denote the elements and the columns of $\boldsymbol{X}$, respectively, as x_{ij} and $\boldsymbol{X}_{\cdot j}$.

The objective is to estimate the response at a point of interest $\boldsymbol{x} = (x_1, \ldots, x_p)^t$ using a sparsity assumption: only a subset of the covariates is indeed relevant for the estimation. We denote as $\boldsymbol{X}^*$ the data matrix $\boldsymbol{X}$ centered at $\boldsymbol{x}$ and augmented with a first column of ones, so that $x^*_{i0} = 1$ for all $i = 1, \ldots, N$. In our approach, the homoscedasticity assumption is not strictly necessary, so we can generically define the variance of e_i as $\mathrm{Var}[e_i|\boldsymbol{x}_i] = s^2(\boldsymbol{x}_i) = \sigma_i^2$. However, as will be apparent below, we make the homoscedasticity assumption $\sigma_i^2 = \sigma^2, \forall i$, for computational reasons.

Multivariate local regression (Cleveland and Devlin, 1988; Loader, 1999) estimates a multivariate regression function valid for some neighborhood of $\boldsymbol{x}$. This function is often linear, corresponding to a first-order Taylor approximation of $m(\cdot)$ at $\boldsymbol{x}$, and it can be defined on the original covariates or on some set of basis functions defined on the original covariates. We consider for simplicity the first case, although the generalization is straightforward. Local regression is appealing from both theoretical and practical sides. On the one hand, it is known to enjoy 100% minimax efficiency for some choice of bandwidth

* Corresponding author. Tel.: +44 0 1865 289300.
E-mail addresses: diego.vidaurre@ohba.ox.ac.uk (D. Vidaurre), mcbielza@fi.upm.es (C. Bielza), pedro.larranaga@fi.upm.es (P. Larrañaga).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2013.01.008

and kernel (Fan, 1993; Ruppert and Wand, 1994). On the other hand, it is computationally fast, easy to implement, flexible, and robust to data design (Hastie and Loader, 1993).

The neighborhood is defined by a kernel function, which assigns weights $\boldsymbol{w} = (w_1, \ldots, w_N)^t$ to the data points in the data set on the grounds of their distance to $\boldsymbol{x}$. The kernel function has a bandwidth parameter, which strongly influences the estimation. High bandwidths increase the bias and decrease the variance of the estimate; low bandwidths do the opposite. The simplest approach is to use a single-value bandwidth for all regressors, and the most general setting is to use a full-matrix bandwidth, which provides flexible smoothing on all orientations. While the first usually leads to a severely biased estimate, the second can imply the estimation of a large number of parameters. A convenient compromise is a bandwidth vector or diagonal bandwidth, denoted as $\boldsymbol{h} = (h_1, \ldots, h_p)^t$, which permits adaptive smoothing at each coordinate direction. The kernel function is defined as

$$w_i^2 = K_{\boldsymbol{h}}(\boldsymbol{x}_i - \boldsymbol{x}) = \prod_{j=1}^{p} \frac{1}{h_j} K\left(\frac{x_{ij} - x_j}{h_j}\right) = \prod_{j=1}^{p} \frac{1}{h_j} K\left(\frac{x_{ij}^*}{h_j}\right), \quad i = 1, \ldots, N, \tag{1}$$

where $K(\cdot)$ is a univariate, symmetric, and non-negative function with a compact support, such that $\int K(t)dt = 1$. In this paper, we use the well-known Gaussian kernel $K(t) = (2/\pi)^{1/2} \exp(-t^2/2)$.

Then, the estimated local linear regression function $g(\cdot)$ is defined by a vector of local regression coefficients $\hat{\boldsymbol{\beta}}(\boldsymbol{x}) = (\hat{\beta}_1(\boldsymbol{x}), \ldots, \hat{\beta}_p(\boldsymbol{x}))^t$ and an intercept term $\hat{\beta}_0(\boldsymbol{x})$. For simplicity of notation, in the following we denote, unless it is necessary to do otherwise, $\hat{\beta}_0(\boldsymbol{x})$ as $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}(\boldsymbol{x})$ as $\hat{\boldsymbol{\beta}}$. Then, we have

$$\hat{y}_i = g(\boldsymbol{x}_i) = \hat{\beta}_0 + \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}}.$$

Since the data is centered at $\boldsymbol{x}$, we have $\hat{y} = g(\boldsymbol{x}) = \hat{\beta}_0$. In the following, we denote $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ as $\hat{\boldsymbol{\beta}}^*$. We can linearly estimate $\hat{\boldsymbol{\beta}}^*$ as

$$\hat{\boldsymbol{\beta}}^* = \left(\boldsymbol{X}^{*t} \boldsymbol{W} \boldsymbol{X}^*\right)^{-1} \boldsymbol{X}^{*t} \boldsymbol{W} \boldsymbol{y},$$

where $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{w}^2)$. We can interpret $\hat{\boldsymbol{\beta}}$ as an estimate of the gradient $(\partial m(\boldsymbol{x})/\partial X_1, \ldots, \partial m(\boldsymbol{x})/\partial X_p)^t$. For estimates of second derivatives, we would need at least a second-order fit.

Therefore, the cornerstone of (linear) local regression is the estimation of a suitable bandwidth. We focus on the multivariate case, where a direct estimation is not straightforward. This estimation implies unknown functionals which themselves depend on the bandwidth. Typically, the bandwidth is set by direct (plugin) computation (Wand and Jones, 1994; Yang and Tschernig, 1999), selected by cross-validation (Sain et al., 1994; Hall et al., 2007), or found within some type of suboptimal search (Lafferty and Wasserman, 2008). It is known that a suitable plugin estimate of $\boldsymbol{h}$ can improve the cross-validated estimate. In this paper, we work on the basis of a plugin diagonal bandwidth estimate, which, as mentioned above, is a reasonable tradeoff between a scalar bandwidth and a full-matrix bandwidth.

We state that the kernel estimate in the local regression framework should account for the importance of each variable. In other words, if a variable is absolutely irrelevant for the regression function, or noisy, it should not participate in the weights calculation (Vidaurre et al., 2012). Since the best rate of convergence in non-parametric regression is $N^{-4/(4+p)}$ (Györfi et al., 2002), to exploit the sparse nature of $m(\cdot)$ is extremely convenient.

The approach taken by Lafferty and Wasserman (2008), so-called *regularization of derivative expectation operator*, or *rodeo*, also uses a diagonal bandwidth, and is of special interest to us because they consider sparsity in $m(\boldsymbol{x})$. Specifically, they use the estimated gradient of the regression function with respect to the bandwidth, $\partial m(\boldsymbol{x})/\partial \boldsymbol{h}$, to conduct a greedy search, considering that a high value of $\partial m(\boldsymbol{x})/\partial h_j$ is indicative of the relevance of variable X_j. In other words, this gradient tells how the regression function varies with infinitesimal changes of the bandwidth. If it varies little, then the variable is considered to be irrelevant and will be assigned a relatively large bandwidth. Favoring computational speed and applicability in high-dimensional settings, this approach does not generalize for arbitrary nonlinearities. This method assumes a known value of σ^2. If σ^2 is unknown, it has to be separately estimated.

In this paper, we take a adaptive regularized multivariate local regression approach by defining appropriate distributions over the parameters, combining it with an efficient bandwidth estimation method. We call it *sparse bandwidth selector* (*sbase* for short). The method includes the estimation of σ^2 and considers sparsity by analyzing the estimated bandwidths at each step. Several elements of the method are analogous to other approaches (which do not consider sparsity explicitly), as for example the work by Yang and Tschernig (1999). We expect that a suitable application of adaptive ridge regularization will further improve the bias–variance tradeoff of the estimation. We also propose an estimation of the regression coefficients though sampling methods, so that we obtain an estimate of the posterior distribution of the response.

The rest of the paper is organized as follows. Section 2 introduces the Bayesian hierarchical model. Section 3 describes the bandwidth selection procedure. Section 4 details how to obtain a maximum a posteriori (MAP) estimate of the response and the parameters. Section 5 introduces how to obtain a posterior distribution of the regression parameters and the response. Section 6 discusses the complexity of the algorithms. Section 7 provides some empirical examples of the performance of the proposed methods. Finally, we draw some conclusions in Section 8.