Information Sciences 348 (2016) 198–208

Contents lists available at ScienceDirect

Information Sciences

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

The mean shift algorithm and its relation to kernel regression

Youness Aliyari Ghassabeh[a,b,*], Frank Rudzicz[a,b]

[a] Toronto Rehabilitation Institute - UHN, 550 University Avenue, Toronto, Ontario, Canada
[b] Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

ARTICLE INFO

Article history:
Received 21 January 2015
Revised 6 January 2016
Accepted 8 February 2016
Available online 13 February 2016

Keywords:
Nadaraya–Watson kernel regression
Mean shift algorithm
Asymptotic bias
Gaussian kernel
Noisy source vector quantization

ABSTRACT

We investigate the connection between the asymptotic bias of the well-known Nadaraya–Watson kernel regression and the mean shift (MS) vector with the Gaussian kernel. We first show that the asymptotic bias for the univariate Nadaraya–Watson kernel regression can be estimated using the MS scalar. Then, we investigate the general D-dimensional case ($D > 1$) and derive a formula for the asymptotic bias as a function of the MS vector. We show that when the regression function is a linear function of its entries, then the asymptotic bias can be represented as a linear function of the MS vector. The MS algorithm for the univariate and the linear cases can be used to find points where the Nadaraya–Watson kernel regression gives an unbiased estimate. Furthermore, this connection suggests that exploiting the theoretical properties of the asymptotic bias of the Nadaraya–Watson kernel regression may be helpful to show the convergence of the MS algorithm. Through the simulations, we show that how the given theoretical results can be used to estimate the bias of the estimated clean signal by just observing the noisy signal. Having access to the bias of the estimator, enables us to evaluate the accuracy of the estimated clean data, which later will be used to design a quantizer.

1. Introduction

The Nadaraya–Watson kernel regression is a nonparametric, statistical technique that takes a limited set of data points as its input and estimates a continuous dependent variable using a kernel function that specifies the amount of contribution of these points. It is an important method for analyzing and finding relations between pairs of random variables ($\mathbf{X}$, Y), where $\mathbf{X}$ is an explanatory multidimensional vector and Y is a scalar response. Moreover, it provides a general relation between $\mathbf{X}$ and dependent variable Y without assuming any parametric assumptions on the observed data.

It is well-known that for a given value of $\mathbf{X} = \mathbf{x}$ the choice of expectation (i.e., regression) of Y given $\mathbf{x}$ should minimize the expected prediction error [21]. Since in general the density functions are not available, the Nadaraya–Watson estimator uses a set of paired observations $(\mathbf{x}_i, y_i), i = 1, \ldots, n$ for estimating the conditional expectation of the dependent variable, Y, given the independent variable, $\mathbf{X}$ [36]. In addition to fitting a curve to a data set, this form of multivariate kernel regression

* Corresponding author at: Toronto Rehabilitation Institute - UHN, 550 University Avenue, Toronto, Ontario, Canada. Tel.: +1 416 389 1498; fax: +1 416 597 3031.
E-mail addresses: aliyari@cs.toronto.ca, youness.aliyari@gmail.com (Y. Aliyari Ghassabeh), frank@cs.toronto.edu (F. Rudzicz).

http://dx.doi.org/10.1016/j.ins.2016.02.020
0020-0255/

has been used in many applications including image processing [28,35], classification [24], sensor fusion problems [32], and expert trading systems [38] (see [4] for other applications).

The performance of an estimator can be judged by looking at its properties, such as bias, mean square error, consistency, and variance. The asymptotic bias for the local linear least squares kernel estimators was studied by Ruppert and Wand and the leading bias term has been derived [33]. They showed that the leading term of the asymptotic bias does not depend to the observations $\mathbf{x}_1, \dots, \mathbf{x}_n$ [33]. Furthermore, the computed leading bias term in [33] is a function of curvature and smoothing, which implies that the bias term increases at the points with high curvature or when performing high amount of smoothing.

The mean shift (MS) technique is a simple nonparametric approach that was proposed by Fukunaga and Hostetler to estimate the modes (maxima) of a probability density function. Modes of a density function can be used to solve different problems in machine learning including image segmentation [37], clustering [25], object tracking [7,40], and edge detection [18]. A generalized version of the MS algorithm has been introduced by Cheng and its application for clustering has been discussed [5]. The MS algorithm became popular in the machine learning society when Comaniciu and Meer demonstrated its potential usage for feature space analysis [6]. The MS algorithm initialize the mode estimate sequence to an arbitrary point[1] and moves it to a weighted average of the neighboring points in each iteration. The algorithm repeats the previous procedure until the changes in output points become negligible.

We investigate the connection between the MS algorithm and the asymptotic bias of multivariate kernel regression. While Comaniciu and Meer noticed a connection between the MS algorithm and the bias of the kernel regression in one-dimensional space, they did not generalize this connection to multivariate kernel regression, instead focused on practical aspects of the MS algorithm [6]. In this study, we generalize this observation to multivariate kernel regression and show how the asymptotic bias for the multivariate Nadaraya–Watson estimator can be estimated using the MS algorithm with the widely used Gaussian kernel. To achieve this goal, we study the asymptotic bias of the multivariate Nadaraya–Watson kernel regression as the first step. In Theorem 1, we give an estimate for the asymptotic bias of the multivariate Nadaraya–Watson and use it for deriving a connection between the MS algorithm and the asymptotic bias of the Nadaraya–Watson kernel regression. We investigate the simple one-dimensional and general D-dimensional ($D > 1$) cases separately and derive formulas for the asymptotic bias as a function of the MS algorithm. For the one-dimensional and the multivariate cases we show through the simulations that how the provided theoretical results can be used to find the bias for the estimated clean data, when a sequence of noisy observations are available. It is very important problem in noisy source quantization problem, since the output of the estimator will be used later for quantizer design. By computing the estimator's bias, we will be able to judge about reliability of the estimated clean data, i.e., high bias means the estimated clean data is not accurate and we need to change the estimator's parameters.

In Sections 2 and 3, we provide a brief review of the MS algorithm and multivariate kernel regression. In Section 4, we derive the asymptotic bias for the Nadaraya–Watson estimator. The MS algorithm and its relation to kernel regression is discussed in Section 5. The simulation results are given in Section 6, where we illustrate the effectiveness of the theoretical results in the previous sections for estimating the asymptotic bias of the Nadaraya–Watson estimator. Section 7 is devoted to the concluding remarks.

2. Mean shift algorithm

Let $\mathbf{x}_i \in \mathbb{R}^D, i = 1, \dots, n$ denote a sequence of n independent and identically distributed random variables. In its most general form for an arbitrary point $\mathbf{x}$, the D-dimensional kernel density estimate $\hat{f}$ using a kernel $K_{\mathbf{H}}(\mathbf{x})$ is given by Silverman [34]

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i), \tag{1}$$

where $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$, $|.|$ denotes the determinant, and $\mathbf{H} \in \mathbb{R}^{D\times D}$ is a symmetric and positive definite matrix, which is called the bandwidth matrix. The kernel function $K : \mathbb{R}^D \to \mathbb{R}$ is a real-valued non-negative function that integrates to one, i.e.,$\int_{\mathbb{R}^D} K(\mathbf{x})d\mathbf{x} = 1$. Radially symmetric kernels have been widely used for probability density estimation [6] and defined by $K(\mathbf{x}) = c_{k,D}k(\|\mathbf{x}\|^2)$, where the constant $c_{k,\,D}$ is called normalization factor.

The profile function of the kernel K is given by k: $[0, \infty) \to [0, \infty)$ that is a decreasing function with bounded integral, i.e., $\int_0^\infty k(x)dx < \infty$ [5]. Using the bandwidth matrix $\mathbf{H}$ and the profile k, the estimated probability density function changes to the following well-known form [8]

$$\hat{f}(\mathbf{x}) = \frac{c_{k,D}|\mathbf{H}|^{-1/2}}{n}\sum_{i=1}^{n} k(\|\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{x}_i)\|^2). \tag{2}$$

[1] For simplicity, the observed data points are considered as starting points.