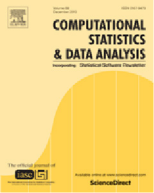
Computational Statistics and Data Analysis 102 (2016) 67–84

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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



On bandwidth selection using minimal spanning tree for kernel density estimation

Sreevani [*], C.A. Murthy

Machine Intelligence Unit, Indian Statistical Institute, Kolkata- 700108, India

ARTICLE INFO

Article history:
Received 1 June 2015
Received in revised form 19 January 2016
Accepted 19 April 2016
Available online 25 April 2016

Keywords:
Kernel density estimation
Bandwidth selection
Euclidean minimal spanning tree
Unbiased estimator

ABSTRACT

The use of kernel density estimation is quite well known in large variety of machine learning applications like classification, clustering, feature selection, etc. One of the major issues in the construction of kernel density estimators is the tuning of bandwidth parameter. Most of the bandwidth selection procedures optimize mean integrated squared or absolute error, which require huge computational time as the size of the data increases. Here, the bandwidth has been taken to be a function of inter-point distances of the data set. It is defined as a function of the length of Euclidean Minimal Spanning Tree of the given sample points. No rigorous theory about the asymptotic properties of the EMST based density estimator has been developed in the literature. Theoretical analysis of the asymptotic properties of the EMST based density estimator has been established and proved that the estimator is asymptotically unbiased to the original density at its every continuity point. Moreover, theoretical analysis has been provided for general kernel. Experiments are conducted using both synthetic and real-life data sets to compare the performance of the EMST bandwidth to those of conventional cross-validation and plug-in bandwidth selectors. It is found that the EMST based estimator achieves the comparative performance, while being simpler and faster than the conventional estimators.

1. Introduction

Many tasks in pattern recognition and machine learning often require the knowledge of underlying densities of the observed data (Menardi and Azzalini, 2014; Brown et al., 2012; Stover and Ulm, 2013; Brox et al., 2007; Ji et al., 2014; Jones and Rehg, 2002; Liu et al., 2007). For example in Bayes classification, the decision rule involves estimation of class conditional probabilities of the training data (Duda et al., 1999; Ramoni and Sebastiani, 2001; Kim and Scott, 2010). And, in model based clustering, every cluster corresponds to 'mode' or 'peak' in the estimated probability density of a given set of points (Li et al., 2007; Hinneburg and Gabriel, 2007; Tang et al., 2015). Estimation of the density can be done either in a parametric or non-parametric way. In parametric estimation, assumptions are made about the structure of the density, whereas in non-parametric estimation no assumptions are made about the form of the density function. Various methods have been studied for non-parametric density estimation such as histogram, kernel density estimator, spline estimators, orthogonal series estimators, etc., (Silverman, 1986; Scott, 2009; Golyandina et al., 2012). The kernel method is perhaps the most popular and well known technique of non-parametric estimation (Parzen, 1962; Cacoullos, 1966).

Throughout this article, we use the following notation. Let $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n \in \mathcal{R}^d$, $d \geq 2$ denote 'n' independent and identically distributed random vectors, and $\underline{X}_i = (X_{i1}, \ldots, X_{id})'$, $(.)'$ represents the transpose. A general vector $\underline{x}$ has the

[*] Corresponding author.
E-mail addresses: sreevani_r@isical.ac.in (Sreevani), murthy@isical.ac.in (C.A. Murthy).

http://dx.doi.org/10.1016/j.csda.2016.04.005
0167-9473/© 2016 Elsevier B.V. All rights reserved.

representation $\underline{x} = (x_1, \ldots, x_d)'$ and E(.) denotes expectation of a random vector. Also, $\int_{\mathcal{R}^d}$ will be shorthand for $\int_{\mathcal{R}} \ldots \int_{\mathcal{R}}$, $d\underline{x}$ will be shorthand for $dx_1 \ldots dx_d$ and $\int (.)\, d\underline{x}$ denotes $\int \ldots \int (.)\, dx_1..dx_d$.

A general d-dimensional kernel density estimator $\hat{f}$, for a random sample $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n$ with common probability density function f, is

$$\hat{f}_n(\underline{x}) = \frac{1}{n} \sum_{i=1}^{n} K_H(\underline{x} - \underline{X}_i),$$

where K_H is the scaled kernel function i.e., $K_H(\underline{x}) = |H|^{-1/2} K(H^{-1/2}\underline{x})$, K is a d-variate kernel function, H is a symmetric positive definite $d \times d$ bandwidth matrix and $| \cdot |$ is the determinant. Traditionally, K is assumed to be symmetric and $\int K(\underline{x})\, d\underline{x} = 1$. Some commonly used kernel functions are uniform, triangle, Epanechnikov, Gaussian, etc. The most widely used kernel is the Gaussian with zero mean and unit variance. From the above equation of $\hat{f}_n(\underline{x})$, it is clear that kernel density estimate at any test point $\underline{x}$ is simply the sum of kernel values caused by all training points $\underline{X}_i$. It is well known that the bandwidth selection is the most important and crucial step to obtain a good estimate. There are mainly two computational challenges associated with KDE; one is the selection of bandwidth, which is estimated using training data and the other is the construction of density at any test point. Note that the issue of bandwidth selection is the only problem considered in this article.

The bandwidth matrix H can be considered as a diagonal positive definite matrix i.e., $H = \text{diag}(h_1^2, \ldots, h_d^2)$, $h_i > 0, \forall\, i$ to simplify the above equation. Further simplification is obtained from the restriction, $h_i = h\ (>0), \forall\, i$, i.e., $H = \text{diag}(h^2, \ldots, h^2)$ and this leads to a single bandwidth kernel density estimator as,

$$\begin{aligned}\hat{f}_n\left(\underline{x}\right) &= \frac{1}{n} \sum_{i=1}^{n} K_h\left(\underline{x} - \underline{X}_i\right) \\ &= \frac{1}{nh^d} \sum_{i=1}^{n} K\left(h^{-1}\left(\underline{x} - \underline{X}_i\right)\right).\end{aligned}$$

A full bandwidth matrix provides more flexibility, but it also introduces more complexity into the estimator since more parameters need to be tuned (Wand and Jones, 1994). Although the selection of bandwidth can be done subjectively, but there is a great demand for automatic selection. Several automatic procedures compute optimal bandwidth value by minimizing the discrepancy between the estimate and target density by using some error criterion. A few such error criteria are given below (Wand and Jones, 1994).

- **Mean Squared Error (MSE)**: $MSE\left(\hat{f}_n(\underline{x})\right) = E\left\{\left(\hat{f}_n(\underline{x}) - f(\underline{x})\right)^2\right\}$.
- **Mean Integrated Squared Error (MISE)**: $MISE\left(\hat{f}_n\right) = E\left\{\int \left(\hat{f}_n(\underline{x}) - f(\underline{x})\right)^2 d\underline{x}\right\}$.
- **Mean Integrated Absolute Error (MIAE)**: $MIAE\left(\hat{f}_n\right) = E\left\{\int \left|\hat{f}_n(\underline{x}) - f(\underline{x})\right| d\underline{x}\right\}$.

Most of the modern bandwidth selectors are motivated by minimizing the MSE or MISE because these two criteria can be decomposed into variance and bias terms, as

$$\begin{aligned}MSE\left(\hat{f}_n(\underline{x})\right) &= \left(E\left[\hat{f}_n(\underline{x})\right] - f(x)\right)^2 + Var\hat{f}_n(\underline{x}) \\ &= \left(Bias\hat{f}_n(\underline{x})\right)^2 + Var\hat{f}_n(\underline{x}), \\ MISE\left(\hat{f}_n\right) &= E\left[\int \left(\hat{f}_n(\underline{x}) - f(\underline{x})\right)^2 d\underline{x}\right] \\ &= \int E\left[\left(\hat{f}_n(\underline{x}) - f(\underline{x})\right)^2\right] d\underline{x} \\ &= \int \text{MSE}\left(\hat{f}_n(\underline{x})\right).\end{aligned}$$

The optimal bandwidth that minimizes MISE, when the underlying density is a d-variate normal and the diagonal bandwidth matrix is employed, can be approximated by Silverman (1986),

$$h_i = \sigma_i \left\{\frac{4}{(d+2)n}\right\}^{\frac{1}{(d+4)}}, \quad \forall\, i = 1, 2, \ldots, d;$$

where σ_i is the standard deviation of the ith variable. This is called "Normal Reference (NR) rule". The most studied automatic bandwidth selector which aims to minimize MISE is the Least Squares Cross Validation (LSCV) (Bowman, 1984). The LSCV