Statistics and Probability Letters 117 (2016) 23–30

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Nonparametric estimation of possibly similar densities

Alan P. Ker

Department of Food, Agricultural and Resource Economics, University of Guelph, Canada

ARTICLE INFO

Article history:
Received 19 October 2015
Received in revised form 16 March 2016
Accepted 16 March 2016
Available online 10 May 2016

Keywords:
Multiple density estimation
Combined estimators
Kernel methods

ABSTRACT

A class of nonparametric methods is developed to estimate a set of *possibly* similar densities that offers greater efficiency if they are similar while seemingly not losing any if they are not. Theoretical properties and finite sample performance are promising.

1. Introduction

The standard univariate kernel density estimator (KDE) can be expressed as

$$\hat{f}(x) = (1/n)\sum K_h(x - X_i) \tag{1}$$

where h is the bandwidth or smoothing parameter, $K_h(u) = (1/h)K(u/h)$, K is the kernel function, and the summation is over the sample $(X_1, \ldots, X_n)$. Throughout, K is assumed to be a square integrable symmetric probability density function with a finite second moment and compact support. Also, denote $\mu_2(K) = \int u^2 K(u)du$ and $R(K) = \int K(u)^2 du$. Letting f be the unknown density of interest, standard properties for second order kernels are

$$\begin{aligned} E\hat{f}(x) - f(x) &= \int K(u)[f(x - hu) - f(x)]du = 1/2h^2\mu_2(K)f''(x) + O(h^4), \\ \mathrm{Var}(\hat{f}(x)) &= (nh)^{-1}f(x)R(K) + o((nh)^{-1}), \end{aligned} \tag{2}$$

and thus

$$\begin{aligned} MSE(\hat{f}(x)) &= (nh)^{-1}f(x)R(K) + 1/4h^4(\mu_2(K))^2(f''(x))^2 + o((nh)^{-1} + h^4), \\ MISE(\hat{f}) &= (nh)^{-1}R(K) + 1/4h^4(\mu_2(K))^2R(f'') + o((nh)^{-1} + h^4). \end{aligned} \tag{3}$$

There exist many empirical situations that require density estimates for multiple units which, to some unknown extent, are similar in structure. This manuscript considers an alternative but not uncommon data environment, one where there exist realizations $\{X_{11}, \ldots, X_{1n}, \ldots, X_{Q1}, \ldots, X_{Qn}\}$ from Q densities $f_1, \ldots, f_Q$ which may possibly be similar. The main objective is to design a nonparametric estimator that has superior performance, relative to the KDE applied separately to the individual samples, when the true densities are identical or similar, while not losing much if they are dissimilar. In addition, the estimator must not require the extent or form of similarity as this is generally unknown in empirical applications.

E-mail address: aker@uoguelph.ca.

http://dx.doi.org/10.1016/j.spl.2016.03.010
0167-7152/

The intuition of the proposed estimator is as follows: if the densities were known to be identical, the logical approach would be to pool the Q samples and estimate a single density. However, if the densities are not identical, this estimator is inconsistent. The idea proposed here is to combine a nonparametric estimate based on the pooled data with a nonparametric estimate based on the individual data in much the same fashion as a parametric estimate is combined with a nonparametric estimate in semiparametric estimation. These combined nonparametric and parametric estimators are designed with the same goal in mind: to offer superior performance if the underlying parametric assumption is correct while not losing much if it is incorrect. Hjort and Glad (1995) introduced an estimator which begins with a parametric estimate and then estimates a nonparametric correction in attempts to reduce bias. If the parametric start is sufficiently close to the true density, the correction factor function will be less rough, and thus estimated nonparametrically with lower bias. This estimator was shown to have promise in finite samples as well as nice asymptotic properties ($O(n^{-1})$ if the parametric assumption is correct and $O(n^{-4/5})$ if it is not). As such, their estimator represents an ideal starting point to develop a combined estimator in an expanded data environment.

2. Proposed estimator

We start with a KDE based on the pooled samples, denote $\hat{g}$, and then multiply a nonparametric estimate of the individual correction function, $r_i(x) = f_i(x)/\hat{g}(x)$. Note, $g(x)$ is a mixture of the Q densities, that is $g(x) = 1/Q \sum f_q(x)$.[1] The motivation is that if the densities are identical or similar, the pooled estimate represents a reasonable start from which to estimate a correction factor function for each individual density. The correction factor function is estimated by $\hat{r}_i(x) = \sum(1/n)K_h(x - X_{ij})/\hat{g}(X_{ij})$ thus leading to the proposed estimator

$$\tilde{f}_i(x) = \hat{g}(x)\hat{r}_i(x) = \sum(1/n)K_h(x - X_{ij})\frac{\hat{g}(x)}{\hat{g}(X_{ij})} \tag{4}$$

where X_i is the sample corresponding to density f_i.

The motivation behind the proposed estimator follows from Hjort and Glad (1995) and generalized by Naito (2004): reduce the global curvature of the underlying function being estimated thereby reducing bias. The correction factor function will have less global curvature if the start is sufficiently close to the unknown density. Unlike the combined parametric and nonparametric estimator, our start is nonparametric which begs the question: where do any possible efficiency gains come from? In contrast to the Hjort and Glad (1995) estimator, our approach makes use of extraneous sample data to estimate the start density. As a result, the total curvature that is being estimated with the individual sample may be reduced yielding a lower bias. Interestingly, the proposed estimator resembles the higher order bias estimator of Jones et al. (1995)

$$\bar{f}(x) = \hat{f}(x)\hat{r}(x) = \frac{1}{n}\sum K_h\left(x - X_j\right)\frac{\hat{f}(x)}{\hat{f}(X_j)}, \tag{5}$$

where $\hat{f}(x)$ is their start which is the KDE based on the individual sample.[2] Note that the three estimators – Jones et al. (1995) estimator, Hjort and Glad (1995) estimator, and the proposed estimator – are identical in form but each uses different start estimates. Hjort and Glad (1995) use a parametric start, Jones et al. (1995) use a nonparametric start based on the individual sample data only, and the proposed estimator uses a nonparametric start based on the pooled sample data. All reduce to the standard kernel when their start is uniform over the support. While this situation would be quite rare for the Jones et al. (1995) and Hjort and Glad (1995) estimators, this is less rare for the proposed estimator. If the underlying densities are quite dissimilar, the smoothing parameter associated with the pooled estimate gets large leading to a fairly uniform start and thus the proposed estimator collapses to the KDE.

3. Theory

Let $\hat{g}(x)$ be a KDE based on the pooled sample data. That is,

$$\hat{g}(x) = \frac{1}{Qn}\sum K_{h_p}(x - X_{iq}), \tag{6}$$

with properties

$$\begin{aligned} E\hat{g}(x) - g(x) &= 1/2h_p^2\mu_2(K)g''(x) + O(h_p^4), \quad \text{and} \\ \mathrm{Var}(\hat{g}(x)) &= (Nh_p)^{-1}g(x)R(K) + o((Nh_p)^{-1}), \end{aligned} \tag{7}$$

[1] The proposed estimator as presented assumes a balanced design. This ensures that g exists as the equally weighted mixture of the density functions $f_1, \ldots, f_Q$. An unbalanced design is inconsequential empirically as $\hat{g}$ is a simple kernel estimate based on the pooled sample. Theoretically however, in an unbalanced design the underlying mixture weights change as the sample size changes and thus g would be a function of the sample size which is unappealing.

[2] Jones et al. (1995) use the term *pilot* instead of *start*.