Journal of Statistical Planning and Inference 153 (2014) 75–86

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A bootstrap procedure for local semiparametric density estimation amid model uncertainties

Mehdi Soleymani, Stephen M.S. Lee *

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong

ARTICLE INFO

Article history:
Received 29 May 2012
Received in revised form
7 May 2014
Accepted 8 May 2014
Available online 22 May 2014

Keywords:
Bootstrap
Kernel density estimator
Local parametric fit
Maximum likelihood
Semiparametric estimation

ABSTRACT

We revisit a semiparametric procedure for density estimation based on a convex combination of a nonparametric kernel density estimator and a parametric maximum likelihood estimator, with the mixing weight locally estimated by the bootstrap method. We establish the asymptotic properties of the resulting semiparametric estimator, and show that undersmoothing at the bootstrap step is necessary if the estimator is to attain a convergence rate faster than that of the kernel density estimator under a good local parametric fit. A simulation study is conducted to investigate the finite-sample performance of the procedure. Exploiting its adaptivity to the goodness of local parametric fit, we propose a double bootstrap algorithm to incorporate into the semiparametric procedure more than one parametric family, and illustrate with a numerical example the benefits gained thereof.

1. Introduction

We consider a semiparametric approach to density estimation based on a convex combination of a nonparametric kernel density estimator and a parametric maximum likelihood estimator. Accuracy of the resulting semiparametric estimator depends on how well the mixing weight adapts to the goodness of fit of the parametric model. It is well known that the kernel density estimator, although robust against model mis-specification, has a slower convergence rate compared to that of the maximum likelihood estimator, should the latter be consistent for the true density. On the other hand, model mis-specification may render the maximum likelihood estimator drastically biased, although its variance is often smaller than that of the kernel estimator. Thus, for optimal performance of the semiparametric estimator, one would intuitively require that a zero weight be attached to the kernel density estimator if the parametric model is correctly specified, and to the maximum likelihood estimator otherwise. A number of empirical mixing weights have been proposed with the aforesaid optimal properties. For example, Olkin and Spiegelman (1987) estimate the mixing weight by maximising a global pseudo-likelihood function defined on the observed sample. Their semiparametric estimator succeeds in approximating the maximum likelihood estimate if the parametric model is correct, and the kernel density estimate if not. In a regression context, Fan and Ullah (1999) consider a similar semiparametric estimator of the regression function, and establish its asymptotic adaptivity to global discrepancies between the model-based and true regression functions. The same problem is studied by Mays et al. (2001), who estimate the mixing weight by cross-validation. In neither of the above works does the estimated mixing weight avail itself of the fact that even a mis-specified model may contain a density function accurately

* Corresponding author.
E-mail addresses: mehdi.soleymani@me.com (M. Soleymani), smslee@hku.hk (S.M.S. Lee).

http://dx.doi.org/10.1016/j.jspi.2014.05.004
0378-3758/

approximating the true density in a local sense. For better adaptivity to local discrepancies between the true density and its "best" parametric fit, a local estimate is therefore required of the mixing weight. That the bootstrap provides just such a local estimate has been explored in different semiparametric contexts: see, for example, Kouassi and Singh (1997) for semiparametric hazard estimation, or Yuan and Yin (2011) for semiparametric estimation of dose–response probabilities. However, neither of the above two works addresses the important issue concerning the theoretical effects the bootstrap may have on the semiparametric estimator. Lee (1994) uses the bootstrap approach to define a semiparametric estimator of a differentiable functional and studies its asymptotic properties under the assumption that the parametric and nonparametric estimators both have the same convergence rate $n^{1/2}$, a condition which does not hold for kernel density estimation.

Application of the bootstrap to estimate the mixing weight locally for semiparametric density estimation remains as yet unexplored, and will be the focus of the present paper. We shall establish theoretical properties of the mixing weights, estimated locally by the bootstrap, and the asymptotic mean squared error of the resulting semiparametric density estimator. To shed light on its local adaptivity properties, our results are deliberated with explicit reference to the local discrepancy between the true density function and the best approximating member of the assumed parametric model. Importantly, we show that for the semiparametric density estimator to capture the fast parametric convergence rate under a good local parametric fit, the bootstrap kernel density estimator must be sufficiently undersmoothed when calculating the estimated mixing weight. As an extension to the semiparametric approach, we explore also the possibility of incorporating more than one parametric family into the estimation procedure. We illustrate how the use of a set of parametric families can "harmlessly" increase the accuracy of density estimation. The rationale lies in the notion that, when a method is able to adapt itself to the uncertainty of a parametric model, one may make further improvement by plugging in a wider range of parametric families to enrich our choice of models, from which a "better" optimal solution may result.

Besides convex combinations, other semiparametric methods have been proposed for density estimation in the literature. For example, Efron and Tibshirani (1996) convert an essentially nonparametric setting to a parametric one by embedding the kernel density estimator in a specially designed exponential family, an approach involving no prior parametric model specification. Hjort and Glad (1995) multiply a parametric density estimate by a correction factor estimated using a kernel method. Hjort and Jones (1996) model the density by a parametric family but estimate the parameter by maximising a local kernel-smoothed likelihood function, thus giving the resulting density estimate a semiparametric flavour. The above three works share a common drawback in that their semiparametric density estimators always converge at the slow nonparametric rate even if a correct parametric family has been specified for the unknown density.

2. Semiparametric density estimator

Let $(X_1,\dots,X_n)$ be a random sample drawn from a univariate distribution F with density function $f=F'$. We consider the problem of estimating $f(x_0)$, for some fixed x_0. Assuming a parametric model for F, that is $F=F_\theta\in\mathcal{F}_\Theta=\{F_\vartheta:\vartheta\in\Theta\}$, where Θ denotes a parameter space in $\mathbb{R}^d$, we would naturally estimate $f(x_0)$ by the maximum likelihood estimator $f_{\hat\theta}(x_0)=F'_{\hat\theta}(x_0)$, where $\hat\theta$ is the maximum likelihood estimator, assumed unique, of θ. If we are uncertain about the plausibility of any parametric model, we may instead estimate $f(x_0)$ by the kernel density estimator $\hat f_{n,h}(x_0)=(nh)^{-1}\sum_{i=1}^n K((x_0-X_i)/h)$, for some kernel function K, which is taken to be a bounded and symmetric density function with zero mean and finite variance, and for some bandwidth $h>0$.

In order to capitalise on the possible benefits derived from both $f_{\hat\theta}(x_0)$ and $\hat f_{n,h}(x_0)$, we consider their convex combination $\hat g_{\epsilon,n,h}(x_0)=\epsilon f_{\hat\theta}(x_0)+(1-\epsilon)\hat f_{n,h}(x_0)$, $\epsilon\in[0,1]$, and choose ϵ to minimise the mean squared error of $\hat g_{\epsilon,n,h}(x_0)$, yielding the optimal mixing weight

$$\epsilon_n(x_0)=\zeta\left(\frac{\mathbb{E}_F[(\hat f_{n,h}(x_0)-f(x_0))(\hat f_{n,h}(x_0)-f_{\hat\theta}(x_0))]}{\mathbb{E}_F[(\hat f_{n,h}(x_0)-f_{\hat\theta}(x_0))^2]}\right),\tag{1}$$

where $\zeta(x)=\min\{1,\max\{0,x\}\}$.

The expression (1) depends on the unknown F and cannot be calculated directly. We propose to estimate ϵ_n by a bootstrap procedure as follows. Let $(X_1^*,\dots,X_n^*)$ be a generic bootstrap sample drawn from $(X_1,\dots,X_n)$ with replacement. Based on each bootstrap sample we obtain the corresponding maximum likelihood estimate $\hat\theta^*$ with respect to the model $\mathcal{F}_\Theta$, and the kernel density estimate $\hat f^*_{n,b}(x_0)=(nb)^{-1}\sum_{i=1}^n K((x_0-X_i^*)/b)$, for some bandwidth b possibly different from h. It has been found that undersmoothing with $b<h$ generally works well when the bootstrap is used for constructing confidence intervals for $f(x_0)$: see, for example, Hall (1992) and Ho and Lee (2008). Denote by $\mathbb{E}^*$ and $\mathbb{P}^*$ the expectation and probability measure, respectively, induced by bootstrap resampling, conditional on $(X_1,\dots,X_n)$. The bootstrap method estimates $\epsilon_n(x_0)$ by

$$\hat\epsilon_n(x_0)=\zeta\left(\frac{\mathbb{E}^*[(\hat f^*_{n,b}(x_0)-\hat f_{n,h}(x_0))(\hat f^*_{n,b}(x_0)-f_{\hat\theta^*}(x_0))]}{\mathbb{E}^*[(\hat f^*_{n,b}(x_0)-f_{\hat\theta^*}(x_0))^2]}\right),$$

which gives rise to our semiparametric estimator $\hat f(x_0)=\hat g_{\hat\epsilon_n(x_0),n,h}(x_0)$. The bandwidth h can be fixed at a conventional choice such as $h\propto n^{-1/5}$, which minimises the order of the mean squared error of $\hat f_{n,h}(x_0)$. The bandwidth b is chosen of order $o(h)$, the rationale behind which will be explained in Section 3.