Computational Statistics and Data Analysis 75 (2014) 248–264

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Computing confidence intervals for log-concave densities

Mahdis Azadbakhsh, Hanna Jankowski *, Xin Gao

Department of Mathematics and Statistics, York University, Toronto, Canada

ARTICLE INFO

Article history:
Received 20 October 2012
Received in revised form 22 January 2014
Accepted 24 January 2014
Available online 15 February 2014

Keywords:
Nonparametric density estimation
Log-concave
Maximum likelihood
Confidence interval

ABSTRACT

In Balabdaoui, Rufibach, and Wellner (2009), pointwise asymptotic theory was developed for the nonparametric maximum likelihood estimator of a log-concave density. Here, the practical aspects of their results are explored. Namely, the theory is used to develop pointwise confidence intervals for the true log-concave density. To do this, the quantiles of the limiting process are estimated and various ways of estimating the nuisance parameter appearing in the limit are studied. The finite sample size behavior of these estimated confidence intervals is then studied via a simulation study of the empirical coverage probabilities.

1. Introduction

The nonparametric maximum likelihood estimator (MLE) of a log-concave density has received much attention in the statistics literature of late. It has been studied, for example, in Walther (2002), Dümbgen and Rufibach (2009, 2011), Chang and Walther (2007), Chen and Samworth (2013), Cule et al. (2010) and Cule and Samworth (2010). For an overview, we recommend the review article of Walther (2009). The appeal of this estimator is that, unlike a kernel-density approach, it does not require a choice of bandwidth. Indeed, the log-concave MLE is not only fully automatic, but also automatically locally adaptive. Furthermore, the simulations in Chen and Samworth (2013, pp. 12–13) show that the log-concave MLE outperforms the kernel-density estimator for larger sample sizes, when the true density is log-concave. For smaller sample sizes, an (automatic) smoothed version of the MLE continues to have improved performance over the kernel-density estimator. (Chen and Samworth, 2013 consider the density on $\mathbb{R}^d$ with $d = 2, 3$.)

Here, we focus on the MLE of a log-concave density on $\mathbb{R}$. That is, let f_0 denote a log-concave density on $\mathbb{R}$ and suppose that we observe $X_1, \ldots, X_n$ independent and identically distributed samples from f_0. Let $\mathcal{F}$ denote the class of log-concave densities on $\mathbb{R}$. Then the nonparametric MLE of a log-concave density on $\mathbb{R}$ is defined as

$$\widehat{f}_n = \underset{f \in \mathcal{F}}{\operatorname{argmax}} \sum_{i=1}^{n} \log f(X_i).$$

Dümbgen and Rufibach (2009) show that this estimator exists and is unique, and also study its consistency. Additional results on consistency can also be found in Pal et al. (2007) and Cule and Samworth (2010). The estimator may be calculated using the active set algorithm, and this has been implemented in the R package *logcondens* (Dümbgen and Rufibach, 2006, 2011). Pointwise asymptotic theory for $\widehat{f}_n$ was developed in Balabdaoui et al. (2009).

* Corresponding author.
E-mail address: hkj@mathstat.yorku.ca (H. Jankowski).

http://dx.doi.org/10.1016/j.csda.2014.01.020

Suppose that the true density f_0 is log-concave with $f_0(x_0) > 0$ and $\varphi_0 = \log f_0$ is twice continuously differentiable in a neighborhood of x_0 with $\varphi_0''(x_0) \neq 0$. One of the main results of Balabdaoui et al. (2009) is that

$$n^{2/5}\left(\widehat{f}_n(x_0) - f_0(x_0)\right) \Rightarrow \left(\frac{f_0^3(x_0)|\varphi_0^{(2)}(x_0)|}{4!}\right)^{1/5} \mathbb{C}(0), \tag{1}$$

where the distribution of $\mathbb{C}(0)$ is known (here, we describe it in Section 2). For a fixed f_0, define

$$c_2(x) = \left(\frac{f_0^3(x)|\varphi_0^{(2)}(x)|}{4!}\right)^{1/5}. \tag{2}$$

If $c_2(x)$ is known, and q_α denotes the quantile such that $P(\mathbb{C}(0) \le q_\alpha) = \alpha$, then the result in (1) implies that

$$\left(\widehat{f}_n(x_0) - \frac{c_2(x_0)}{n^{2/5}}\, q_{1-\alpha/2},\ \widehat{f}_n(x_0) - \frac{c_2(x_0)}{n^{2/5}}\, q_{\alpha/2}\right) \tag{3}$$

forms an asymptotically correct $100(1-\alpha)\%$ confidence interval for $f_0(x_0)$. The main goal of this paper is to provide estimators for the quantiles q_α and $c_2(x_0)$ so that the confidence intervals (3) may be implemented in practice, and to assess the quality of this procedure.

In Section 2 we describe the process $\mathbb{C}(0)$ and provide its quantile estimates based on simulations (the simulations are detailed in Appendix A). In Section 3 we consider estimation of the constant c_2, and in Section 4 we use simulations to understand the empirical performance of the estimated confidence intervals (3). The methods presented here have been implemented in the R package *logcondens* (Dümbgen and Rufibach, 2006).

2. Quantiles of the limiting process

Let $\mathbb{B}(t),\ t \in \mathbb{R}$ denote a two-sided Brownian motion. That is, $\mathbb{B}(t) = \mathbb{B}_1(t), t \ge 0$ and $\mathbb{B}(t) = \mathbb{B}_2(-t), t \le 0$, where $\mathbb{B}_1, \mathbb{B}_2$ are two independent Brownian motions with $\mathbb{B}_1(0) = \mathbb{B}_2(0) = 0$. Let

$$\mathbb{Y}(t) = \begin{cases} \int_0^t \mathbb{B}(s)ds - t^4, & t \ge 0 \\ \int_t^0 \mathbb{B}(s)ds - t^4, & t < 0, \end{cases}$$

and let $\mathbb{H}$ be the almost surely unique process such that

1. $\mathbb{H}(t) \le \mathbb{Y}(t)$ for all $t \in \mathbb{R}$,
2. $\mathbb{H}''(t)$ is concave,
3. $\mathbb{H}(t) = \mathbb{Y}(t)$ if the slope of $\mathbb{H}''(t)$ is strictly decreasing at t.

The process $\mathbb{H}$ thus defined exists and is unique (Balabdaoui et al., 2009, Theorem 2.1). Let $\mathbb{C}(t) = \mathbb{H}''(t)$ for all $t \in \mathbb{R}$, then the quantity of interest, $\mathbb{C}(0)$, is simply $\mathbb{C}(t)$ evaluated at $t = 0$.

The process $\mathbb{H}$, or rather its close relative, was first shown to exist in Groeneboom et al. (2001a), and we refer to Appendix A for further details. Using their approach, one could show that $\mathbb{C}(t) = \lim_{m\to\infty} \mathbb{C}_m(t)$, where $\mathbb{C}_m$ is defined as:

$$\mathbb{C}_m = \underset{\varphi \in \mathcal{C}_m}{\operatorname{argmin}} \left\{ \int_{-m}^{m} \varphi^2(t)dt - 2\int_{-m}^{m} \varphi(t)d\left(\mathbb{B}(t) - 4t^3\right) \right\},$$

where $\mathcal{C}_m$ denotes the class of concave functions with the restriction that $\varphi(-m) = \varphi(m) = -12m^2$. Thus, we can think of $\mathbb{C}(t)$ as the concave regression on the function $-12t^2$ plus white noise.

Our interest here is limited to the value of $\mathbb{C}(t)$ at $t = 0$, and below we present some observed properties based on $n = 100\,000$ independent samples. Details on the algorithm used to generate these samples is given in Appendix A. Fig. 1 shows the estimate of the density of $\mathbb{C}(0)$. Visually, two things are immediately striking: first, the density appears to be asymmetric (right-skewed), and second, the density appears to be log-concave. We address these two questions in Sections 2.1 and 2.2.

Moment (plus median) estimates of $\mathbb{C}(0)$ are given in Table 1 while quantile estimates are given in Table 2. Table 2 gives four different quantile estimates based on

(A) the empirical distribution function,
(B) the kernel density estimate,
(C) the log-concave MLE,
(D) the normal approximation.

The last column of the table gives standard errors of the values in column (A), (see Shorack and Wellner, 1986, Example 1, p. 639). The Gaussian approximation is given for reference only. Our simulations indicate that $E[\mathbb{C}(0)] = 0$, as expected.