Proceedings of the 2011 Winter Simulation Conference
S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, eds.

A CROSS-VALIDATION APPROACH TO BANDWIDTH SELECTION FOR A KERNEL-BASED ESTIMATE OF THE DENSITY OF A CONDITIONAL EXPECTATION

Athanassios N. Avramidis

School of Mathematics, University of Southampton
Highfield, Southampton
SO17 1BJ, UNITED KINGDOM

ABSTRACT

To estimate the density f of a conditional expectation $\mu(Z) = \mathbb{E}[X|Z]$, Steckley and Henderson (2003) sample independent copies $Z_1, \ldots, Z_m$; then, conditional on Z_i, they sample n independent samples of X, and their sample mean $\bar{X}_i$ is an approximate sample of $\mu(Z_i)$. For a kernel density estimate $\hat{f}$ of f based on such samples and a bandwidth (smoothing parameter) h, they consider the mean integrated squared error (MISE), $\int(\hat{f}(x) - f(x))^2 \mathrm{d}x$, and find rates of convergence of m, n and h that optimize the rate of convergence of MISE to zero. Inspired by the cross-validation approach in classical density estimation, we develop an estimate of MISE (up to a constant) and select the h that minimizes this estimate. While a convergence analysis is lacking, numerical results suggest that our method is promising.

1 INTRODUCTION

The problem of dealing with uncertainty in simulation models has received considerable attention in the proceedings of the Winter Simulation Conference.

The problem motivating this paper is studied in Steckley and Henderson (2003) and is as follows. Let Z and X be random variables, and put $\mu(z) = \mathbb{E}[X|Z = z]$, where $\mathbb{E}$ denotes expectation. Provided that the random variable $\mu(Z)$ has a density (with respect to Lebesgue measure), say f, the problem of interest is to estimate f. It is assumed that the function $\mu()$ is unknown, but it can be estimated, as it is possible to: (i) sample Z; and (ii) sample X from the conditional distribution $\mathbb{P}(X \in \cdot|Z = z)$ for any possible realization z. This problem arises, for example, when Z is a parameter that is uncertain and a system's performance is modeled as the expectation $\mathbb{E}[X|Z]$. Then, the density f summarizes the uncertainty in performance due to uncertainty about Z.

The paper is organized as follows. The approach of Steckley and Henderson (2003) is reviewed in Section 2.1. In Section 2.2 we develop a related approach, where the focus is on the bandwidth, which is set differently. We review briefly the least-squares *cross-validation* approach to bandwidth selection in the classical setting where exact samples from the unknown density are available. Then we adapt this approach to our setting, where the samples are not exact. In Section 3 we give numerical results, comparing the MISE of our estimate to that of Steckley and Henderson (2003) in examples taken from these authors.

2 METHODOLOGY

2.1 Background

This section follows closely the method in Steckley and Henderson (2003). We include some of their assumptions, not aiming for completeness, but instead to suit our later needs. It is assumed:

A1. Z has density g
A2. $\mu()$ is strictly monotone and its inverse, μ^{-1}, is differentiable.

Note that A1–A2 imply that the conditional expectation $\mu(Z)$ has a density (see Billingsley (1986), equations (20.16) and (20.20)). Sampling is as follows:

1. (Outer sampling) $Z_1,\ldots,Z_n$ are independent samples of Z.
2. (Inner sampling) Conditional on $Z_1,\ldots,Z_n$, $(X_{i,j} : i = 1,\ldots,n, j = 1,\ldots,m)$ are mutually independent and such that $X_{i,1},\ldots,X_{i,m}$ are samples from the distribution $\mathbb{P}(X \in \cdot | Z = Z_i)$ for each i.

Steckley and Henderson (2003) propose the kernel estimate of f

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K(x - \bar{X}(Z_i);\, h^2) \tag{1}$$

where

$$\bar{X}_i = \bar{X}(Z_i) = \frac{1}{m}\sum_{j=1}^{m} X_{i,j}$$

and the chosen kernel is $K(x;\, h^2) = e^{-x^2/2h^2}/(h\sqrt{2\pi})$, that is, the density at x of a mean-zero Normal distribution whose standard deviation h is called the bandwidth. Further assume:

A3. The conditional distribution of the inner sample mean $\bar{X}(Z)$ given Z is Normal with mean $\mu(Z)$ and variance $\sigma^2(Z)/m$, where $\sigma^2(z) = \mathrm{Var}(X|Z = z)$ is the *variance function*.

The mean integrated squared error (MISE) is $\mathbb{E}L$, where

$$L = L(h) = L(n,m,h) = \int (\hat{f}(x) - f(x))^2 \mathrm{d}x \tag{2}$$

is the integrated squared error, and n, m, and h are functions of a computer budget b. Under A1-A3 and further assumptions, Steckley and Henderson (2003) develop an expansion of $\mathbb{E}L$. To restate their result, assume the variance function is constant, say σ^2. Their equation (1) expands MISE into: an integrated squared bias term of size $O(h^2 + \sigma^2/m)$; an integrated variance term of size $O(1/nh)$; and a remainder term. In the limit where $h \to 0$, $m \to \infty$, and $nh \to \infty$, they show that MISE tends to zero. Subject to the computing-budget constraint $nm = b$, and with the bandwidth set as $h = am^{-1/\delta}$ for constants a and δ, they show that as $b \to \infty$, the best *rate* of convergence of $\mathbb{E}L$ to zero is attained by $\delta = 2$ and $m = rb^{2/7}$ for a constant r (see page 387, where r is called d and is a function of a and properties of f) which implies $h = O(b^{-1/7})$, $n = O(b^{5/7})$, and $\mathbb{E}L = O(b^{-4/7})$. This regime gives optimal growth rates, up to the unrestricted constant r (or a).

We will consider a density estimator that is of the form (1), where n and m grow as specified above, but the bandwidth h is set differently, as explained in the next section.

2.2 Bandwidth Selection

By *standard setting*, we mean that–unlike our setting–a random sample $X_1,\ldots,X_n$ is available directly from a target density f. Here, and with $\mathbb{E}L$ being the quality measure, an established method for bandwidth selection is *least-squares cross-validation*. We review the basic idea very briefly, following Silverman (1986), Section 3.4.3. Write

$$L_0(h) = L - \int f^2 = \int \hat{f}^2 - 2\int \hat{f} f.$$

(Integrals may be written as above to lighten notation.) The basic idea is to construct an estimate of $L_0(h)$ and then to select h to minimize this estimate. Define $\hat{f}_{-i}$ as the density estimate constructed from all data points *except* i, that is, $\hat{f}_{-i}(x) = (n-1)^{-1}\sum_{j:j\neq i} K(x - X_j;\, h^2)$. It is easy to see that $n^{-1}\sum_i \hat{f}_{-i}(X_i)$ is an unbiased estimate of $\mathbb{E}\int \hat{f} f$. Thus, $M_0(h) = \int \hat{f}^2 - 2n^{-1}\sum_i \hat{f}_{-i}(X_i)$ is an unbiased estimate of $\mathbb{E}L_0(h)$ for

each h. The minimizer of $M_0(h)$ is the selected bandwidth. It is hoped that this minimizer is close to the minimizer of $\mathbb{E}L_0(h)$, which coincides with the minimizer of $\mathbb{E}L(h)$, as the term $\int f^2$ is independent of h. This leads to good large-sample properties (Hall 1983, Stone 1984).

In our setting, the mean $\bar{X}_i$ of any inner sample i is no longer an exact sample from f, so the work above does not seem to apply directly. Motivated by the work in the standard setting, we seek an unbiased estimate of $\mathbb{E}\int \hat{f}f$. We observe that $\int \hat{f}(x)f(x)\mathrm{d}x = n^{-1}\sum_i \int K(x-\bar{X}(Z_i);\, h^2)f(x)\mathrm{d}x$ and examine its expectation:

$$\begin{aligned}\mathbb{E}\int \hat{f}(x)f(x)\mathrm{d}x &= \mathbb{E}\int K(x-\bar{X}(Z_1);\, h^2)f(x)\mathrm{d}x \\ &= \int \mathbb{E}\left[\int K(x-\bar{X}(Z_1);\, h^2)f(x)\mathrm{d}x \middle| Z_1 = z\right] g(z)\mathrm{d}z. \end{aligned} \tag{3}$$

Now

$$\begin{aligned}\mathbb{E}\left[\int K(x-\bar{X}(Z_1);\, h^2)f(x)\mathrm{d}x \middle| Z_1 = z\right] &= \int\int K(x-y;\, h^2)f(x)K\left(y-\mu(z);\, \frac{\sigma^2(z)}{m}\right)\mathrm{d}x\mathrm{d}y \\ &= \int K\left(x-\mu(z);\, h^2+\frac{\sigma^2(z)}{m}\right)f(x)\mathrm{d}x.\end{aligned}$$

The first step above uses A3, the normality of the sample mean. In the second step, the integral with respect to y is a convolution of normal densities, giving a normal density whose variance is the sum of variances of the densities being convoluted. (Similar observations are made in Steckley and Henderson (2003), but with somewhat different aims.) Inserting into (3),

$$\begin{aligned}\mathbb{E}\int \hat{f}(x)f(x)\mathrm{d}x &= \int\int K\left(x-\mu(z);\, h^2+\frac{\sigma^2(z)}{m}\right)f(x)g(z)\mathrm{d}x\mathrm{d}z \\ &= \mathbb{E}K\left(\mu(Z_2)-\mu(Z_1);\, h^2+\frac{\sigma^2(Z_1)}{m}\right)\end{aligned} \tag{4}$$

where Z_1, Z_2 are independent random variables each having density g.

Let $i \neq k$, and consider the quantity $K\left(\bar{X}_k-\bar{X}_i;\, h^2+S_i^2/m\right)$, where

$$S_i^2 = (m-1)^{-1}\sum_{j=1}^{m}(X_{i,j}-\bar{X}_i)^2$$

is the i-th inner sample variance. This is a natural estimate of (4), and averaging these estimates over all such i,k leads to

$$\frac{1}{n(n-1)}\sum_i\sum_{k:k\neq i} K\left(\bar{X}_k-\bar{X}_i;\, h^2+\frac{S_i^2}{m}\right) \tag{5}$$

as a (cross-validation) estimate of (4). Observing that $\int \hat{f}^2(x)\mathrm{d}x = n^{-2}\sum_i\sum_j K(\bar{X}_i-\bar{X}_j;\, 2h^2)$ leads to

$$M(h) = \frac{1}{n^2}\sum_i\sum_j K(\bar{X}_i-\bar{X}_j;\, 2h^2) - \frac{2}{n(n-1)}\sum_i\sum_{k:k\neq i} K\left(\bar{X}_k-\bar{X}_i;\, h^2+\frac{S_i^2}{m}\right)$$

as an estimate of $\mathbb{E}L(h)-\int f^2$.

For a computing budget b, the proposed estimate of f is as in (1), where: $m = \max(2, [rb^{2/7}])$ and $n = \max(2, [b/m])$, where $[x]$ is the integer closest to x; h is the minimizer of $M(h)$; and r is a user-selected constant. If we additionally restrict the minimizer of $M(h)$ to be of order $m^{-1/2}$ (or equivalently of order $b^{-1/7}$), then we know that under the assumptions in Steckley and Henderson (2003), MISE obtains the optimal convergence rate; and it is then hoped to reduce MISE compared to these authors' choice $h = am^{-1/2}$, where guidance on a seems to be lacking. We adopt this restriction and give the details later.

3 NUMERICAL RESULTS

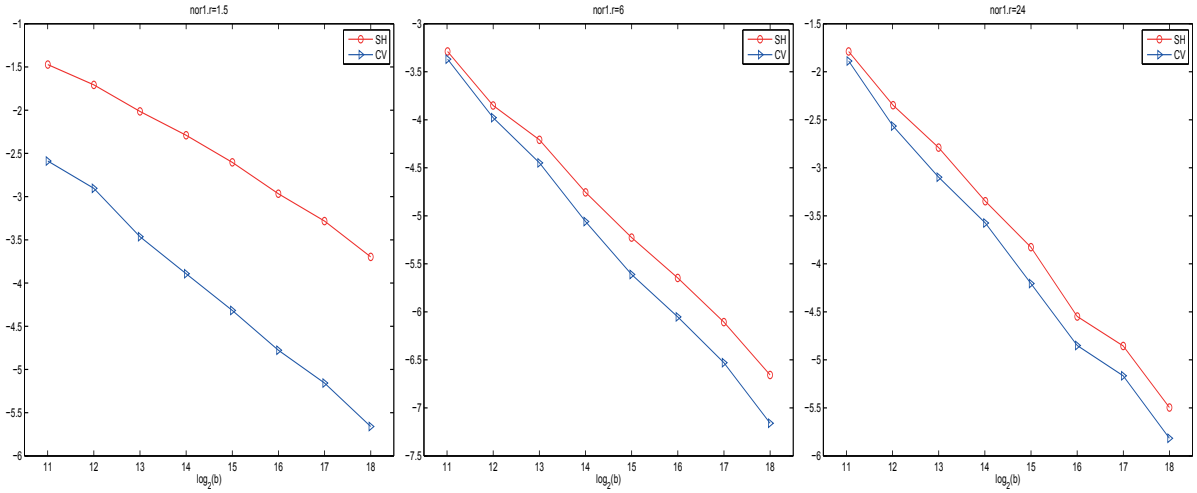
We report on Examples 1 and 3 of Steckley and Henderson (2003). Preliminary experiments suggested that $M(h)$ has a unique minimizer on $(0,\infty)$ that is roughly of order $m^{-1/2}$, which makes it plausible that the minimizer obeys the same convergence rate as prescribed by the asymptotic optimality framework of Steckley and Henderson (2003). We wanted to compare empirically the choice $h = h_{\mathrm{SH}} = m^{-1/2}$ of Steckley and Henderson (2003) to the choice

$$h = h_{\mathrm{CV}} = \operatorname*{argmin}_{m^{-1/2}/32 \le h \le 32m^{-1/2}} M(h). \tag{6}$$

For simplicity we solve the minimization problem approximately: we discretize the above interval into 201 equally-spaced points in logarithmic scale and call the minimizer $\hat{h}_{\mathrm{CV}}$. Although the constant "32" above is arbitrary, we observed that $\hat{h}_{\mathrm{CV}}$ was never an endpoint of the specified interval. This observation and the denseness of the discretization gave us confidence that the MISEs associated to $\hat{h}_{\mathrm{CV}}$, h_{CV}, and $\mathrm{argmin}_{h>0} M(h)$ were nearly the same for the purpose of comparison against the MISE of h_{SH}.

We report the MISEs $\mathbb{E}L(h_{\mathrm{SH}})$ and $\mathbb{E}L(\hat{h}_{\mathrm{CV}})$, each estimated as the average of 100 independent replications (25 replications in Example 1, case $r = 1.5$ below) for budgets $b \in \{2^{11}, 2^{12}, \ldots, 2^{18}\}$. We recall that r controls a trade-off between outer and inner sample size. In preliminary experiments, we found values of r that roughly minimize the (estimated) MISE of each method at the largest budget. These numbers, denoted r^*_{SH} and r^*_{CV}, help us show MISE at its near-optimum and away from it.

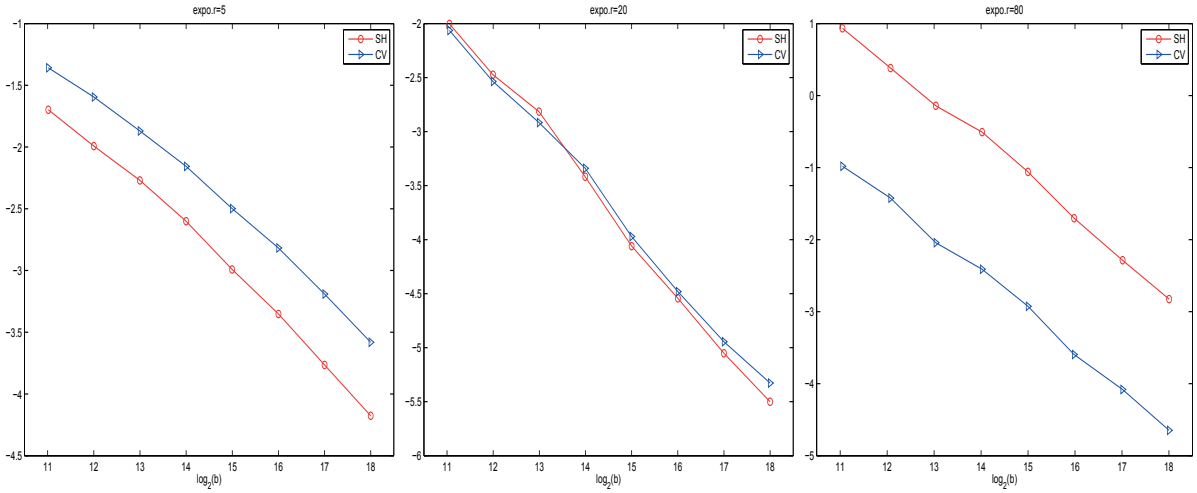
Example 1. Here $Z \sim \mathrm{Beta}(4,4)$. (A $\mathrm{Beta}(a,b)$ random variable has density on (0,1) proportional to $x^{a-1}(1-x)^{b-1}$.) Conditional on $Z = z$, X has the Normal distribution with mean z and variance 0.5. Thus f is the Beta(4,4) density. Here, we found $r^*_{\mathrm{SH}} \approx r^*_{\mathrm{CV}} \approx 6$; then we set $r \in \{1.5, 6, 24\}$, that is, at the optimum, and also up and down from the optimum by a factor of four. Figure 1 shows $\log_2(MISE)$ versus $\log_2(b)$.



(a) $r = 1.5$ (b) $r = 6$ (c) $r = 24$

Figure 1: Mean Integrated Squared Error of the Two Methods for Example 1 and Several r.

Example 2. Here Z has a Beta(4,4) density shifted to the right by one unit, so the support is (1,2). Conditional on $Z = z$, $X \sim \mathrm{Expon}(z)$, the exponential distribution with mean z. Thus f is the Beta(4,4) density. Here, we found $r^*_{\mathrm{SH}} \approx 16$ and $r^*_{\mathrm{CV}} \approx 24$, and we set r to 20 and also up and down from this by a factor of four. Figure 2 shows $\log_2(MISE)$ versus $\log_2(b)$.



(a) $r = 5$ (b) $r = 20$ (c) $r = 80$

Figure 2: Mean Integrated Squared Error of the Two Methods for Example 2 and Several r.

The results shown are representative of a larger set of experiments. Neither estimate showed a uniform advantage. The new density estimate yielded smaller MISE more often than not.

4 CONCLUSION

We developed an approach to choosing the bandwidth when estimating the density of a conditional expectation by a kernel-based method, where outer and inner sample sizes are prescribed by optimal asymptotics established in Steckley and Henderson (2003). The method is based on a cross-validation estimate of the expected squared error and determines the bandwidth by minimizing this estimate. In our experiments, the new density estimate yielded notably smaller MISE often, but not always.

REFERENCES

Billingsley, P. 1986. *Probability and Measure*. second ed. New York: Wiley.

Hall, P. 1983. "Large Sample Optimality of Least Squares Cross-Validation in Density Estimation". *Annals of Statistics* 11 (4): 1156–1174.

Silverman, B. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Steckley, S. G., and S. G. Henderson. 2003. "A kernel approach to estimating the density of a conditional expectation". In *Proceedings of the 2003 Winter Simulation Conference*, Edited by Stephen E. Chick, Paul J. Sanchez, David M. Ferrin, and Douglas J. Morrice, 383–391. Piscataway, New Jersey: Institute of Eletrical and Electronics Engineers, Inc.

Stone, C. J. 1984. "An asymptotically Optimal Window Selection Rule for Kernel Density Estimates". *The Annals of Statistics* 12 (4): 1285–1297.

AUTHOR BIOGRAPHY

ATHANASSIOS (THANOS) N. AVRAMIDIS is Lecturer in Operational Research in the School of Mathematics at the University of Southampton, United Kingdom. His main research interests are Monte Carlo and discrete-event simulation with emphasis on efficiency improvement and the interface to probability and statistics. Another research area is stochastic modeling of industrial and service systems. His recent research articles are available on-line from his web page: http://www.personal.soton.ac.uk/aa1w07.