BU-1223-M

A Note on Bandwidth Selection in Kernel Distribution Function Estimation

Naomi Altman *
Biometrics Unit
Cornell University

Christian Léger †
Département IRO
Université de Montréal

September 1993

Abstract

Sarda (1993) introduced a data dependent method to choose the bandwidth of a kernel distribution function estimator and showed this method was asymptotically optimal. In analogy with "least squares cross-validation" for density estimation, the bandwidth selector uses a leave-one-out estimator of the distribution function. In this note we show that optimality is also achieved when no observations are removed. Unfortunately, simulations show that neither method works in practice, even for samples of size as large as 1000.

AMS subject classification: Primary 62 G05 Secondary 62 G20
Keywords and phrases: Distribution function;nonparametric estimation;smoothing parameter selection;cross-validation;leave-one-out estimator.

*Supported by Hatch Grant 151410 NYF

†Supported by NSERC (Canada) and FCAR (Québec)

1 Introduction

Sarda (1993) introduced a data dependent procedure to choose the bandwidth of a kernel distribution function estimator (KDFE) and showed that this procedure is asymptotically optimal in a certain sense. The procedure uses a leave-one-out KDFE to estimate the average squared error of the kernel estimator. In this note, we show that it is not necessary to use a leave-one-out KDFE; the bandwidth which minimizes estimates of the average squared error based on the full-data KDFE is also asymptotically optimal. However, simulations show that neither procedure is practical: both procedures almost always choose the smallest possible bandwidth, even for samples as large as 1000.

We describe the kernel distribution function estimator and the estimators of average squared error in the next section. In Section 3, we show that the asymptotic optimality result of Sarda (1993) also holds for our leave-none-out estimator. Section 4 contains simulation results that show that neither method is useful in practice.

2 The leave-none-out procedure

Let $X_1, \ldots, X_n$ be distributed identically and independently from distribution function F. The kernel distribution function estimator $\hat{F}_h$ was introduced by Nadaraya (1964) and is defined by

$$\hat{F}_h(x) = n^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.1}$$

where K is the distribution function of a kernel k, $K(x) = \int_{-\infty}^{x} k(t)\,dt$ and h is the bandwidth.

A measure of quality of the kernel estimator is the Mean Integrated Squared Error

$$MISE(h) = E \int (\hat{F}_h(x) - F(x))^2 w(x) dF(x), \tag{2.2}$$

where w is a nonnegative weight function. A discrete approximation of $MISE$ is the average squared error

$$ASE(h) = n^{-1} \sum_{j=1}^{n} [\hat{F}_h(X_j) - F(X_j)]^2 w(X_j). \tag{2.3}$$

Sarda's estimator of $ASE(h)$ is

$$CV(h) = n^{-1} \sum_{j=1}^{n} [\hat{F}_{h;-j}(X_j) - F_n(X_j)]^2 w(X_j), \tag{2.4}$$

where F_n is the empirical distribution function and $\hat{F}_{h;-j}$ is the kernel estimator computed by leaving out X_j. To choose the bandwidth, $CV(h)$ is minimized.

The estimator $CV(h)$ has been called a cross-validation criterion by Sarda. In a review of bandwidth selection procedures, Altman and Léger (1993) argue that it may be more appropriate to call it a leave-one-out estimator and to reserve cross-validation for procedures that estimate prediction risk by validating a predictor on independent future observations as in Stone (1974). We introduce a leave-none-out estimator of $ASE(h)$:

$$LNO(h) = n^{-1} \sum_{j=1}^{n} [\hat{F}_h(X_j) - F_n(X_j)]^2 w(X_j). \tag{2.5}$$

Sarda considered such an estimator, but argued that "the resulting score function will produce a very small bandwidth." The simulations in Section 4 confirm this statement, but using a leave-one-out estimator does not solve the problem. In fact, the two criteria are so close that the leave-none-out estimator $LNO(h)$ is also asymptotically optimal, although neither criterion is useful in practice.

3 Asymptotic optimality of $LNO(h)$

To establish the asymptotic optimality of $LNO(h)$, we use conditions (A.1)–(A.5) of Sarda (1993). First,

w is bounded and supported on a compact set. (A.1)

The set of bandwidths considered is

$$H_n = [C_1 n^{-a}, C_2 n^{-b}], \qquad 1/4 < b \leq a < 1/2. \tag{A.2}$$

The function K is absolutely continuous and

$$\lim_{x \to -\infty} K(x) = 0 \text{ and } \lim_{x \to \infty} K(x) = 1. \tag{A.3}$$

For $k = K'$, we have

$$\int x k(x)\, dx = 0 \text{ and } \int x^2 k(x)\, dx < \infty \tag{A.4}$$

Also, F verifies

F is twice differentiable and

F and $|f'|$ are bounded from below on the support of w. (A.5)

Let $d(\hat{F}_h, F)$ be either $MISE(h)$, $ASE(h)$ or $ISE(h)$ where

$$ISE(h) = \int (\hat{F}_h(x) - F(x))^2 w(x)\, dF(x), \tag{3.1}$$

and let

$$d'(\hat{F}_h, F) = d(\hat{F}_h, F) - V_1 n^{-1}, \tag{3.2}$$

with $V_1 = \int F(x)(1 - F(x))w(x)\, dF(x)$. We then have the following theorem.

Theorem 1 *Under the assumptions (A.1)-(A.5), the bandwidth $\hat{h}$ minimizing $LNO(h)$ over H_n is asymptotically optimal with respect to d' in the sense that*

$$lim_{n\to\infty} \frac{d'(\hat{F}_{\hat{h}}, F)}{\inf_{h\in H_n} d'(\hat{F}_h, F)} = 1 \qquad a.s. \tag{3.3}$$

Proof: After using algebra similar to the proof in Sarda, we have

$$\begin{aligned} CV(h) \;-\; LNO(h) &= n^{-1}\sum_{j=1}^{n}[\hat{F}_{h;j}(X_j) - \hat{F}_h(X_j)]^2 w(X_j) + \\ &\quad 2n^{-1}\sum_{j=1}^{n}(\hat{F}_{h;j}(X_j) - \hat{F}_h(X_j))(\hat{F}_h(X_j) - F_n(X_j))w(X_j) \\ &= n^{-1}\sum_{j=1}^{n} A_j^2 w(X_j) + 2n^{-1}\sum_{j=1}^{n} A_j(\hat{F}_h(X_j) - F_n(X_j)w(X_j) \\ &= (\overline{ASE}(h) - ASE(h)) + 2n^{-1}\sum_{j=1}^{n} A_j(F(X_j) - F_n(X_j))w(X_j) \\ &= T_1(h) + T_2(h), \end{aligned}$$

where as in Sarda, $A_j = (n-1)^{-1}(\hat{F}_h(X_j) - K(0))$ and $\overline{ASE}(h) = n^{-1}\sum_{j=1}^{n}[\hat{F}_{h;j}(X_j) - F(X_j)]^2 w(X_j)$.

Using Sarda's theorem, all we need to show is

$$\sup_{h\in H_n}\left|\frac{CV(h) - LNO(h)}{MISE'(h)}\right| \to 0 \qquad a.s.$$

From the proof of Lemma 2 of Sarda,

$$\sup_{h\in H_n}\left|\frac{T_1(h)}{MISE'(h)}\right| \to 0, \qquad a.s.$$

Also, using inequality (3.2) of Sarda, assumption (A.2), and for $a < 1/2$, we have

$$\begin{aligned}\sup_{h\in H_n}\left|\frac{T_2(h)}{MISE'(h)}\right| &\leq Cn^{1+a}\sup_{h\in H_n}|T_2(h)| \\ &\leq 4C'n^a\sup_x|F_n(x)-F(x)|,\end{aligned}$$

since w is bounded from (A.1) and $|A_j| \leq 2/(n-1)$. Now, Smirnov's law of the iterated logarithm (Shorack and Wellner, 1986, p. 504) states that

$$\overline{\lim_{n\to\infty}}\frac{\sqrt{n}\sup_x|F_n(x)-F(x)|}{\sqrt{2\log_2 n}} = 1/2 \qquad a.s.$$

Hence since $a < 1/2$, we have that

$$\sup_{h\in H_n}\left|\frac{T_2(h)}{MISE'(h)}\right| \to 0 \qquad a.s.$$

Remark: Unlike the density estimation problem, it is unnecessary to leave out an observation because the bandwidth h is not in the denominator of $\hat{F}_h$. Hence $\hat{F}_{h;j}(X_j) - \hat{F}_h(X_j) = (n-1)^{-1}\hat{F}_h(X_j) - (n-1)^{-1}K(0)$ which is small compared to $MISE'(h)$ whereas in the density estimation problem $\hat{f}_{h;j}(X_j) - \hat{f}_h(X_j) = (n-1)^{-1}\hat{f}_h(X_j) - [(n-1)h]^{-1}k(0)$ and the second term which is $O(1/nh)$ is as large as the mean integrated squared error.

4 Simulations

A small simulation study was run, primarily to determine if $CV(h)$ or $LNO(h)$ was preferable in small samples. Unfortunately, the results suggest that neither criterion performs well for samples of sizes 50 through 1000, selecting bandwidths which are far too small.

Four target distributions were selected from the examples of normal mixtures given in Marron and Wand (1992). These were: standard normal distribution, skewed unimodal distribution #2, asymmetric bimodal distribution #8 and claw distribution #10. The cumulative distribution functions are listed in Table 1. For each distribution, 100 samples of sizes 50, 500 and 1000 were generated. For each sample, the minimizer of $CV(h)$, $LNO(h)$ and $ASE(h)$ was determined over a logarithmic grid of bandwidths from .004 to 3.73. The weight function used was $w(X_{(j)}) = 1$ for $j = 4,\ldots,n-3$ and $w(X_{(j)}) = 0$ for $j = 1,2,3,n-2,n-1,n$ where

Name	Distribution Function
Standard Normal Distribution	$N(0,1)$
Skewed Unimodal Distribution #2	$\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2}, \frac{4}{9}) + \frac{3}{5}N(\frac{13}{12}, \frac{25}{81})$
Skewed Bimodal Distribution #8	$\frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2}, \frac{1}{9})$
Claw Distribution #10	$\frac{1}{2}N(0,1) + \sum_{i=0}^{4} \frac{1}{10}N(\frac{i}{2} - 1, \frac{1}{100})$

Table 1: Distribution functions used in the simulation study. Plots of the corresponding densities are in Marron and Wand, 1992.

$X_{(j)}$ is the j^{th} order statistic. As well, the minimizer of mean squared error (MSE), i.e., the expectation of ASE, was estimated.

The results for the standard normal distribution are displayed in Figure 1 - results for the other 3 distributions are virtually identical. $CV(h)$ and $LNO(h)$ consistently choose the smallest available bandwidth. (Slight spread for sample size 50 is due to flatness of the criteria near the minimum.)

Equations (2.4) and (2.5) show that $CV(h)$ and $LNO(h)$ are identically 0 at $h = 0$. However, Figure 2 shows that choice of the smallest bandwidth available is *not* due to allowing the bandwidths to approach too close to zero, but is in fact inherent in the selection criteria. The Figure shows $CV(h)$, $LNO(h)$ and $ASE(h)$ for 5 samples of size n=1000 in the range $h = .004$ to $h = .25$. $CV(h)$ and $LNO(h)$ are clearly monotone increasing in this region, while $ASE(h)$ has a clear minimum. (All the functions increase in the range $h = .25$ to $h = 3.73$. The restricted range displayed on the plots was chosen to make the minimum apparent on the vertical scale.)

The simulation results clearly show that neither $CV(h)$ nor $LNO(h)$ are sufficiently good for bandwidth selection in the range of sample sizes observed in practice.

References

Altman, N. and Léger, C. (1993). Cross-validation, the bootstrap, and related methods for tuning parameter selection. Tech. Rep. BU-1216-M, Biometrics Unit, Cornell University.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** 712–736.

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions. *J. Statist. Plann. Inference* **35** 65–75.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147.

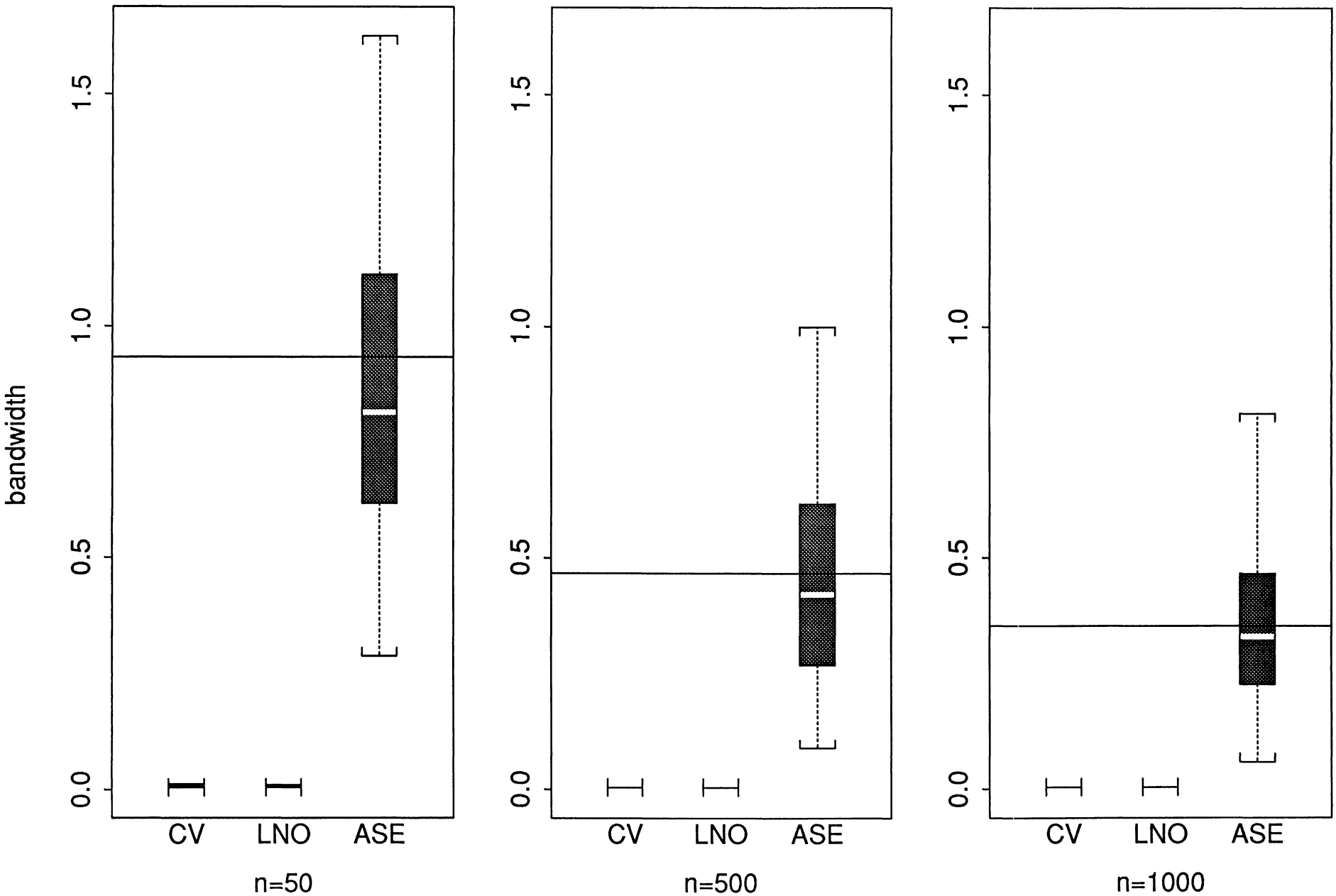

FIGURE 1: The bandwidths minimizing the CV, LNO, and ASE criteria for 100 samples of size n from a Normal distribution. The horizontal line is the bandwidth minimizing MSE.

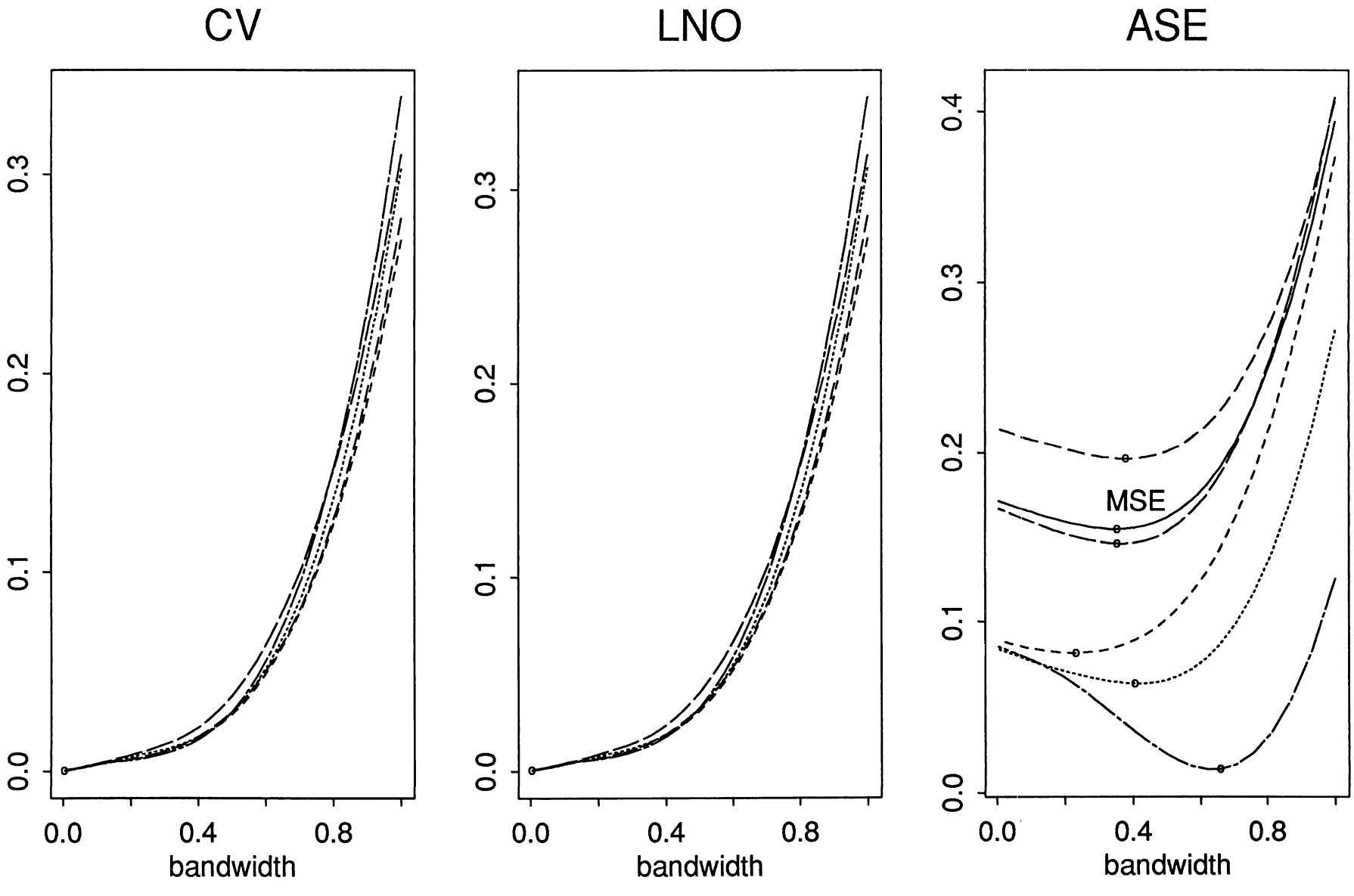

FIGURE 2: Bandwidth selection criteria (CV, LNO and ASE) as a function of bandwidth for 5 realizations. The minimum for each curve is marked with a small circle. Notice that the minimum of CV and LNO is at the smallest available bandwidth for each realization while the minimum of ASE varies, but is in the range 0.2 to 0.8. The solid curve on the ASE plot denotes the MSE curve, with minimum at the target bandwidth 0.354.