

Cross Validated SNP Density Estimates

Mark Coppejans
Duke University
Durham NC, USA

A. Ronald Gallant ¹
University of North Carolina
Chapel Hill NC, USA

First draft: May 1999

This draft: April 2000

¹Corresponding author: Mark Coppejans, Department of Economics, Duke University, Box 90097, Durham NC 27708-0097 USA; phone 1-919-660-1804; e-mail mtc@econ.duke.edu. Research supported by the National Science Foundation. The most recent version of the paper is available by anonymous ftp as the PostScript file an.ps at ftp.econ.duke.edu in directory pub/arg/papers.

ABSTRACT

We consider cross-validation strategies for the SNP nonparametric density estimator, which is a truncation (or sieve) estimator based upon a Hermite series expansion. Our main focus is on the use of SNP density estimators as an adjunct to EMM structural estimation. It is known that for this purpose a desirable truncation point occurs at the last point at which the MSE curve of the SNP density estimate declines abruptly. We study the determination of the MSE curve on a per sample basis for iid data by means of leave-one-out cross-validation and hold-out-sample cross-validation through an examination of their performance over the Marron-Wand test suite and models related to asset pricing and auction applications. We find that both methods are informative as to the location of abrupt drops. The hold-out-sample method is cheaper to compute because it requires fewer nonlinear optimizations. The minimum of the hold-out-sample cross-validation curve also seems to be a better indicator of the minimum of the true MSE curve. We consider the asymptotic justification of hold-out-sample cross-validation. For this purpose, we establish rates of convergence of the SNP estimator under the Hellinger norm that are of interest in their own right.

1 Introduction and Summary

We consider cross-validation strategies for the SNP nonparametric density estimator introduced by Gallant and Nychka (1987), which is a truncation (or sieve) estimator based upon a Hermite series expansion. SNP density estimators were introduced to model the semiparametric component of nonlinear structural models. Nonlinear models require a numerical optimization procedure to estimate the parameters of the parametric part of the model. A series expansion is the natural choice for modeling the nonparametric part because determining the coefficients of series expansion can be accomplished by just including the coefficients among the parameters to be determined in the numerical optimization. Moreover, if one ignores the sieve origin of the additional parameters and computes tests and confidence intervals as if they were part of the parametric model, inference is often reasonably accurate. Provably so in some instances (Eastwood and Gallant, 1991; Fan, Zhang, and Zhang, 2000).

An open question in these applications is how the truncation point should be determined in a given sample. Early semiparametric applications such as Davidian and Gallant (1992, 1993) and nonparametric applications such as Gallant, Hsieh, and Tauchen (1991), Gallant, Rossi, and Tauchen (1992, 1993) used likelihood based inference. It therefore seemed natural to use standard maximum likelihood truncation rules such as BIC (Schwarz, 1978) to determine the truncation point. This tendency has become habitual and BIC is now the standard choice criterion, regardless of application. We provide results that are relevant to nonparametric and semiparametric likelihood based applications in this paper.

However, our main interest is in the choice of the truncation point for SNP estimators used in connection with EMM estimation (Gallant and Tauchen, 1999). In this application it is known that if the truncation point of the SNP estimator is properly selected, then the EMM estimator is fully efficient (Gallant and Long, 1997; Chumacero, 1997; Andersen, Chung, and Sorensen, 1999; and Ng and Michaelides, 2000).

For iid data, more is known. It is known that BIC will most likely truncate too soon (Fenton and Gallant, 1996) and therefore BIC is not likely to suggest the correct truncation point in EMM applications. This same study suggests that more aggressive truncation strategies such as AIC are unlikely to work either. Further, it is known that in typical EMM

applications it is only necessary to make sure that the truncation point is large enough to achieve nearly full efficiency. Once this point is reached, further EMM efficiency gains are marginal, regardless of sample size (Gallant and Tauchen, 1999). Moreover, the mean square of the error in the SNP fit to the true density plotted against the truncation point seems to provide a reasonably reliable suggestion as to where this point might be. The MSE error curve has a few abrupt declines that correspond to when the estimator captures the gross features of the true density. The correct truncation point to achieve nearly full efficiency lies just to the right of an abrupt decline (Gallant and Tauchen, 1999).

In this paper we study the determination by cross-validation of the SNP mean squared error function on a per sample basis for iid data. We study two methods: the more traditional leave-one-out cross-validation method, and an average hold-out-sample method. We find that both methods are informative as to the location of abrupt drops. The hold-out-sample method is cheaper to compute, because it requires fewer nonlinear optimizations. It also seems to indicate the minimum of the true MSE curve more reliably and can be justified theoretically.

The plan of the paper is as follows. In Section 2, we describe the leave-one-out and hold-out-sample cross-validation procedures and examine their behavior over the densities of the Marron-Wand test suite and for two densities representative of asset pricing and auction applications, namely, the second largest order statistic from the log normal distribution and a scale mixture of normals. It is from this examination that we conclude that the estimated mean square error functions provides adequate guidance for applications. In Section 4 we study the Hellinger distance properties of the SNP estimator using methods devised by Wong and Shen (1995), which we believe to be the sharpest results available to date. In Section 5 we relate Hellinger distance to mean squared error, and derive results on for hold-out-sample cross-validation. Section 6 concludes.

2 Test Cases

The SNP estimator is based on the class of densities

$$\mathcal{F}_K = \left\{ f_K : f_K(x, \xi) = \left[\sum_{i=0}^K \xi_i x^i \right]^2 e^{-x^2/2} + \epsilon_o \phi(x), \xi \in \Xi_K \right\}$$

$$\Xi_K = \left\{ \xi : \xi = (\xi_0, \xi_1, \dots, \xi_K), \int f_K(x, \xi) dx = 1 \right\}$$

where ϕ denotes the standard normal density, ϵ_o is a small positive number, and $K = 0, 1, \dots$

Estimation is by quasi maximum likelihood

$$\hat{f}_K = \underset{\substack{f \in \mathcal{F}_K \\ -\infty < \mu < \infty \\ 0 < \sigma < \infty}}{\operatorname{argmax}} \sum_{x_i \in \mathcal{X}} \log \left[\frac{1}{\sigma} f \left(\frac{x_i - \mu}{\sigma} \right) \right]$$

where

$$\mathcal{X} = \{x_1, x_2, \dots, x_n\}$$

is a random sample of size n from the unknown true density f_o . In the later sections we shall suggest specific relationships between K and n of the form $K = K_n$ for some function on the integers K_n . When such a relationship is in force, we will use \mathcal{F}_n and \hat{f}_n to denote \mathcal{F}_K and \hat{f}_K , respectively.

Some structural aspects of the SNP estimator deserve comment. If $K = 0$ then \mathcal{F}_K contains only the normal density which implies that the normal density is the leading term of the SNP expansion. Other choices of the weight function $e^{-x^2/2}$ that multiplies the polynomial term and the density $\phi(x)$ that multiplies ϵ_o are permitted by the theory developed in Gallant and Nychka (1987) and will result in a different leading term. Here, we shall only consider the choices $e^{-x^2/2}$ and $\phi(x)$ because they are particularly convenient and are standard in applications. The term $\epsilon_o \phi(x)$ acts as a lower bound that insures that integrals such as $\int_{-\infty}^{\infty} \log f f_o dx$ exist for all $f \in \mathcal{F}_K$. It also serves to keep terms such as $\log f$ from going out of range during optimizations. In the results reported here, $\epsilon_o = 10^{-5}$. Other nonparametric estimators based on Hermite expansions have been proposed (Devroye and Györfi, 1985). The SNP estimator differs from these others in three respects: it is positive, it is location and scale invariant, and the coefficients are determined by quasi maximum likelihood rather than by method of moments.

The mean squared error of the estimator is

$$\begin{aligned} \operatorname{MSE}(\hat{f}_K) &= \int \hat{f}_K^2 dx - 2 \int \hat{f}_K f_o dx + \int f_o^2 dx \\ &= M_{(1)} - 2M_{(2)} + M_{(3)} \end{aligned}$$

The first term on the right can be computed directly and the third term is constant and therefore not needed to determine the shape of a plot of $\text{MSE}(\hat{f}_K)$ against K . This leaves only

$$M_{(2)} = \int \hat{f}_K f_o dx$$

to be determined.

To this end, for a given proportion α , let $\mathcal{X}_{j,\alpha}$ denote the j th block of size $[\alpha n]$ from the sample \mathcal{X} , where $[\alpha n]$ is either the integer part of αn or 1, whichever is the larger; leftover points go in the last block. That is,

$$\begin{aligned} \mathcal{X}_{j,\alpha} &= \{x_{(j-1)[\alpha n]+1}, \dots, x_{j[\alpha n]}\} & j = 1, \dots, J-1 \\ \mathcal{X}_{J,\alpha} &= \{x_{(J-1)[\alpha n]+1}, \dots, x_n\}, \end{aligned}$$

where J is the largest positive integer with $J[\alpha n] \leq n$. Let $\hat{f}_{j,K}$ denote the estimate obtained from the sample points that remain after deletion of the j th block. That is,

$$\hat{f}_{j,K} = \underset{\substack{f \in \mathcal{F}_K \\ -\infty < \mu < \infty \\ 0 < \sigma < \infty}}{\text{argmax}} \sum_{x_i \in \tilde{\mathcal{X}}_{j,\alpha}} \log \left[\frac{1}{\sigma} f \left(\frac{x_i - \mu}{\sigma} \right) \right],$$

where $\tilde{\mathcal{X}}_{j,\alpha} = \mathcal{X} \setminus \mathcal{X}_{j,\alpha}$. An estimate of $M_{(2)}$ is

$$\hat{M}_{(2)} = \frac{1}{J} \sum_{j=1}^J \left[\frac{1}{[\alpha n]} \sum_{x_i \in \mathcal{X}_{j,\alpha}} \hat{f}_{j,K}(x_i) \right] \doteq \frac{1}{n} \sum_{j=1}^J \sum_{x_i \in \mathcal{X}_{j,\alpha}} \hat{f}_{j,K}(x_i).$$

If $\alpha = 1/n$, then

$$\text{CVL} = M_{(1)} - 2\hat{M}_{(2)}$$

is the traditional leave-one-out cross-validation formula, which requires n nonlinear optimizations to compute. See Marron (1987) for a discussion of leave-one-out cross-validation and its relation to other bandwidth strategies for delta sequence density estimators which include kernel estimators, histogram estimators, and orthogonal series estimators (with coefficients determined differently than here, as mentioned above).

The hold-out-sample cross-validation formula is obtained by putting α to some fixed proportion such as $\alpha = 0.1$, which is the value used in the figures that follow. This α does

not change as n increases, which is the basis for a distinction between hold-out-sample and leave-one-out cross-validation. Rather than the formula above, we find it convenient to use

$$\text{CVH} = \hat{M}_{(1)} - 2\hat{M}_{(2)}$$

for hold-out-sample cross-validation, where

$$\hat{M}_{(1)} = \frac{1}{J} \sum_{j=1}^J \int \hat{f}_{j,K}^2(x) dx.$$

This formula requires J nonlinear optimizations as opposed to the $J + 1$ optimizations that would be required were $M_{(1)}$ to replace $\hat{M}_{(1)}$.

In the work reported in this section, we know f_o and therefore can compute both $\text{MSE}(\hat{f}_K)$ and $M_{(3)}$. To facilitate comparisons, we actually compute and plot $\text{CVL} + M_{(3)}$ and $\text{CVH} + M_{(3)}$ against K . Of course this cannot be done in applications, nor is it necessary because $M_{(3)}$ does not depend on K .

Here we report computations for five densities: a stochastic volatility model, the second largest order statistic from the log normal density, and three densities from the Marron-Wand (1992) test suite. Computations for the remainder of the Marron-Wand test suite are available by anonymous ftp from host ftp.econ.duke.edu in directory pub/arg/papers as file mwsuite.ps or by clicking on “browse ftp site” at www.unc.edu/~arg. These five densities are plotted in Figure 1.

Figure 1 about here

The stochastic volatility model is one of the standard models for statistical analysis of financial market data. A recent summary of the literature is Ghysels, Harvey, and Renault (1995). Although specialized estimators for various specifications of the stochastic volatility model have been proposed, the EMM estimator is often used instead because it permits one to change specifications with little effort and provides informative diagnostic tests of tentative specifications. A recent example is Liu (2000). As discussed above, EMM is usually implemented with an SNP score which requires determination of K . Also, as discussed above,

the computations reported in Gallant and Tauchen (1999) suggest (i) that a typical MSE curve in applications will drop sharply and then flatten, (ii) that a reasonable choice of K is the last point at which MSE takes a sharp drop, and (iii) that finding the K that produces the best MSE will have little payoff, even if successful.

A standard stochastic volatility model, following Tauchen and Pitts (1983), is a lognormal scale mixture of normals. Its density is

$$p(y|\rho) = \int_{-\infty}^{\infty} n(y|\rho_1, e^{2u}) n(u|\rho_2, \rho_3^2) du$$

where $n(u|\alpha, \beta^2)$ denotes the normal density with mean α and variance β^2 . The mean μ , variance σ^2 , and raw kurtosis $\kappa = \frac{\mathcal{E}_\rho(y-\mu)^4}{(\sigma^2)^2}$ of $p(y|\rho)$ determine ρ as follows

$$\begin{aligned} \rho_1 &= \mu \\ \rho_2 &= \log \sigma - \rho_3^2 \\ \rho_3 &= [\log(\kappa/3)/4]^{1/2}. \end{aligned}$$

Our choice is $(\mu, \sigma^2, \kappa) = (0, 1/4, 8)$; the considerations in Section 4 of Gallant and Tauchen (1999) suggest that this choice is representative and that results are likely to be robust to changes in these choices. The density is plotted in the first panel of Figure 1.

For this choice, MSE, CVL, and CVH are plotted against K in Figure 2 for sample sizes $n = 400, 900, 1600, 2500 = 20^2, 30^2, 40^2, 50^2$. The plots of either CVL or CVH appear to be adequate for a determination of the point at which MSE drops sharply and begins to flatten. This finding holds for our remaining four cases, as we shall see, and for those cases not reported here, but available by ftp from the site indicated earlier.

Figure 2 about here

Some interesting subsidiary information is displayed in Figure 2. The K chosen by BIC, by minimum CVL, and by minimum CVH are shown as vertical bars. Interestingly, BIC seems to give a reliable indication of a lower bound for K : one ought not choose K to the left of that indicated by BIC. This finding is robust across cases considered. The minimum

values of the CVL and CVH curves do not give a reliable indication of the minimum of the true MSE curve for small sample sizes, with CVH seemingly the more reliable of the two. However, as noted earlier, it is more important to locate the points at which the MSE curve drops abruptly than to locate its exact minimum.

As suggested by the theory that follows, because SNP is a truncation estimator and can therefore automatically adapt to the smoothness of f_o , SNP ought to be able to achieve better MSE than kernel estimators, which cannot automatically adapt. In each plot shown in Figure 2, the upper horizontal dotted line shows the MSE for the leave-one-out cross-validated kernel bandwidth and the lower dotted line shows the best MSE achievable by kernel estimators. In most instances, these two horizontal lines are too close together to be distinguished visually. Inspection of the plots bears out the conjecture because the solid MSE line drops below the horizontal dotted lines. Most often, the difference is slight. It could be made more apparent by left-truncating the plots and changing scale, but it is nonetheless true that the difference is probably too slight to be of practical importance. This finding is also robust across cases considered.

An area of current interest in simulation estimation is the analysis of auction data; see for instance, Laffont, Ossard, and Vuong (1995). Under an independent private value assumption, the values assigned to an object by N bidders are a random sample of size N from a common valuation distribution. In either a single object second price auction or a single object English (oral ascending bid) auction, the observed selling price is the second largest order statistic in this sample. In an sealed bid first price auction, and other auction formats, the winning bid is a functional of the order statistics of the valuation distribution. These observations suggest that estimating the density of the second largest order statistic in a sample of size N has relevance to the application of EMM estimators to auction data. Here we consider the case when the common valuation distribution is lognormal.

The second largest order statistic from a sample of size $\rho_1 = N$ from the lognormal distribution has density function

$$p(y|\rho) = \frac{N(N-1)}{y} \left[\Phi\left(\frac{\log y - \rho_2}{\rho_3}\right) \right]^{N-2} \left[1 - \Phi\left(\frac{\log y - \rho_2}{\rho_3}\right) \right] \phi\left(\frac{\log y - \rho_2}{\rho_3}\right)$$

where $y > 0$ and ϕ and Φ denote the standard normal density and distribution functions,

respectively.

The lognormal distribution is nearly normal for small values of ρ_3 and and departs from the normal as ρ_3 increases (Johnson and Kotz, 1994, Chapter 14). The choice $\rho_3 = 1$ appears reasonable judging from the tables in Johnson and Kotz (1994, Chapter 14). The considerations in Section 5 of Gallant and Tauchen (1999) suggest that $\rho_1 = N$ and $\rho_2 = -3$ are both representative and that results are likely to be robust to changes in these choices. The density is plotted in the second panel of Figure 1.

Figure 3 about here

MSE, CVL, and CVH are plotted in Figure 3. As above, the plots of either CVL or CVH appear to be adequate for a determination of the point at which MSE drops sharply and begins to flatten and BIC gives a reliable indication of a lower bound for K .

Marron and Wand (1992) proposed a test suite as a battery for use in evaluating density estimators. It is designed to evaluate the ability of an estimator to track complex behavior at the center of the distribution. The densities in the suite are mixtures of normals; the mixing proportions $\{\pi_j\}_{j=1}^J$ and the corresponding means $\{\mu_j\}_{j=1}^J$ and standard deviations $\{\sigma_j\}_{j=1}^J$ of the constituent normal densities are given in Marron and Wand (1992). Here, we consider three densities from the suite: the trimodal, the Gaussian, and the smooth comb. They are plotted in the panels three through five of Figure 1. The trimodal is a reasonably representative density from the test suite and the Gaussian and smooth comb are at opposite extremes. The Gaussian is actually in \mathcal{F}_K and the smooth comb is a very difficult density to estimate.

Figure 4 about here

For the trimodal density, MSE, CVL, and CVH are plotted in Figure 4. As before, the plots of either CVL or CVH appear to be adequate for a determination of the point at which

MSE drops sharply and begins to flatten and BIC gives a reliable indication of a lower bound for K .

Figure 5 about here

Figure 5, which shows overplots of \hat{f}_K and f_o for the case $n=900$, indicates the source of the sharp declines in MSE, CVL, and CVH seen in Figure 4. As seen from Figure 5, the abrupt declines are the points at which the gross features of f_o are first captured by \hat{f}_K ; specifically for the trimodal, they are the points at which modes are detected. Once the gross features are identified, relatively extreme values of K are required to determine fine detail. The computations reported in Gallant and Tauchen (1999) suggest that there is little payoff in terms of the efficiency of the EMM estimator to capturing fine detail. Determination of the gross features seems to be adequate.

Figure 6 about here

Figure 6 plots MSE, CVL, and CVH for the Gaussian density, which, as remarked above, is actually in \mathcal{F}_K . As one would expect, all methods for determining K eventually detect this fact. Interestingly, minimum CVH does better than minimum CVL.

Figure 7 about here

Lastly, Figure 7 plots MSE, CVL, and CVH for the smooth comb density. This is a density that is very hard to estimate. Correspondingly, the MSE shows less abrupt drops and all three methods for determining K shown in the figure get driven inexorably higher as sample size increases. Nonetheless, the plots CVL and CVH do seem to give a reliable indication of the situation one faces.

The numerical methods employed are as follows. All code is in C++. The major software components are available by anonymous ftp from host ftp.econ.duke.edu in directories pub/arg/npe and pub/arg/libcpp or by clicking on “browse ftp site” at www.unc.edu/~arg. The algorithms used to compute \hat{f}_K and the methods employed to compute and simulate from the Marron-Wand test suite are described in Section 4.3 of Fenton and Gallant (1996); Sections 4 and 5 of Gallant and Tauchen (1999) describe computation of and simulations from the stochastic volatility model and the second largest order statistic from the log normal. Briefly, f_K is represented as a sum of Hermite functions, as described in Section 3, and optimization is carried out by a quasi Newton method with line search using the outer product form of the information matrix to approximate the Hessian; the requisite matrix algebra is by means of a matrix class found in pub/arg/libcpp at ftp.econ.duke.edu. The integrals $\text{MSE}(\hat{f}_K)$ and $M_{(3)}$ are computed as Reimann sums, e.g., $M_{(3)} = \sum f_o^2(x_i)\Delta x_i$ for finely spaced x_i , and $M_{(1)}$ is computed by Gauss-Hermite quadrature.

3 Hermite Expansions

It is convenient to rewrite the SNP density in terms of normalized Hermite polynomials whose definitions and properties are collected together in the following lemma.

LEMMA 1 The Hermite polynomials

$$H_{e_i}(x) = \sum_{m=0}^{\lceil i/2 \rceil} (-1)^m \frac{i!}{m!2^m(i-2m)!} x^{i-2m},$$

where $\lceil i/2 \rceil$ is the integer part of $i/2$, and their normalized counterpart

$$\bar{H}_{e_i}(x) = (\sqrt{2\pi} i!)^{-1/2} H_{e_i}(x)$$

satisfy the following properties for all real x and all positive integers m and i .

- (a) $\int_{-\infty}^{\infty} H_{e_i}(x)H_{e_m}(x)e^{-x^2/2} dx = \begin{cases} \sqrt{2\pi}i! & i = m \\ 0 & i \neq m \end{cases}$
- (b) $\int_{-\infty}^{\infty} \bar{H}_{e_i}(x)\bar{H}_{e_m}(x)e^{-x^2/2} dx = \begin{cases} 1 & i = m \\ 0 & i \neq m \end{cases}$
- (c) Differential equation: $\frac{d^2}{dx^2}H_{e_i}(x) - x\frac{d}{dx}H_{e_i}(x) + iH_{e_i}(x) = 0$

- (d) Recurrence relation: $H_{e_{i+1}}(x) = xH_{e_i}(x) - iH_{e_{i-1}}(x)$
- (e) Differential relation: $\frac{d}{dx}H_{e_i}(x) = iH_{e_{i-1}}(x)$
- (f) Rodrigues formula: $H_{e_i}(x) = (-1)^i e^{x^2/2} \frac{d^i}{dx^i} e^{-x^2/2}$
- (g) Upper bound: $|\overline{H}_{e_i}(x)e^{-x^2/4}| < \tilde{B} \doteq 0.6862127$.

Proof Abramowitz and Stegun (1972, Chapter 22). □

In this section and hereafter, K will depend explicitly on the sample size n according to $K = K_n$, where K_n is a function defined on the integers that is specified later. We write \mathcal{F}_n for \mathcal{F}_K and f_n for f_K to emphasize that this restriction is in force. Using Lemma 1, the class \mathcal{F}_n can be written in terms of normalized Hermite polynomials as

$$\mathcal{F}_n = \left\{ f_n : f_n(x, \theta) = \left[\sum_{i=0}^{K_n} \theta_i \overline{H}_{e_i}(x) \right]^2 e^{-x^2/2} + \epsilon_o \phi(x), \theta \in \Theta_n \right\}$$

$$\Theta_n = \left\{ \theta : \theta = (\theta_0, \theta_1, \dots, \theta_{K_n}), \sum_{i=0}^{K_n} \theta_i^2 + \epsilon_o = 1 \right\}.$$

The family of orthonormal functions $\left\{ \overline{H}_{e_m}(x) \right\}_{m=0}^{\infty}$ is complete for $L_2(e^{-x^2/2})$ where $L_2(e^{-x^2/2})$ denotes those functions g with

$$\|g\|_H = \left(\int_{-\infty}^{\infty} g^2(x) e^{-x^2/2} dx \right)^{1/2} < \infty$$

(Sansone, 1991, p. 351). Thus every g in $L_2(e^{-x^2/2})$ has the expansion

$$g(x) = \sum_{i=0}^{\infty} \theta_i \overline{H}_{e_i}(x)$$

where $\theta_i = \int_{-\infty}^{\infty} g(x) \overline{H}_{e_i}(x) e^{-x^2/2} dx$ and equality is with respect to $\|\cdot\|_H$. Moreover $\int_{-\infty}^{\infty} g^2(x) e^{-x^2/2} dx = \sum_{i=0}^{\infty} \theta_i^2$. Suppose f is a density that can be written as

$$f(x) = g^2(x) e^{-x^2/2} + \epsilon_o \phi(x)$$

for some g . Because $\int f dx = \int \phi dx = 1$, it follows that g is in $L_2(e^{-x^2/2})$ and that g has a Hermite expansion whose coefficients satisfy

$$\sum_{i=0}^{\infty} \theta_i^2 + \epsilon_o = 1$$

Consequently, the class of functions that can be attained as limits of functions from \mathcal{F}_n is

$$\mathcal{F}_\infty = \{f : f = g^2 e^{-x^2/2} + \epsilon_o \phi, g \in L_2(e^{-x^2/2}), \int f dx = 1\} = \overline{\left(\bigcup_{n=1}^{\infty} \mathcal{F}_n \right)}$$

where the overbar indicates closure with respect to the norm $\|f\|_{\mathcal{F}} = \|g\|_H$ defined for functions of the form $f = g^2 e^{-x^2/2} + \epsilon_o \phi$ with $g \in L_2(e^{-x^2/2})$.

The following lemma states the rate at which the Hermite coefficients decrease.

LEMMA 2 If g is k -times differentiable with $g^{(j)} \in L_2(e^{-x^2/2})$ for $j = 0, 1, \dots, k$ then the Hermite coefficients of g satisfy

$$\sum_{i=K}^{\infty} \theta_i^2 = o(K^{-k}).$$

Proof If g is k -times differentiable with $g^{(k)} \in L_2(e^{-x^2/2})$ then $g^{(k)}$ has a Hermite expansion with coefficients

$$\theta_i^{(k)} = \int_{-\infty}^{\infty} g^{(k)}(x) \overline{H}_{e_i}(x) e^{-x^2/2} dx = (\sqrt{2\pi} i!)^{-1/2} \int_{-\infty}^{\infty} g^{(k)}(x) H_{e_i}(x) e^{-x^2/2} dx.$$

Using Parts (d) and (e) of Lemma 1 to integrate by parts we have the recursion

$$\int_{-\infty}^{\infty} g^{(k)}(x) H_{e_i}(x) e^{-x^2/2} dx = \int_{-\infty}^{\infty} g^{(k-1)}(x) H_{e_{i+1}}(x) e^{-x^2/2} dx$$

which can be iterated to yield

$$\theta_i^{(k)} = [(i+k) \cdots (i+1)]^{1/2} \int_{-\infty}^{\infty} g(x) \overline{H}_{e_{i+k}}(x) e^{-x^2/2} dx = [(i+k) \cdots (i+1)]^{1/2} \theta_{i+k}.$$

It follows that

$$0 \leq \sum_{i=0}^{\infty} \theta_{i+k}^2 < \sum_{i=0}^{\infty} i \theta_{i+k}^2 < \sum_{i=0}^{\infty} i^2 \theta_{i+k}^2 < \cdots < \sum_{i=0}^{\infty} i^k \theta_{i+k}^2 < \sum_{i=0}^{\infty} [\theta_i^{(k)}]^2 < \infty$$

which implies $\sum_{i=0}^{\infty} (i+k)^k \theta_{i+k}^2 < \infty$ because $(i+k)^k$ is a polynomial in i of degree k . From

$$0 \leq \lim_{K \rightarrow \infty} \sum_{i=K}^{\infty} K^k \theta_i^2 \leq \lim_{K \rightarrow \infty} \sum_{i=K}^{\infty} i^k \theta_i^2 = 0$$

we can conclude that

$$\sum_{i=K}^{\infty} \theta_i^2 = o(K^{-k}).$$

□

4 A Rate on Hellinger Distance

In this section we apply a result of Wong and Shen (1995) to obtain a rate of convergence on the Hellinger distance between the true density and its estimate. This rate is used in the next section to establish the asymptotic validity of cross-validation formulae.

Write the true density

$$f_o(x) = [g_o(x)]^2 e^{-x^2/2} + \epsilon_o \phi(x)$$

as

$$f_\infty(x, \theta_o) = \left[\sum_{i=0}^{\infty} \theta_{oi} \overline{H}_{e_i}(x) \right]^2 e^{-x^2/2} + \epsilon_o \phi(x)$$

where $\theta_o = (\theta_{o0}, \theta_{o1}, \theta_{o2}, \dots)$. When θ_o is truncated for use in the formula

$$f_n(x, \theta) = \left[\sum_{i=0}^{K_n} \theta_i \overline{H}_{e_i}(x) \right]^2 e^{-x^2/2} + \epsilon_o \phi(x),$$

the renormalization

$$\theta_n = \frac{\sqrt{(1 - \epsilon_o)}}{\sqrt{\sum_{i=0}^{K_n} (\theta_{oi})^2}} (\theta_{o0}, \theta_{o1}, \dots, \theta_{oK_n})$$

is required to get θ_n in Θ_n . Thus, there is a distinction between $f_n(x, \theta_o)$ and $f_n(x, \theta_n)$.

To apply the Wong and Shen result, we need to compute the bracketing Hellinger metric entropy of the estimator and the chi-squared truncation error, which we do next, under the following assumption.

ASSUMPTION 1 Let f_o indicate the true density, the density from which the observed data $\{x_t\}$ is a random sample. We shall require f_o to be in \mathcal{F}_∞ , to at be least once continuously differntiable over $(-\infty, \infty)$, and to have a largest mode. Writing $f_K(x) = g_K^2(x)e^{-x^2/2} + \epsilon_o \phi(x)$ for the truncation of $f_o(x) = g_o^2(x)e^{-x^2/2} + \epsilon_o \phi(x)$, we require that

$$\lim_{K \rightarrow \infty} \sup_{-\infty < x < \infty} \frac{g_o^2(x) + g_K^2(x)}{g_K^2(x) + \epsilon_o / \sqrt{2\pi}} \leq C$$

for some $C < \infty$.

The SNP estimator is location and scale invariant. We will exploit this fact in the theoretical development by assuming that the mode given by Assumption 1 occurs at $x = 0$ with $f(0) = 1$ and considering the simplified estimator

$$\hat{f}_n = f_n(x, \hat{\theta}_n)$$

where

$$\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta_n} \frac{1}{n} \sum_{t=1}^n \log[f_n(x_t, \theta)],$$

K_n increases with n at a specified rate, and $\{x_t\}$ is a random sample from the density $f_o(x)$.

The condition $\lim_{K \rightarrow \infty} \sup_{-\infty < x < \infty} [g_o^2(x) + g_K^2(x)] / [g_K^2(x) + \epsilon_o / \sqrt{2\pi}] \leq C$ restricts the tail behavior of f_o . If f_o has normal tails beyond some points well to the left and right of the 1% and 99% quantiles of f_o , the condition is satisfied. If g_o is any polynomial of finite degree plus a bounded function, the condition is satisfied. Restrictions on tail behavior are common in nonparametric density estimation, and this condition seems relatively mild as such restrictions go. Note that Assumption 1 is a restriction on f_o only, we have not restricted \mathcal{F}_n .

The Hellinger metric $d(\cdot, \cdot)$ is defined by

$$d(f_1, f_2) = \left\{ \int [f_1^{1/2}(x) - f_2^{1/2}(x)]^2 dx \right\}^{1/2} = \|f_1^{1/2} - f_2^{1/2}\|.$$

A Hellinger u -bracketing of \mathcal{F}_n is a set of N pairs of functions $\{(f_j^L, f_j^U) : j = 1, 2, \dots, N\}$ from \mathcal{F}_n such that (i) $d(f_j^L, f_j^U) \leq u$ for $j = 1, 2, \dots, N$ and (ii) for any $h \in \mathcal{F}_n$, there is a j such that $f_j^L \leq h \leq f_j^U$. Let N^* denote the the smallest such N . The Hellinger metric entropy of \mathcal{F}_n is the function $H(\cdot, \mathcal{F}_n)$ defined by

$$H(u, \mathcal{F}_n) = \log N^*.$$

If f_1 are f_2 both in \mathcal{F}_n , then

$$\begin{aligned} d^2(f_1, f_2) &= \int (f_1^{1/2} - f_2^{1/2})^2 dx \\ &= \int \frac{(f_1 - f_2)^2}{(f_1^{1/2} + f_2^{1/2})^2} dx \\ &= \int \frac{[\sum_{i=0}^{K_n} (\theta_{1i} - \theta_{2i}) \bar{H}_{e_i}(x) e^{-x^2/2}]^2 [(f_1 - \epsilon_0 \phi)^{1/2} + (f_2 - \epsilon_0 \phi)^{1/2}]^2}{(f_1^{1/2} + f_2^{1/2})^2} dx \\ &\leq \int \left[\sum_{i=0}^{K_n} (\theta_{1i} - \theta_{2i}) \bar{H}_{e_i}(x) e^{-x^2/2} \right]^2 dx \\ &= \sum_{i=0}^{K_n} (\theta_{1i} - \theta_{2i})^2. \end{aligned}$$

Therefore, $\left[\sum_{i=0}^{K_n} (\theta_{1i} - \theta_{2i})^2\right]^{1/2} \leq u$ implies $d(f_1, f_2) \leq u$. By Gallant and Souza (1991, Lemma 1), the number of open balls of radius u required to cover Θ_n is bounded by $2(K_n + 1)(2/u + 1)^{K_n}$. Therefore, $N^* \leq 2(K_n + 1)(1/u + 1)^{K_n}$ and

$$H(u, \mathcal{F}_n) \leq \log(2K_n + 2) + K_n \log(1/u + 1).$$

To apply the Wong and Shen result, we are required to find ϵ that satisfies

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} H^{1/2}(u, \mathcal{F}_n) du \leq c_4 n^{1/2} \epsilon^2$$

for a given c_4 . Because both the square root and log functions are concave, Jensen's inequality implies, for $a = \sqrt{2}\epsilon - \epsilon^2/2^8$, that

$$\begin{aligned} \int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} H^{1/2}(u, \mathcal{F}_n) du &\leq a \left[\log(2K_n + 2) + K_n \log \left(a^{-1} \int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \frac{1}{u} du + 1 \right) \right]^{1/2} \\ &= a \left[\log(2K_n + 2) + K_n \log \left(a^{-1} \log \frac{\sqrt{2}\epsilon}{\epsilon^2/2^8} + 1 \right) \right]^{1/2}. \end{aligned}$$

Suppose that $K_n = n^\alpha$ and $\epsilon = n^{-\beta}$ for some $\alpha, \beta > 0$. Then from the expression above it follows that given any $\delta > 0$ there are bounds B_i such that

$$\begin{aligned} \int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} H^{1/2}(u, \mathcal{F}_n) du &\leq B_1 n^{-\beta} \left[\alpha \log n + n^\alpha \log \left(n^\beta \log n^{-\beta} \right) \right]^{1/2} \\ &\leq B_2 n^{-\beta} \left(n^\alpha \log n^\beta \right)^{1/2} \\ &\leq B_3 n^{-\beta + \alpha/2 + \delta/2} \end{aligned}$$

Because δ is arbitrary, we may assume that $B_3 \leq c_4$. If we put $\beta = (1 - \alpha - \delta)/2$ so that

$$\epsilon = n^{-\beta} = n^{(\alpha + \delta - 1)/2},$$

then this inequality implies

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} H^{1/2}(u, \mathcal{F}_n) du \leq c_4 n^{1/2} \epsilon^2$$

as required.

Next we compute the rate of decrease of the chi-squared truncation error defined as

$$\delta_n(1) = \inf_{f \in \mathcal{F}_n} \rho_1(f_o, f).$$

Using Assumption 1 and writing $f_n = f(\cdot, \theta_n)$ for the truncation of f_o normalized to integrate to one, we have for some $B > 0$ that

$$\begin{aligned}
\delta_n(1) &\leq \int \frac{(f_o - f_n)^2}{f_n} dx \\
&= \int \frac{(g_o^2 - g_n^2)^2 e^{-x^2/2}}{g_n^2 + \epsilon_o/\sqrt{2\pi}} dx \\
&= \int \frac{(g_o - g_n)^2 (g_o + g_n)^2 e^{-x^2/2}}{g_n^2 + \epsilon_o/\sqrt{2\pi}} dx \\
&\leq \int \frac{2(g_o^2 + g_n^2)}{g_n^2 + \epsilon_o/\sqrt{2\pi}} (g_o - g_n)^2 e^{-x^2/2} dx \\
&\leq B \int (g_o - g_n)^2 e^{-x^2/2} dx \quad \text{for large } K \\
&= B \int \left[\sum_{i=0}^{\infty} \theta_{oi} \bar{H}_{e_i} - \sum_{i=0}^{K_n} \theta_i \bar{H}_{e_i} \right]^2 e^{-x^2/2} dx \\
&= B \int \left[\sum_{i=K_n+1}^{\infty} \theta_{oi} \bar{H}_{e_i} + B \sum_{i=0}^{K_n} (\theta_i - \theta_{oi}) \bar{H}_{e_i} \right]^2 e^{-x^2/2} dx \\
&\leq 2B \int \left\{ \left[\sum_{i=K_n+1}^{\infty} \theta_{oi} \bar{H}_{e_i} \right]^2 + B \left[\sum_{i=0}^{K_n} (\theta_i - \theta_{oi}) \bar{H}_{e_i} \right]^2 \right\} e^{-x^2/2} dx \\
&= 2B \sum_{i=K_n+1}^{\infty} (\theta_{oi})^2 + 2B \sum_{i=0}^{K_n} (\theta_i - \theta_{oi})^2 \\
&= 2B \sum_{i=K_n+1}^{\infty} (\theta_{oi})^2 + 2B \frac{[\sqrt{\sum_{i=1}^{\infty} (\theta_{oi})^2} - \sqrt{\sum_{i=1}^{K_n} (\theta_{oi})^2}]^2}{\sum_{i=0}^{K_n} (\theta_{oi})^2} \sum_{i=0}^{K_n} (\theta_{oi})^2 \\
&= 2B \sum_{i=K_n+1}^{\infty} (\theta_{oi})^2 + 2B \left[\sum_{i=1}^{\infty} (\theta_{oi})^2 + \sum_{i=1}^{K_n} (\theta_{oi})^2 - 2\sqrt{\sum_{i=1}^{\infty} (\theta_{oi})^2} \sqrt{\sum_{i=1}^{K_n} (\theta_{oi})^2} \right] \\
&\leq 2B \sum_{i=K_n+1}^{\infty} (\theta_{oi})^2 + 2B \left[\sum_{i=1}^{\infty} (\theta_{oi})^2 - \sum_{i=1}^{K_n} (\theta_{oi})^2 \right] \\
&= 4B \sum_{i=K_n+1}^{\infty} (\theta_{oi})^2 \\
&= o(K_n^{-k}).
\end{aligned}$$

Therefore, the rate of decrease of the chi-squared truncation error is $\delta_n(1) = o(K_n^{-k})$.

For every $\theta \in \Theta_n$ and $c_1 > 0$,

$$\left\{ \|f_n^{1/2}(\cdot, \hat{\theta}_n) - f_o^{1/2}\| \geq \epsilon \right\} \subset A(n, \theta)$$

where

$$A(n, \theta) = \left\{ \sup_{\|f^{1/2} - f_o^{1/2}\| \geq \epsilon, f \in \mathcal{F}_n} \prod_{i=1}^n \frac{f(x_i)}{f_n(x_i, \theta)} \geq \exp\left(-\frac{1}{2}c_1 n \epsilon^2\right) \right\}.$$

Using Theorem 3(i) of Wong and Shen, we have

$$P \left\{ \|\hat{f}_n^{1/2} - f_o^{1/2}\| \geq \epsilon \right\} \leq \inf_{\theta \in \Theta_n} P[A(n, \theta)] \leq 5 \exp(-c_2 n \epsilon^2) + \exp \left[-n c_1 \epsilon^2 / 2 + n \delta_n(1) \right].$$

Substituting $\epsilon = n^{(\alpha+\delta-1)/2}$ and $\delta_n(1) = o(n^{-\alpha k})$ from above we have

$$P \left\{ \|\hat{f}_n^{1/2} - f_o^{1/2}\| \geq n^{(\alpha+\delta-1)/2} \right\} \leq 5 \exp(-c_2 n^{\alpha+\delta}) + \exp \left[-c_1 n^{\alpha+\delta} / 2 + o(n^{-\alpha k+1}) \right].$$

The best rate is achieved when $n^{\alpha+\delta} \approx n^{-\alpha k+1}$; that is, when $\alpha = \frac{1-\delta}{k+1}$. This gives,

$$P \left\{ \|\hat{f}_n^{1/2} - f_o^{1/2}\| \geq n^{-k(1-\delta)/(2k+2)} \right\} \leq 5 \exp(-c_2 n^{(1+\delta k)/(k+1)}) + \exp \left\{ -n^{(1+\delta k)/(k+1)} [(c_1/2) - o(1)] \right\}.$$

Application of the Borel-Cantelli lemma yields

$$\|\hat{f}_n^{1/2} - f_o^{1/2}\| = n^{-k(1-\delta)/(2k+2)}$$

almost surely where $\delta > 0$ is arbitrary.

If we were to further restrict f_o so that $\sum_{i=K}^{\infty} \theta_i^2 \approx K^{-k-\xi}$ for some small $\xi > 0$, rather than using $\sum_{i=K}^{\infty} \theta_i^2 = o(K^{-k})$ given by Lemma 2, then we would have $\delta_n(1) \approx K^{-k-\xi}$ and could set $\delta = 0$ in the expression for $\|\hat{f}_n^{1/2} - f_o^{1/2}\|$ above.

5 Cross-Validation

As seen earlier, for the applications we have in mind, we require an estimate of the mean square error curve that is uniform in the truncation point of the SNP estimator. In this section, we shall derive this result by pursuing the more traditional goal of trying to find K_n that minimizes the mean square error on a per sample basis and obtain a uniform estimate of the mean square error curve as a by-product.

By per sample basis, we mean that for each j and each n , $\hat{f}_{j,n}$ is held fixed and the only sampling variation that is taken into account is the sampling variation over $\mathcal{X}_{j,\alpha}$. That is, for each j , all expectations and probability statements are conditional on $\tilde{\mathcal{X}}_{j,\alpha}$ and the structure with respect to n is that of a triangular array of random variables with rows

$$\mathcal{X}_{j,\alpha} = \{x_{(j-1)[\alpha n]+1}, \dots, x_{j[\alpha n]}\}.$$

Location and scale parameters of certain random variables defined over a row $\mathcal{X}_{j,\alpha}$ will depend on $\hat{f}_{j,n}$. To avoid clutter, we do not use a special notation to indicate this structure.

We must first show that a sequence K_n that drives the mean square error to zero exists. This may be done by relating the L_2 norm

$$\|f_1 - f_2\| = \left[\int (f_1 - f_2)^2 dx \right]^{1/2}$$

to Hellinger distance

$$d(f_1, f_2) = \left[\int (f_1^{1/2} - f_2^{1/2})^2 dx \right]^{1/2}$$

and applying the results of the previous section. Assumption 1 implies $f_o(x) \leq 1$ whereas an estimate \hat{f}_n may not satisfy $\hat{f}_n \leq 1$. If we replace \hat{f}_n by $\tilde{f}_n = \min\{\hat{f}_n, 1\}$, we improve the accuracy of an approximation because $\|f_o - \tilde{f}_n\| \leq \|f_o - \hat{f}_n\|$ and $d(f_o, \tilde{f}_n) \leq d(f_o, \hat{f}_n)$. It may be that $\int \tilde{f}_n dx < 1$ due to this truncation, but this is of no concern in this section. Throughout this section we shall merely presume that truncation has been applied as necessary to guarantee that densities are bounded by one without using a special notation. In consequence, the only density that we may presume integrates to one is the true density f_o .

The next Lemma shows that the Hellinger distance is bounded from below by the L_2 norm times a constant.

LEMMA 3 Under Assumption 1

$$\int [f_1^{1/2}(x) - f_2^{1/2}(x)]^2 dx \geq 2 \int [f_1(x) - f_2(x)]^2 dx.$$

Proof Note that

$$\begin{aligned} \int [f_1^{1/2}(x) - f_2^{1/2}(x)]^2 dx &= \int [f_1^{1/2}(x) - f_2^{1/2}(x)]^2 \frac{[f_1^{1/2}(x) + f_2^{1/2}(x)]^2}{[f_1^{1/2}(x) + f_2^{1/2}(x)]^2} dx \\ &= \int \frac{[f_1(x) - f_2(x)]^2}{[f_1^{1/2}(x) + f_2^{1/2}(x)]^2} dx \\ &\geq 2 \int [f_1(x) - f_2(x)]^2 dx. \end{aligned}$$

The last line uses the fact that $f_1, f_2 \leq 1$. □

Replacing f_1 with \hat{f}_n and f_2 with f_o , we see this result establishes the existence of a sequence of a functions \hat{f}_n for which $L_2 \rightarrow 0$ a.s. because the results of the previous section establish the existence of such a sequence for Hellinger distance.

We also need to establish a similar type of inequality in the other direction. Here and throughout the remainder of this section, C denotes a generic upper bound.

LEMMA 4 Given the conditions in Lemma 3 and some M , $0 < M < \infty$,

$$\int_{-M}^M [f_1^{1/2}(x) - f_2^{1/2}(x)]^2 dx \leq C \int [f_1(x) - f_2(x)]^2 dx.$$

Proof Similar to the proof of Lemma 3, we have

$$\begin{aligned} \int_{-M}^M [f_1^{1/2}(x) - f_2^{1/2}(x)]^2 dx &= \int_{-M}^M \frac{[f_1(x) - f_2(x)]^2}{[f_1^{1/2}(x) + f_2^{1/2}(x)]^2} dx \\ &\leq \frac{1}{4\epsilon_o\phi(M)} \int_{-M}^M [f_1(x) - f_2(x)]^2 dx. \end{aligned}$$

□

This result guarantees that as $L_2^2 \rightarrow 0$, $\int_{-M}^M [f_1^{1/2}(x) - f_2^{1/2}(x)]^2 \rightarrow 0$ at the same rate for $M < \infty$ arbitrarily large. This latter distance is just the Hellinger distance defined on a strict subset of the underlying support. A result which covers the entire support as $n \rightarrow \infty$ is given in the lemma below.

LEMMA 5 Suppose that for some sequence of densities, $\{f_{1,n}, f_{2,n}\}$,

$$\int [f_{1,n}(x) - f_{2,n}(x)]^2 \leq Cn^{-\tau}, \quad \tau > 0.$$

Set $M_n = (2 \log(n^{\tau_M}))^{1/2}$, $0 < \tau_M < \tau$. Then under the conditions in Lemma 3,

$$\int_{-M_n}^{M_n} [f_{1,n}^{1/2}(x) - f_{2,n}^{1/2}(x)]^2 dx \leq Cn^{-\tau+\tau_M}.$$

Proof The proof follows from Lemma 4 since $\phi(M_n) = n^{-\tau_M}/\sqrt{2\pi}$. □

The goal now is to construct a data driven procedure that selects K . Assumption 1 together with the results of the previous section imply that the number of terms should not grow faster than $\mathcal{O}(n^{1/2})$. Therefore, we choose some positive constant C^* and then search for K over the set

$$\mathcal{K}_n = \{1, 2, \dots, \lceil C^* n^{1/2} \rceil\},$$

where $\lceil a \rceil$ is the integer part of a .

Recall that for a given proportion α , $\mathcal{X}_{j,\alpha}$ is the j th block of size $[\alpha n]$ from the sample \mathcal{X} , $\tilde{\mathcal{X}}_{j,\alpha} = \mathcal{X} \setminus \mathcal{X}_{j,\alpha}$ is the remainder of the sample, and $\hat{f}_{j,K}$ is the estimate obtained from the remainder, namely,

$$\hat{f}_{j,K} = \operatorname{argmax}_{f \in \mathcal{F}_K} \sum_{x_i \in \tilde{\mathcal{X}}_{j,\alpha}} \log f(x_i).$$

Recall also that all probability statements below are conditional on $\tilde{\mathcal{X}}_{j,\alpha}$.

We will next establish uniform convergence over \mathcal{K}_n . To do this, set

$$\begin{aligned} \hat{Q}_{j,n}(K) &= \int \hat{f}_{j,K}^2(x) dx - \frac{2}{[\alpha n]} \sum_{x_i \in \mathcal{X}_{j,\alpha}} \hat{f}_{j,K}(x_i) + \int f_o^2(x) dx, \\ Q_{j,n}(K) &= \int \hat{f}_{j,K}^2(x) dx - 2 \int \hat{f}_{j,K}(x) f_o(x) dx + \int f_o^2(x) dx, \end{aligned}$$

and

$$\begin{aligned} \hat{S}_{j,n}(K) &= \hat{Q}_{j,n}(K) + \frac{2}{[\alpha n]} \sum_{x_i \in \mathcal{X}_{j,\alpha}} f_o(x_i), \\ S_{j,n}(K) &= Q_{j,n}(K) + 2 \int f_o^2(x) dx. \end{aligned}$$

In Section 2, choosing \hat{K}_n to minimize $CVH = \hat{M}_{(1)} + 2\hat{M}_{(2)}$ was proposed as a possible truncation point of the SNP estimator. Minimizing, CVH is equivalent to minimizing either $(1/J) \sum_{j=1}^J \hat{Q}_{j,n}(K)$ or $(1/J) \sum_{j=1}^J \hat{S}_{j,n}(K)$ because these two functions only differ from CVH by additive terms that do not depend on K . The population analog of this choice – in the sense of replacing $\hat{M}_{(2)}$ by $M_{(2)}$ — is

$$K_n = \operatorname{argmin}_{K \in \mathcal{K}_n} \frac{1}{J} \sum_{j=1}^J Q_{j,n}(K) = \operatorname{argmin}_{K \in \mathcal{K}_n} \frac{1}{J} \sum_{j=1}^J S_{j,n}(K)$$

The next Assumption is used to help show that $\sup_{K \in \mathcal{K}_n} |\hat{Q}_{j,n}(K) - Q_{j,n}(K)|$ tends to zero at a faster rate than $Q_{j,n}(K_n)$. The rate is needed to establish uniform convergence.

ASSUMPTION 2 For large n ,

$$Q_{j,n}(K_n) \geq Cn^{-1+\zeta},$$

where $\zeta > 0$ is arbitrarily small.

The Assumption rules out the case when $Q_{j,n}(K_n)$ decreases to zero at a rate n^{-1} . An example of this is when the underlying density is a finite sum of Hermite polynomials. We rule out this purely parametric case since valid model selection criteria, such as BIC, has already been investigated extensively in the literature. In contrast, Stone (1982) has shown that in the typical nonparametric setting, $Q_{j,n}(K_n)$ can not converge to zero faster than $n^{-2k/(2k+1)}$, which is slower than $n^{-1+\zeta}$ for small ζ .

We are now ready to state and prove the uniform convergence result.

LEMMA 6 Under Assumptions 1 through 2

$$\sup_{K \in \mathcal{K}_n} \left| \frac{\hat{S}_{j,n}(K) - S_{j,n}(K)}{Q_{j,n}(K)} \right| = o_s(1), \quad j = 1, \dots, J.$$

Proof Fix $K \in \mathcal{K}_n$ and $j \in \{1, \dots, J\}$, and define

$$\begin{aligned} U_{j,n}(K) &= \hat{S}_{j,n}(K) - S_{j,n}(K) \\ &= \frac{2}{[\alpha n]} \sum_{x_i \in \mathcal{X}_{j,\alpha}} [f_o(x_i) - \hat{f}_{j,K}(x_i)] - 2 \int [f_o(x) - \hat{f}_{j,K}(x)] f_o(x) dx, \end{aligned}$$

and note that $\mathcal{E}U_{j,n}(K) = 0$, $|U_{j,n}(K)|$ is bounded, and $V_{j,n}(K) = [\alpha n] \text{Var}[U_{j,n}(K)] \leq CQ_{j,n}(K)$. Then for large n , we have by Bernstein's inequality (Pollard, 1984, p. 193) for small $\eta > 0$,

$$\begin{aligned} P[|U_{j,n}(K)| \geq \eta Q_{j,n}(K)] &\leq 2 \exp \left[-\frac{\frac{1}{2}[\alpha n] \eta^2 Q_{j,n}^2(K)}{V_{j,n}(K) + C\eta Q_{j,n}(K)} \right] \\ &\leq 2 \exp \left[-Cn\eta^2 Q_{j,n}(K) \right] \\ &\leq 2 \exp \left[-Cn^\zeta \eta^2 \right], \end{aligned}$$

where the last inequality follows from Assumption 2. Now we will no longer hold K fixed. For an overall probabilistic bound, observe that

$$P \left[\sup_{K \in \mathcal{K}_n} |U_{j,n}(K)/Q_{j,n}(K)| \geq \eta \right] \leq (\#\mathcal{K}_n) \sup_{K \in \mathcal{K}_n} P[|U_{j,n}(K)/Q_{j,n}(K)| \geq \eta],$$

where $\#\mathcal{K}_n$ is the cardinality of \mathcal{K}_n , which is $C^*n^{1/2}$. This implies a bound of

$$2C^*n^{1/2} \exp(-Cn^\zeta \eta^2) \leq \exp(-Cn^{\zeta/2} \eta^2)$$

for large n . The conclusion follows from the Borel-Cantelli Lemma. \square

THEOREM 1 Under Assumptions 1 through 2

$$\sup_{K \in \mathcal{K}_n} \left| \frac{\sum_{j=1}^J \hat{S}_{j,n}(K) - \sum_{j=1}^J S_{j,n}(K)}{\sum_{j=1}^J Q_{j,n}(K)} \right| = o_s(1)$$

Proof Because

$$\frac{Q_{j,n}(K)}{\sum_{j=1}^J Q_{j,n}(K)} < 1,$$

we have

$$\begin{aligned} \left| \frac{\sum_{j=1}^J \hat{S}_{j,n}(K) - \sum_{j=1}^J S_{j,n}(K)}{\sum_{j=1}^J Q_{j,n}(K)} \right| &\leq \sum_{j=1}^J \frac{|\hat{S}_{j,n}(K) - S_{j,n}(K)|}{\sum_{j=1}^J Q_{j,n}(K)} \\ &= \sum_{j=1}^J \left[\frac{Q_{j,n}(K)}{\sum_{j=1}^J Q_{j,n}(K)} \right] \left[\frac{|\hat{S}_{j,n}(K) - S_{j,n}(K)|}{Q_{j,n}(K)} \right] \\ &\leq \sum_{j=1}^J \frac{|\hat{S}_{j,n}(K) - S_{j,n}(K)|}{Q_{j,n}(K)} \end{aligned}$$

□

We can now show that \hat{K}_n is consistent.

THEOREM 2 Under Assumptions 1 through 2

$$\frac{\sum_{j=1}^J Q_{j,n}(\hat{K}_n)}{\sum_{j=1}^J Q_{j,n}(K_n)} = 1 \quad \text{a.s.}$$

Proof Applying Theorem 1 we have

$$\begin{aligned} 0 &\leq \sum_{j=1}^J Q_{j,n}(\hat{K}_n) - \sum_{j=1}^J Q_{j,n}(K_n) \\ &= \sum_{j=1}^J S_{j,n}(\hat{K}_n) - \sum_{j=1}^J S_{j,n}(K_n) \\ &= \sum_{j=1}^J \hat{S}_{j,n}(\hat{K}_n) - \sum_{j=1}^J S_{j,n}(K_n) + o_s \left[\sum_{j=1}^J Q_{j,n}(\hat{K}_n) \right] \\ &\leq \sum_{j=1}^J \hat{S}_{j,n}(K_n) - \sum_{j=1}^J S_{j,n}(K_n) + o_s \left[\sum_{j=1}^J Q_{j,n}(\hat{K}_n) \right] \\ &= \sum_{j=1}^J S_{j,n}(K_n) - \sum_{j=1}^J S_{j,n}(K_n) + o_s \left[\sum_{j=1}^J Q_{j,n}(\hat{K}_n) \right] + o_s \left[\sum_{j=1}^J Q_{j,n}(K_n) \right] \\ &= o_s \left[\sum_{j=1}^J Q_{j,n}(\hat{K}_n) \right] + o_s \left[\sum_{j=1}^J Q_{j,n}(K_n) \right] \end{aligned}$$

$$\begin{aligned}
&\leq 2o_s \left[\sum_{j=1}^J Q_{j,n}(\hat{K}_n) \right] \\
&= Z_n.
\end{aligned}$$

Then

$$\begin{aligned}
1 &\leq \frac{\sum_{j=1}^J Q_{j,n}(\hat{K}_n)}{\sum_{j=1}^J Q_{j,n}(K_n)} \\
&= \frac{\sum_{j=1}^J [Q_{j,n}(\hat{K}_n) - Q_{j,n}(K_n) + Q_{j,n}(K_n)]}{\sum_{j=1}^J [Q_{j,n}(K_n) - Q_{j,n}(\hat{K}_n) + Q_{j,n}(\hat{K}_n)]} \\
&= \frac{\sum_{j=1}^J Q_{j,n}(K_n) + Z_n}{\sum_{j=1}^J Q_{j,n}(\hat{K}_n) - Z_n} \\
&\leq \frac{\sum_{j=1}^J Q_{j,n}(\hat{K}_n) + Z_n}{\sum_{j=1}^J Q_{j,n}(\hat{K}_n) - Z_n} \\
&= \frac{[1 + 2o_s(1)] \sum_{j=1}^J Q_{j,n}(\hat{K}_n)}{[1 - 2o_s(1)] \sum_{j=1}^J Q_{j,n}(\hat{K}_n)} \\
&\rightarrow 1 \quad \text{a.s..}
\end{aligned}$$

□

6 Conclusion

Through the use of several examples covering a wide range of cases, we uncovered the phenomenon that the MSE curve (in terms of K) tends to rapidly decrease until a certain point, and then it abruptly levels off. Given this, an ideal methodology that selects K would be one that accurately predicts an area of the MSE that is nearly constant, and at the same time, it should be computationally inexpensive. Two of the more common bandwidth procedures in the literature are BIC (and slight deviations thereof) and CVL, but BIC often selects a K which is too small and CVL is relatively expensive to compute because it requires n nonlinear optimizations. As a consequence, we proposed to choose K by minimizing CVH, which only required $n/10$ nonlinear optimizations here. This method did as well as, or better, at predicting the minimum of the MSE curve than CVL, and its theoretical justification was also provided. As a final check on its robustness, we showed that SNP under CVH compares favorably to the optimal kernel estimates.

7 References

- Abramowitz, Milton, and Irene A. Stegun (1972), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York.
- Andersen, Torben G., Hyung-Jin Chung, and Bent E. Sorensen (1999), Efficient Method of Moments Estimation of a Stochastic Volatility Model: A Monte Carlo Study, *Journal of Econometrics* 91,61–87.
- Chumacero, R. (1997), “Finite Sample Properties of the Efficient Method of Moments,” *Studies in Nonlinear Dynamics and Econometrics* 2, 35–51.
- Davidian, Marie, and A. Ronald Gallant (1992), “Smooth Nonparametric Maximum Likelihood Estimation for Population Pharmacokinetics, with Application to Quinidine,” *Journal of Pharmacokinetics and Biopharmaceutics* 20, 531–558.
- Davidian, Marie, and A. Ronald Gallant (1993), “The Nonlinear Mixed Effects Model with a Smooth Random Effects Density,” *Biometrika* 80, 475–488.
- Devroye, Luc, and Laszlo Györfi (1985), *Nonparametric Density Estimation, The L_1 View*, Wiley, New York.
- Eastwood, Brian J., and A. Ronald Gallant (1991), “Adaptive Rules for Semiparametric Estimators That Achieve Asymptotic Normality,” *Econometric Theory* 7, 307–340.
- Fan, Jianqing, Chunming Zhang, and Jian Zhang (2000), “Sieve Likelihood Ratio Statistics and Wilks Phenomenon,” Manuscript, Department of Statistics, University of North Carolina, Chapel Hill NC 27599-3260, USA.
- Fenton, Victor M., and A. Ronald Gallant (1996), “Qualitative and Asymptotic Performance of SNP Density Estimators,” *Journal of Econometrics* 74, 77–118.
- Gallant, A. Ronald, David A. Hsieh, and George E. Tauchen (1991), “On Fitting a Recalcitrant Series: The Pound/Dollar Exchange Rate, 1974–83,” in Barnett, William A., James Powell, and George E. Tauchen, eds., *Nonparametric and Semiparametric*

- Methods in Econometrics and Statistics, Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, Cambridge University Press, Cambridge, Chapter 8, 199–240.
- Gallant, A. Ronald, and Jonathan R. Long (1997), “Estimating Stochastic Differential Equations Efficiently by Minimum Chi-Squared,” *Biometrika* 84, 125–141.
- Gallant, A. Ronald, and Douglas W. Nychka (1987), “Seminonparametric Maximum Likelihood Estimation,” *Econometrica* 55, 363–390.
- Gallant, A. Ronald, Peter E. Rossi, and George Tauchen (1992), “Stock Prices and Volume,” *The Review of Financial Studies* 5, 199–242.
- Gallant, A. Ronald, Peter E. Rossi, and George Tauchen (1993), “Nonlinear Dynamic Structures,” *Econometrica* 61, 871–907.
- Gallant, A. Ronald, and Geraldo Souza (1991), “On the Asymptotic Normality of Fourier Flexible Form Estimates,” *Journal of Econometrics* 50, 329–353.
- Gallant, A. Ronald, and George Tauchen (1999), “The Relative Efficiency of Method of Moments Estimators,” *Journal of Econometrics* 92, 149–172.
- Ghysels, E., A. Harvey, and E. Renault (1995), “Stochastic Volatility,” in ed. G. S. Maddala, *Handbook of Statistics, Vol. 14, Statistical Methods in Finance*, Amsterdam: North Holland.
- Johnson, N. L., and S. Kotz (1994), *Continuous Univariate Distributions-1, 2nd ed.*, New York: Wiley.
- Laffont, J.-J., H. Ossard, and Q. Vuong (1995), “The Econometrics of First Price Auctions,” *Econometrica* 63, 953–980.
- Liu, Ming, (2000), “Modeling Long Memory in Stock Market Volatility,” *Journal of Econometrics*, forthcoming.
- Marron, J. S. (1987), “A Comparison of Cross-Validation Techniques in Density Estimation,” *The Annals of Statistics* 15, 152–162.

- Marron, J. S., and M. P. Wand (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics* 20, 712–736.
- Ng, S., and A. Michaelides (2000), "Estimating the Rational Expectations model of Speculative Storage: a Monte Carlo Comparison of Three Simulation Estimators," *Journal of Econometrics*, forthcoming.
- Pollard, David (1984), *Convergence of Stochastic Processes*, Springer-Verlag, New York.
- Sansone, G. (1991), *Orthogonal Functions*, Dover, New York.
- Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.
- Stone, C. J. (1982), "Optimal Global Rates of Convergence of Nonparametric Regression," *Annals of Statistics* 10, 1040–1053.
- Tauchén, G., and M. Pitts (1983), "The Price Variability-Volume Relationship on Speculative Markets," *Econometrica* 51, 485–505.
- Wong, Wing Hung, and Xiaotong Shen (1995), "Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES," *The Annals of Statistics* 23, 339–362.

Figures

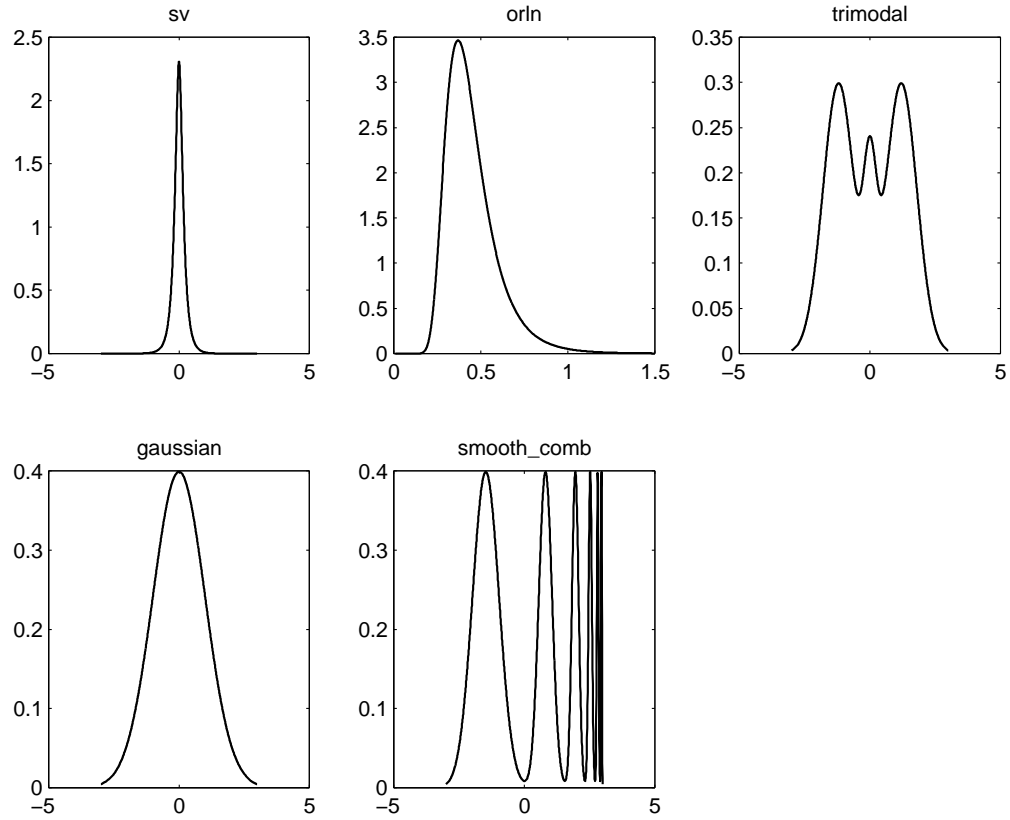


Figure 1. Densities considered. The plot labeled sv is the density of a scale mixture of normals with parameters chosen such that the density has mean 0, variance $1/4$, and raw kurtosis 8; orln is the density of the second largest order statistic in a sample of size 100 from the log normal with location parameter -3 and scale parameter 1. The densities trimodal, gaussian, and smooth_comb are densities from the Marron-Wand test suite.

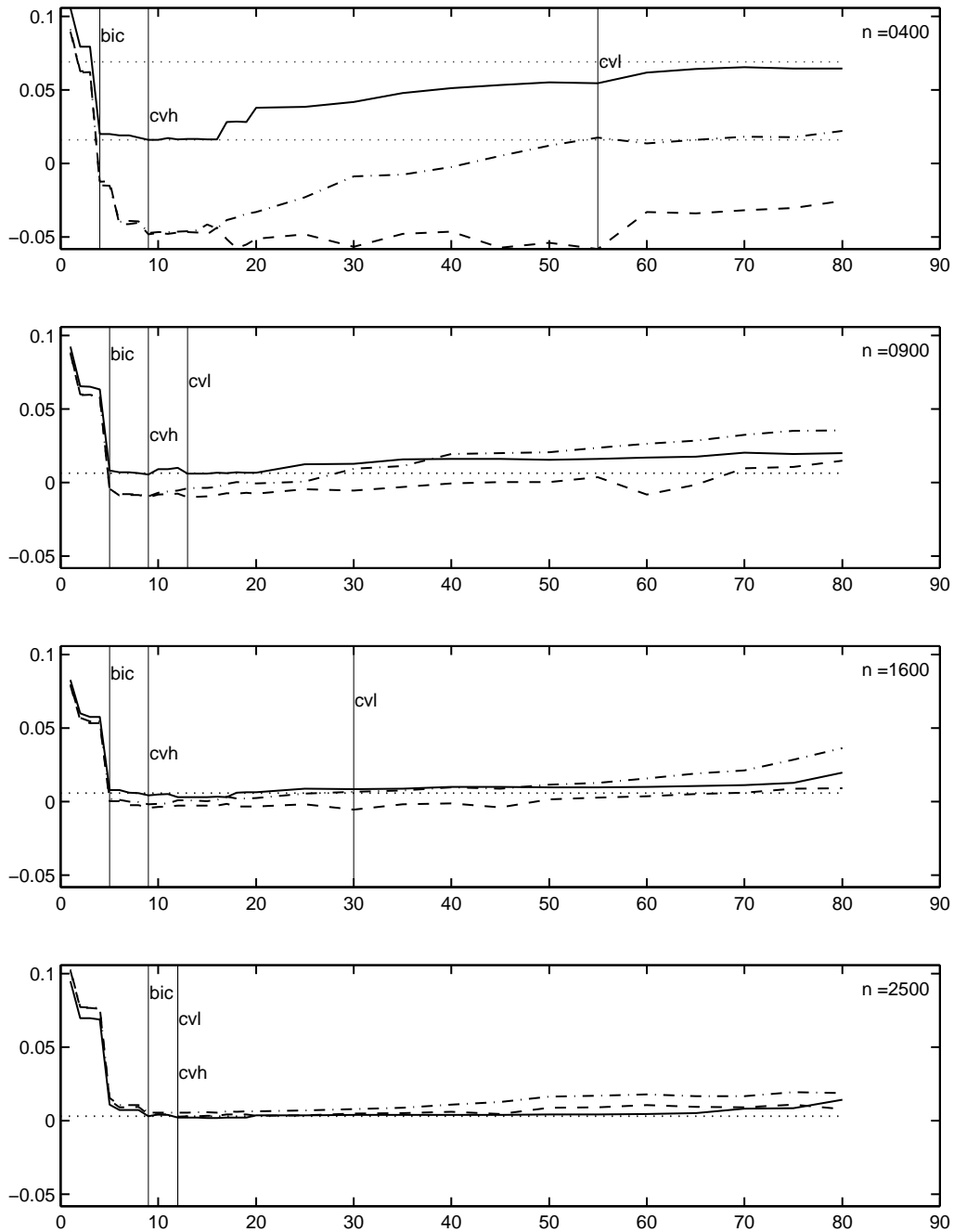


Figure 2. Scale Mixture of Normals. Plotted is the mean squared error (MSE) and its cross validated estimate (CV) for a realization of size n , as shown in each plot, from the density $p(y|\rho) = \int_{-\infty}^{\infty} n(y|\rho_1, e^{2u}) n(u|\rho_2, \rho_3^2) du$ with ρ chosen so that the density has mean 0, variance 1/4, and raw kurtosis 8. Solid line is MSE, dashed line is its leave-one-out CV estimate (CVL), and dashed and dotted line is the average of ten, 10% hold-out-sample CV estimates (CVH). Upper dotted horizontal line is MSE achieved by a cross-validated kernel estimate and lower dotted line is best kernel MSE for this realization. Vertical lines indicate BIC, CVL, and CVH choices of K , as marked.

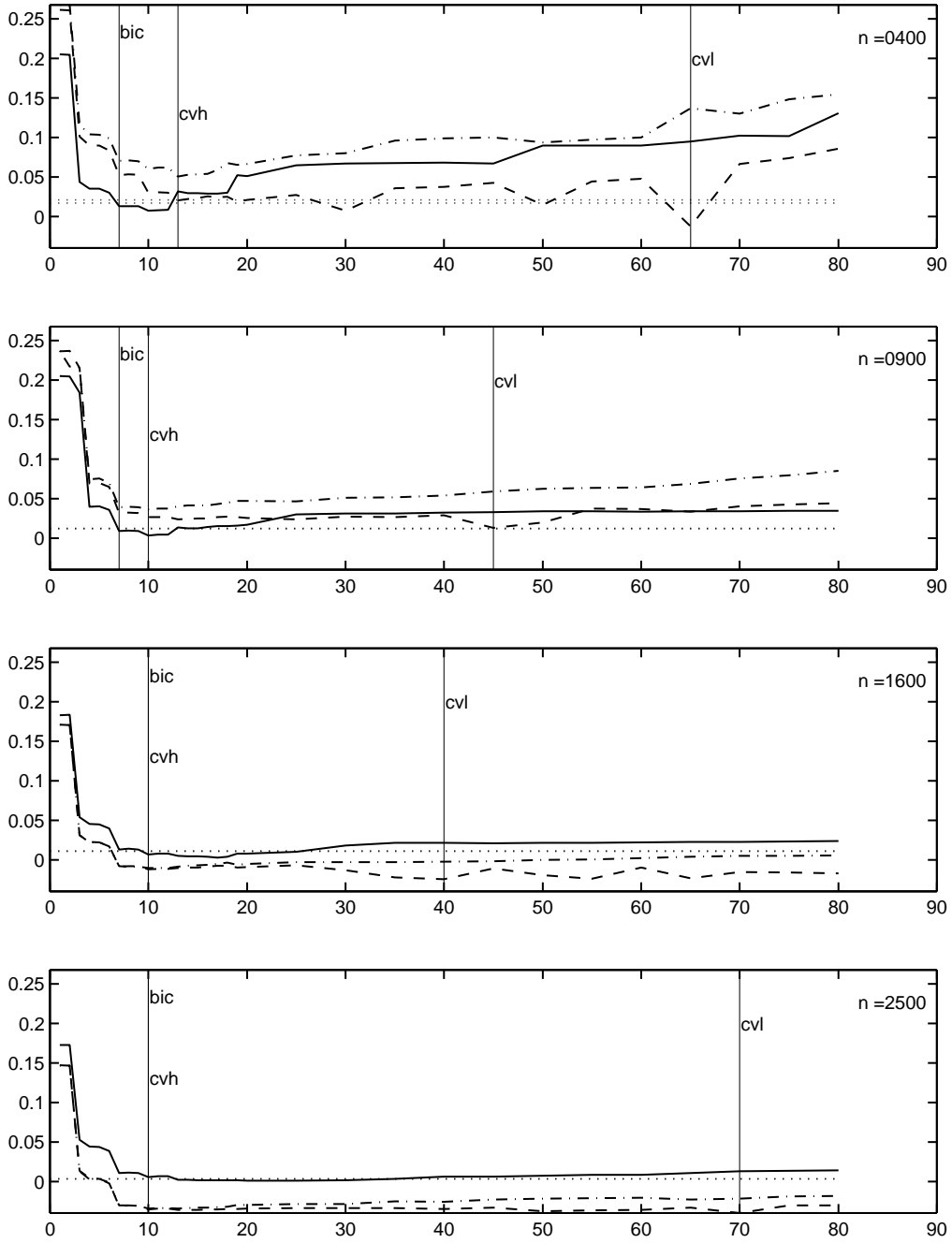


Figure 3. Second Largest Order Statistic of the Lognormal. Plotted is the mean squared error (MSE) and its cross validated estimate (CV) for a realization of size n , as shown in each plot, from the density $p(y|\rho) = \frac{N(N-1)}{y} \left[\Phi\left(\frac{\log y - \rho_2}{\rho_3}\right) \right]^{N-2} \left[1 - \Phi\left(\frac{\log y - \rho_2}{\rho_3}\right) \right] \phi\left(\frac{\log y - \rho_2}{\rho_3}\right)$ where $y > 0$, ϕ and Φ denote the standard normal density and distribution functions, respectively, and $(N, \rho_2, \rho_3) = (100, -3, 1)$. Solid line is MSE, dashed line is its leave-one-out CV estimate (CVL), and dashed and dotted line is the average of ten, 10% hold-out-sample CV estimates (CVH). Upper dotted horizontal line is MSE achieved by a cross-validated kernel estimate and lower dotted line is best kernel MSE for this realization. Vertical lines indicate BIC, CVL, and CVH choices of K , as marked.

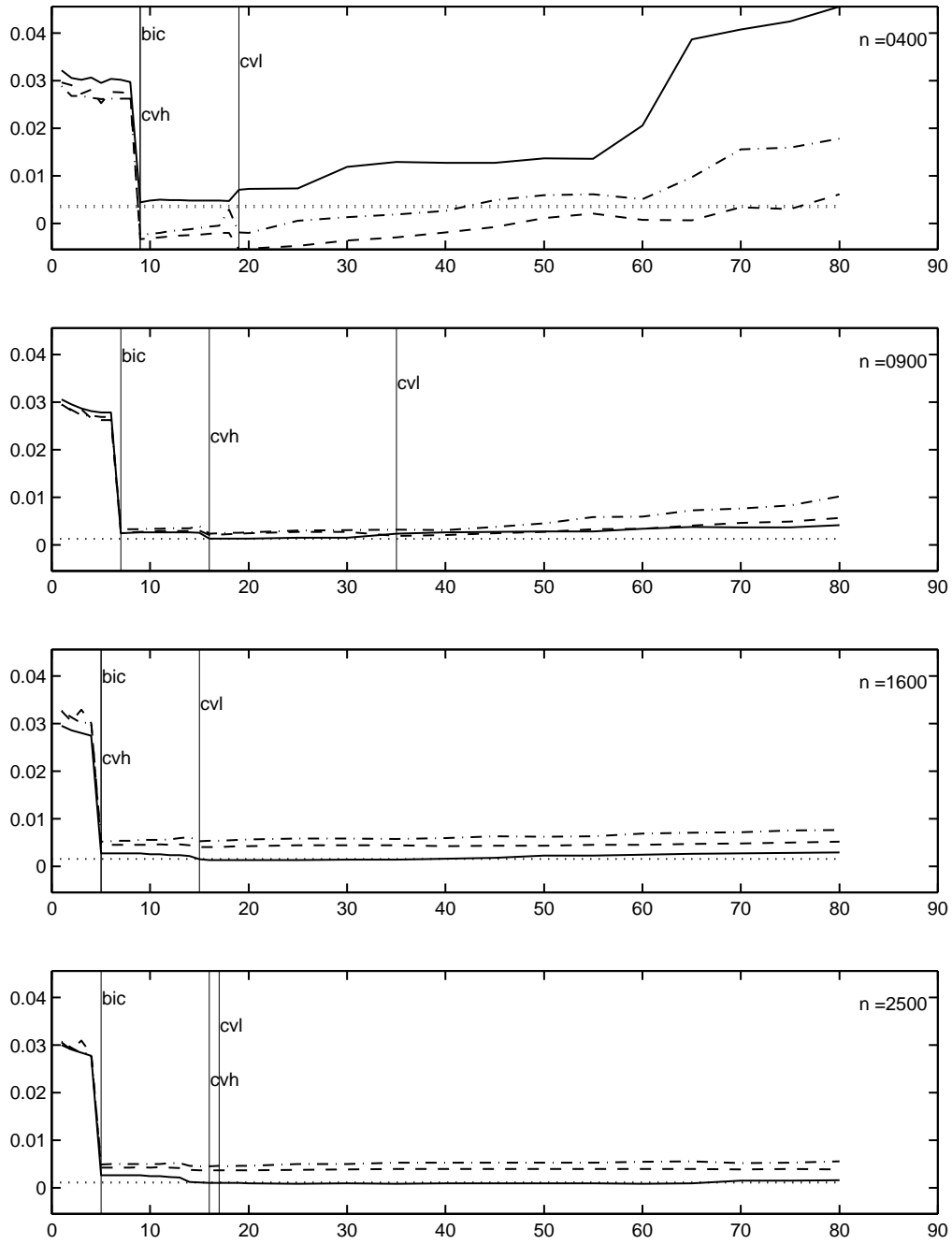


Figure 4. Trimodal. Plotted is the mean squared error (MSE) and its cross validated estimate (CV) for a realization of size n , as shown in each plot, from the trimodal density of the Marron-Wand test suite. Solid line is MSE, dashed line is its leave-one-out CV estimate (CVL), and dashed and dotted line is the average of ten, 10% hold-out-sample CV estimates (CVH). Upper dotted horizontal line is MSE achieved by a cross-validated kernel estimate and lower dotted line is best kernel MSE for this realization. Vertical lines indicate BIC, CVL, and CVH choices of K , as marked.

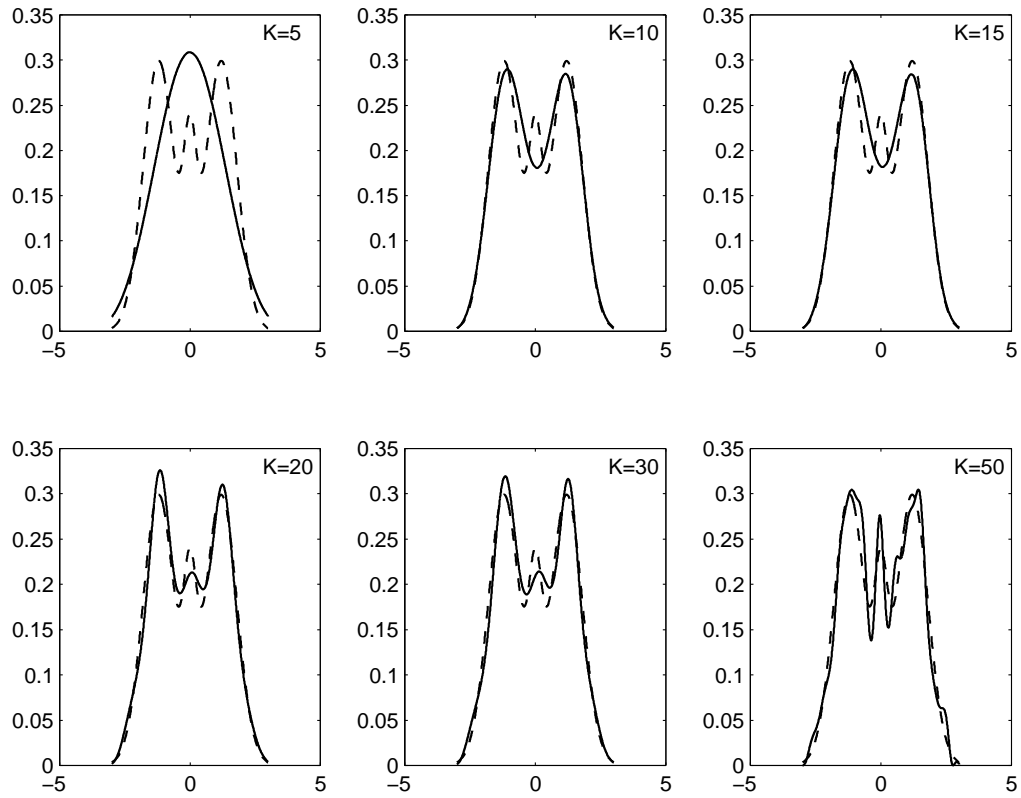


Figure 5. Trimodal. Plotted are SNP density estimates from a realization of size 900 and values of K as shown in each plot, from the trimodal density of the Marron-Wand test suite. Solid line is the estimate, dashed line is true density.

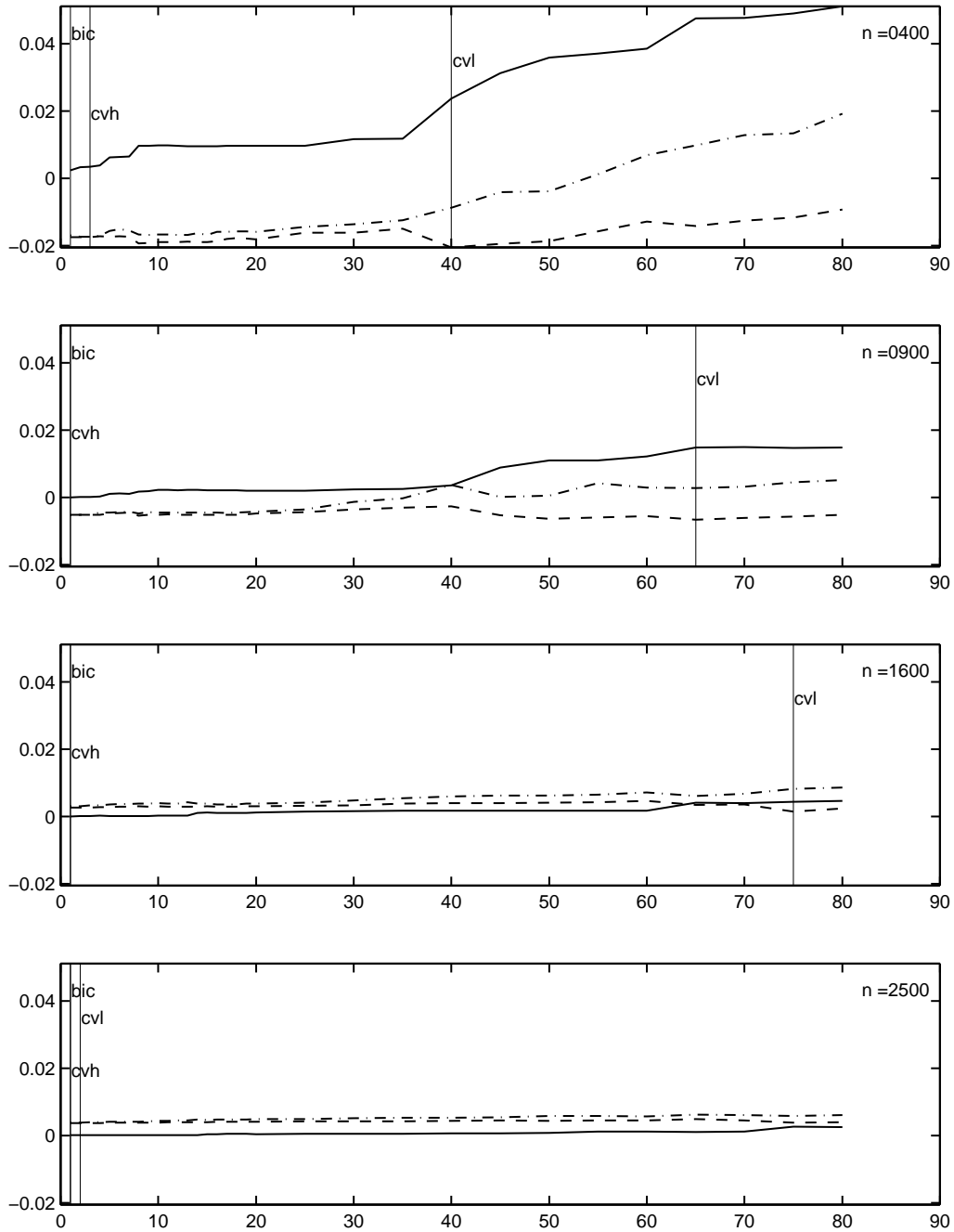


Figure 6. Gaussian. Plotted is the mean squared error (MSE) and its cross validated estimate (CV) for a realization of size n , as shown in each plot, from the gaussian density of the Marron-Wand test suite. Solid line is MSE, dashed line is its leave-one-out CV estimate (CVL), and dashed and dotted line is the average of ten, 10% hold-out-sample CV estimates (CVH). Vertical lines indicate BIC, CVL, and CVH choices of K , as marked.

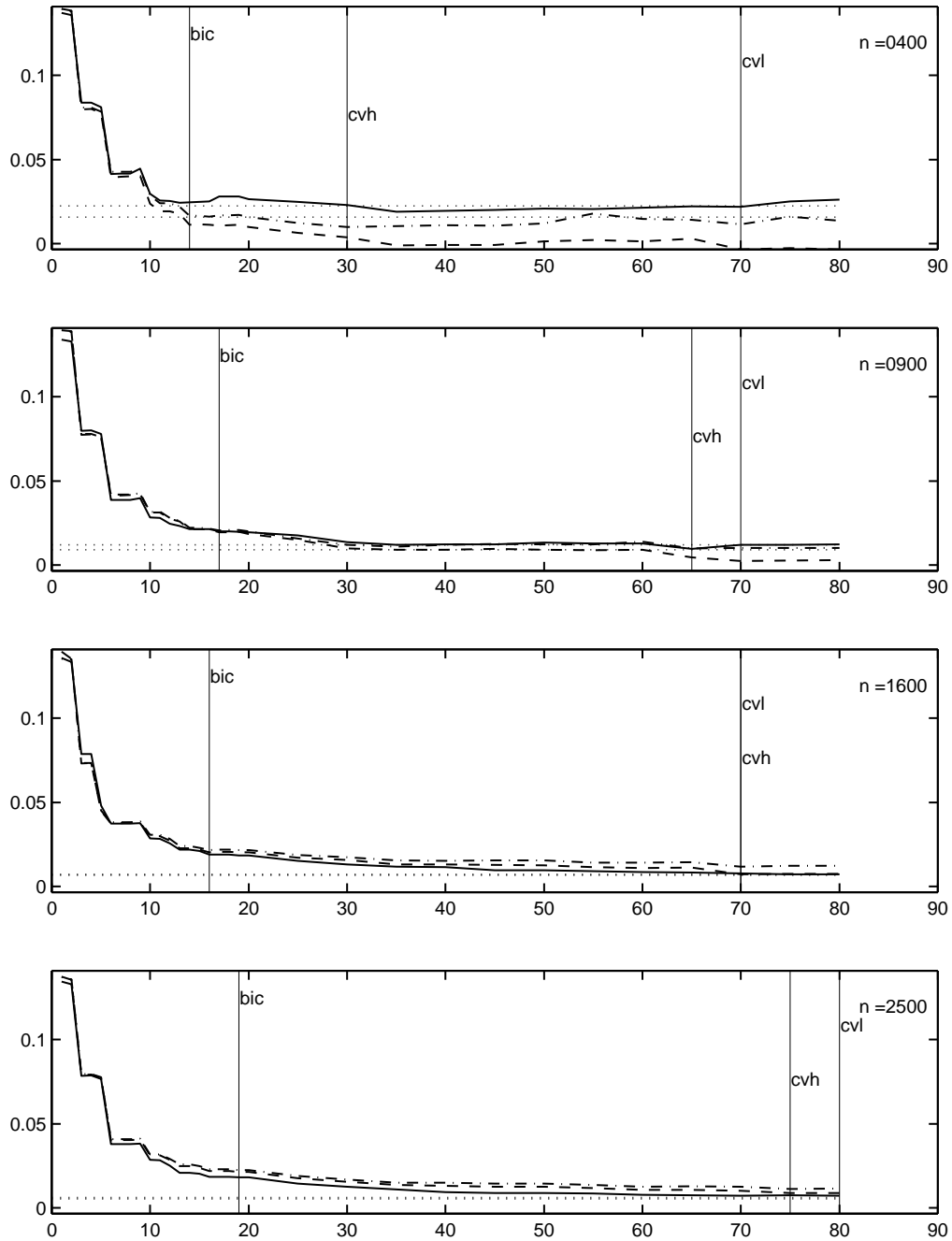


Figure 7. Smooth Comb. Plotted is the mean squared error (MSE) and its cross validated estimate (CV) for a realization of size n , as shown in each plot, from the smooth comb density of the Marron-Wand test suite. Solid line is MSE, dashed line is its leave-one-out CV estimate (CVL), and dashed and dotted line is the average of ten, 10% hold-out-sample CV estimates (CVH). Upper dotted horizontal line is MSE achieved by a cross-validated kernel estimate and lower dotted line is best kernel MSE for this realization. Vertical lines indicate BIC, CVL, and CVH choices of K , as marked.