Statistica Sinica **8**(1998), 607-646

MODEL INDEXING AND SMOOTHING PARAMETER SELECTION IN NONPARAMETRIC FUNCTION ESTIMATION

Chong Gu

Purdue University

Abstract: Smoothing parameter selection is among the most intensively studied subjects in nonparametric function estimation. A closely related issue, that of identifying a proper index for the smoothing parameter, is however largely neglected in the existing literature. Through heuristic arguments and simple simulations, we show that most current working indices are conceptually "incorrect", in the sense that they are not interpretable across-replicate in repeated experiments. As a con sequence, a few popular working concepts, such as expected mean square error and "degrees of freedom", appear vulnerable under close scrutiny. Due to technical constraints, the arguments are mainly developed in the penalized likelihood setting, but conceptual parallels can be drawn to other settings as well. In the light of our findings, simulations and discussion are also presented to compare the relative merits of the simple cross-validation method versus the more sophisticated plug-in method for smoothing parameter selection, and to explore related issues. The development stems from an attempt to understand the well-publicized negative correlation between optimal and cross-validation smoothing parameters, which however turns out to bear little statistical relevance.

Key words and phrases: Constraint, cross-validation, kernel method, negative correlation, penalized likelihood, plug-in method.

1. Introduction

Smoothing parameter selection plays an important role in practical nonparametric function estimation. In spite of the ever growing number of procedures being proposed and theorems being proved, there remain some basic concepts to be clarified, some gaps between theory and practice to be patched, and some counter-intuitive phenomena to be understood. Through the assessment of the interpretability of the smoothing parameters commonly in use, we illustrate in this article that most working indices of smoothing parameter are conceptually "incorrect", in the sense that they are not interpretable across-replicate in repeated experiments, and consequently the commonly accepted intuitions are somewhat distorted. Our conclusions concerning the philosophies of smoothing parameter selection and the relative merits of the cross-validation method versus the plug-in method are at odds with some of the recent literature.

We consider a regression problem for simple exposition, but the arguments readily apply to other problems. Given observations

$$Y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $x_i \in [0,1]$ and $\epsilon_i \sim N(0, \sigma^2)$, one is to estimate $f(x)$. The issues under discussion are the statistical models behind nonparametric estimates, their proper indexing, and the ramifications in smoothing parameter selection.

Our arguments will be developed under the setting of penalized likelihood estimation. Assume $f(x)$ is smooth in the sense that its second derivative exists and is small. A popular approach to the estimation of $f(x)$ is through minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \lambda \int_0^1 \ddot{f}^2(x)dx, \tag{2}$$

where the least squares term discourages lack of fit, the smoothness functional $\int_0^1 \ddot{f}^2(x)dx$ penalizes roughness, and the smoothing parameter λ controls the tradeoff. The minimizers of (2) with $\lambda \in (0, \infty)$, known as the cubic smoothing splines, define a continuum of estimates. When $\lambda \to \infty$, one obtains the simple linear regression line; when $\lambda \to 0$, one computes the minimum curvature interpolant. The practicability of the method hinges on a good choice of λ, the selection of a good model from a continuum of available models. A comprehensive treatment of (2) and related methods can be found in Wahba (1990).

An alternative derivation of smoothing splines is through a constrained least squares problem, which minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(x_i))^2, \qquad s.t. \int_0^1 \ddot{f}^2(x)dx \le \rho \tag{3}$$

for some $\rho \ge 0$. The solution of (3) usually falls on the boundary $\int_0^1 \ddot{f}^2(x)dx = \rho$ and, by the Lagrange method, it can be calculated as the minimizer of (2) with an appropriate Lagrange multiplier λ. Thus, up to the choice of λ and ρ, a penalized likelihood problem with a penalty proportional to $\int_0^1 \ddot{f}^2(x)dx$ is equivalent to a constrained maximum likelihood problem subject to a soft constraint of the form $\int_0^1 \ddot{f}^2(x)dx \le \rho$. See, e.g., Schoenberg (1964).

Given the least squares functional $(1/n)\sum_{i=1}^{n}(Y_i - f(x_i))^2$, which is dependent on the data Y_i, the mapping from ρ to λ is one-to-one, but an important fact is that the mapping changes with the least squares functional. That is, for a fixed constraint $\int_0^1 \ddot{f}^2(x)dx \le \rho$, the Lagrange multiplier λ varies with the data Y_i; conversely, a fixed λ in (2) implies *different* binding constraints on the estimates for different data. This simple observation, that ρ and λ are *not* equivalent as model indices, is a key to an understanding of the discussion.

In Section 2 we present some heuristics and a simple simulation to show that the λ index of smoothing splines is not interpretable across-replicate while the ρ index is, and consequently any across-replicate concepts directly indexed by λ are likely to mislead. In section 3 we discuss the ramifications of model indexing in smoothing parameter selection, and demonstrate that the counter-intuitive negative correlation between the optimal and the cross-validation smoothing parameters is actually an illusion due to improper model indexing. In Section 4 the relative merits of the cross-validation method and the plug-in method for smoothing parameter selection are compared in the context of kernel density estimation. There we show that the former is as competitive as the latter, that the latter also demonstrates a negative correlation, and that the published theory may have little to do with the practical performance of the plug-in method. In Section 5 we briefly summarize our findings and discuss the implications.

2. Model Indexing

Statistical estimation can be viewed as a compromise between the data and the model, the assumptions one makes about the scheme in which the data are generated. In classical parametric estimation a statistical model often consists of two parts: a random part represented by the likelihood function, and a systematic part characterized by a certain constraint. For example, a parametric model $f(x) = f(x, \theta)$ for the systematic part $f(x)$ in (1) simply represents a rigid constraint. For a general statistical procedure, it may not always be possible to explicitly describe the effective constraint, and the constraint actively in force may not comport with the stated assumptions. It seems always possible, however, to perceive conceptually some effective constraint with which the data make compromise in an estimation procedure. With such effective constraint in mind, we have the following heuristic.

Heuristic 1. *The model behind an estimate is characterized by the constraint to which the estimate is subject.*

For smoothing spline estimates, it is clear that the effective constraints are characterized by the ρ index in the constrained least squares formulation of (3). For other procedures, the effective constraint may remain an abstract notion impossible to quantify, yet the mere awareness of such notion may caution one to stay away from otherwise tempting conceptual pitfalls.

The discrepancies between the estimates and the truth are usually measured via loss or risk functions. Intuitively, the performance of an estimate relative to other estimates based on the same data should be largely determined by how close the effective model (i.e., constraint) is to the state of nature, as compared to the effective models behind the other estimates. The state of nature does not change over replicates in an experiment with a fixed stochastic structure except for minor random fluctuations, hence there should be a single optimal model yielding the (nearly) best-performing estimates for all replicates, provided that the same set of effective constraints are reproduced by the procedure over replicates. This leads to our second heuristic.

Heuristic 2. *The optimal models should remain largely invariant over replicated data from the same stochastic structure.*

For (1), Heuristic 2 means that the optimal strategy among given choices should only depend on the true $f(x)$ and the stochastic behavior of ϵ_i, but not on the specific realization of ϵ_i.

Now consider a simple simulation. On $x_i = (i - .5)/100$, $i = 1, \ldots, 100$, we generated 100 replicates of data from (1) with $f(x) = 1+3\sin(2\pi x-\pi)$ and $\sigma^2 = 1$. For λ on a fine grid of $\log_{10} n\lambda = (-5)(.05)(-1)$, we calculated the minimizers of (2) for each of the replicates, and determined retrospectively the effective constraint an estimate $\hat{f}(x)$ had been subject to by calculating $\rho = \int_0^1 \ddot{\hat{f}}^2(x)dx$. The best-performing estimate on the grid was identified for each of the replicates, with the performance of $\hat{f}(x)$ as an estimate of $f(x)$ being measured by the mean square error at the sampling points: $(1/100)\sum_{i=1}^{100}(\hat{f}(x_i) - f(x_i))^2$. The grid was broad enough to bracket the best-performing estimates for all the 100 replicates.

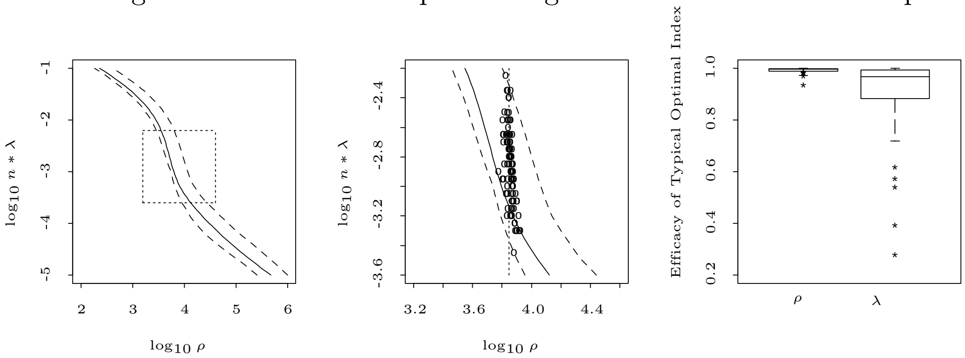

Figure 1. The ρ and λ indices of smoothing spline estimates in simulation. Left and center: Empirical relation between ρ and λ and the optimal indices. Right: Efficacies of typical optimal ρ and λ indices.

The left frame of Figure 1 depicts the mapping between the λ index and the ρ index in our simulation, where the solid curve plots the mapping for the first replicate and the dashed lines sketch an envelope surrounding the bundle of 100 such curves. The window marked by the dotted lines is amplified in the center frame of Figure 1, where the indices of the best-performing estimates are superimposed as circles and the ρ of the true function $\int_0^1 \ddot{f}(x)dx = 10^{3.846}$ is marked by the vertical dotted line. It is reassuring to see that the optimal models scatter around the dotted line. To comprehend the magnitudes of the scatters of the optimal indices, we examine the relative performance of some typical optimal index at the middle of the cloud. We pick $\log_{10} \rho = 3.846$ as a typical optimal ρ and $\log_{10} n\lambda = -2.8$ as a typical optimal λ, and assess their efficacy by calculating for each replicate the ratio of the mean square error of the best-performing estimate to that of the estimate with the typical optimal index. The right frame of Figure 1 summarizes these ratios in box plots, which indicate that the scatter of the optimal λ indices is an order of magnitude greater than the scatter of the optimal ρ indices. By Heuristic 1, the ρ index for models behind the estimates has a clear statistical meaning as it characterizes the constraints to which the estimates are subject; Heuristic 2 extends further support to the ρ index through the above simulation. In contrast, the λ index is *not* statistically interpretable across-replicate in this setting, although it sometimes helps replicate-specific calculations as we shall note shortly.

Denote by f_λ the minimizer of (2) for fixed λ, and by f_ρ the solution of (3) for fixed ρ. A tempting criterion for the assessment of penalized likelihood estimates is the expected mean square error of f_λ indexed by λ:

$$R(\lambda) = E\frac{1}{n}\sum_{i=1}^{n}(f_\lambda(x_i) - f(x_i))^2, \tag{4}$$

where the expectation is with respect to ϵ_i. This seemingly natural criterion is unfortunately defective: a fixed λ implies different models for different realizations of ϵ_i, so the expectation is mixing apples with oranges. Concepts based on the exact quantification of (4), such as the minimizer of $R(\lambda)$ as the across-replicate "optimal" λ, are hence misleading. One may nevertheless legitimately define an expected mean square error for f_ρ indexed by ρ, and discuss the across-replicate optimal ρ, although analysis of the constrained problem is less tractable.

Caution 1. *Working with an index that is not interpretable across-replicate, concepts based on a risk function can mislead.*

Despite the conceptual defect in $R(\lambda)$, the right-hand-side of (4) can be useful in determining the rates, but not the exact quantifications, of the asymptotic behavior of the minimizers of (2). One may calculate a rate $E(1/n)\sum_{i=1}^{n}(f_\lambda(x_i) -f(x_i))^2 = O(K)$, with K an expression in n and λ, and then convert the rate to $(1/n)\sum_{i=1}^{n}(f_\lambda(x_i) - f(x_i))^2 = O_p(K)$, which concerns a replicate-specific loss function.

Denote the fitted value by $\hat{Y}_i = f_\lambda(x_i)$. Fixing λ, the minimizer of (2) forms a so-called linear smoother in the sense that $\hat{\boldsymbol{Y}} = A(\lambda)\boldsymbol{Y}$, where $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ are vectors of Y_i and $\hat{Y}_i$, respectively, and $A(\lambda)$ is a so-called smoothing matrix or hat matrix indexed by λ; see, e.g., Buja, Hastie and Tibshirani (1989) and Wahba (1990). A popular concept in data smoothing is the so-called "degrees-of-freedom", defined as the trace of $A(\lambda)$ or that of a related matrix. Given x_i, $\lambda \leftrightarrow A(\lambda)$ is one-to-one, so the "degrees-of-freedom" index of models is, unfortunately, a repackaging of the λ index. In the above simulation, the trace of $A(\lambda)$ corresponding to the optimal λ index ranges from 5.08 to 9.14.

Caution 2. *The popular notion of "degrees-of-freedom" in nonparametric regression does not seem to convey the proper intuition for model complexity.*

In parametric regression, the trace of the hat matrix happens to match the dimension of the model space, which provides an intuitive characterization of the binding effect of the model. The concept of degrees-of-freedom rests only with the dimension, but not with the trace. For example, there is no hat matrix in parametric density estimation, yet there still is a degrees-of-freedom.

3. Smoothing Parameter Selection

For practical estimation one has to choose a particular ρ or λ to calculate an estimate, and it is rarely the case that a good choice of ρ or λ can be determined *a priori.* The practice of using a linear smoother with predetermined "degrees-of-freedom", or using the minimizer of (2) with a fixed λ, is hardly a defensible strategy, for the choice of ρ would then be up to the specific realization of ϵ_i in (1). Unless a proper value of $\rho = \int_0^1 \ddot{f}^2(x)dx$ can be assumed, which is not too far from a parametric assumption, effective data-driven model selection procedures are necessary for the method to be of any practical use.

For data-specific calculations, the ρ index and the λ index are equivalent. Because the penalized problem is much easier to deal with, the λ index is most convenient for operational purposes. The objective of model selection is thus to locate a *data-specific* optimal λ, say the minimizer of

$$L(\lambda|\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n}(f_{\lambda|\boldsymbol{Y}}(x_i) - f(x_i))^2,$$

where the dependence of f_λ on the data is made explicit, and it is necessary to keep any λ selection procedure data-specific. As a side remark, we note that naive resampling procedures should *not* be used in λ-indexed model selection without proper justification, for the optimal λ for a resample may not necessarily be good for the observed data.

Caution 3. *Working with an index that is not interpretable across-replicate, a proper model selection method should be data-specific.*

An effective model selection procedure for regression is Craven and Wahba's (1979) generalized cross-validation, which selects the minimizer of

$$V(\lambda|\boldsymbol{Y}) = \frac{\boldsymbol{Y}^T(I - A(\lambda))^2\boldsymbol{Y}/n}{[\mathrm{trace}(I - A(\lambda))/n]^2}$$

for use in (2), where the matrix $A(\lambda)$ is as defined in Section 2. The score $V(\lambda|\boldsymbol{Y})$ is data-specific and its minimizer λ_* can be shown to approximately minimize the data-specific loss function $L(\lambda|\boldsymbol{Y})$, in the sense that $1 - \min_\lambda L(\lambda|\boldsymbol{Y})/L(\lambda_*|\boldsymbol{Y}) = o_p(1)$, of course under conditions; see Li (1986). Note that the data in $V(\lambda|\boldsymbol{Y})$ and $L(\lambda|\boldsymbol{Y})$ have to be the same to make this work. Actually, Li's (1986) result can be expressed as

$$V(\lambda|\boldsymbol{Y}) - L(\lambda|\boldsymbol{Y}) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p(L(\lambda|\boldsymbol{Y})), \tag{5}$$

where $\boldsymbol{\epsilon}$ is the vector of unobservable noise ϵ_i in (1). $V(\lambda|\boldsymbol{Y})$ is thus a consistent estimate of the relative loss $L(\lambda|\boldsymbol{Y}) + n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$. With small probability the term $o_p(L(\lambda|\boldsymbol{Y}))$ may not be negligible relative to $L(\lambda|\boldsymbol{Y})$, and this is when the method may fail to identify a (data-specific) good λ.

We now continue the simulation of Section 2 by evaluating the performance of generalized cross-validation on the 100 replicates. Plotted in the left frame of Figure 2 are the loss of the cross-validated estimates $L(\lambda_*|\boldsymbol{Y})$ versus the loss of the best-performing estimates $\min_\lambda L(\lambda|\boldsymbol{Y})$ for each of the replicates. A point on the dotted line indicates a perfect performance of the procedure. As explained above the method may malfunction with small probability, and indeed it worked rather poorly on a few of the replicates. The general performance however appears satisfactory.

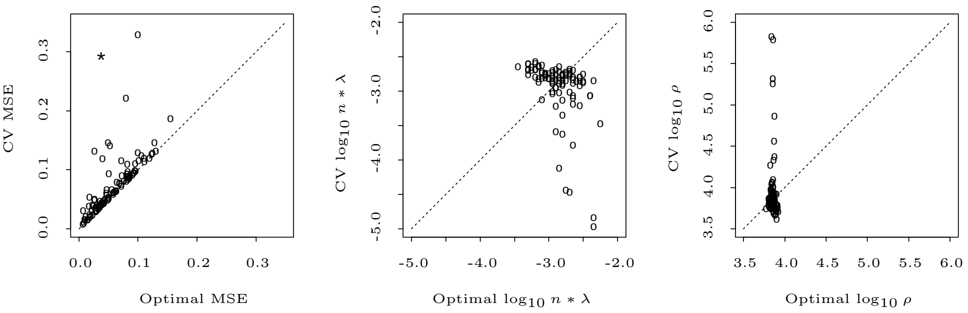

Figure 2. The performance of generalized cross-validation in simulation. Left: Cross-validation MSE versus optimal MSE. Center: Cross-validation λ versus optimal λ. Right: Cross-validation ρ versus optimal ρ.

To visually perceive how bad things can be, we take a closer look as the replicate plotted as the star in the left frame of Figure 2. It recorded the lowest efficacy $\min_\lambda L(\lambda|\boldsymbol{Y})/L(\lambda_*|\boldsymbol{Y}) = .133$ among the 100 replicates. Plotted in Figure 3 are the data from the worst replicate as the circles, the cross-validated estimate as the solid line, the best performing estimate as the dashed line, and the test function as the dotted line. There seem to be some indications that the method was "fooled" by the data. For example, the local dip near $x = .75$ was apparently responding to the data pattern at that locale.

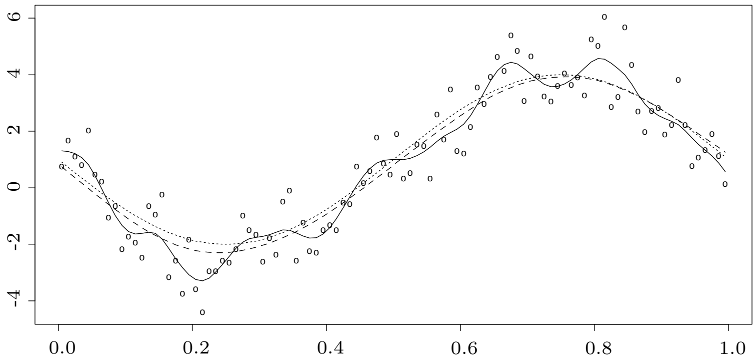

Figure 3. A poor performance of generalized cross-validation. The circles are the data, the solid line the cross-validated estimate, the dashed line the best-performing estimate, and the dotted line the test function.

In the course of the above simulation, we have collected sufficient information to reproduce the well-publicized negative correlation between the optimal and the cross-validation smoothing parameters in the center frame of Figure 2, where the λ index of the cross-validated estimates is plotted against that of the best-performing estimates. Scott and Terrell (1987) and Hall and Johnstone (1992) made the observation concerning a few versions of cross-validation under various problem settings, and charged cross-validation for performing counter-intuitively. Were the λ index interpretable across-replicate, as was usually perceived, the negative correlation would indeed signal an alarm against the use of cross-validation in practice. In the light of our previous discussion, however, the points in the center frame of Figure 2 are *not* comparable with each other, and hence the whistle can be a false alarm. Plotting the more relevant ρ index of the cross-validated estimates versus that of the best-performing estimates in the right frame of Figure 2, we see that the negative correlation no longer exists. There is nearly a single optimal model which generalized cross-validation tries to adopt, but due to the error in the estimation of the relative loss $L(\lambda|\boldsymbol{Y})+n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ by $V(\lambda|\boldsymbol{Y})$, the models actually adopted are scattered nearby, except for a few wild failures when the error term $o_p(L(\lambda|\boldsymbol{Y}))$ in (5) gets out of control.

Caution 4. *Working with an index that is not interpretable across-replicate, the relation between the optimal index and the cross-validation index has no logical bearing on the statistical performance of cross-validation.*

Cross-validation may not be the final word for smoothing parameter selection, but whatever a better procedure is going to be, as long as it is λ-indexed, it should be after a data-specific loss function. Also, a procedure can not be expected to work all the time, as the decisions have to be based on stochastic data.

In settings other than Gaussian regression, such as density estimation, a strategy for data-specific λ selection in penalized likelihood estimation can be found in Gu (1992) and (1993). In simulations similar to that reported above, the procedure demonstrates the same qualitative performance as that of generalized cross-validation as depicted in the three frames of Figure 2, including the negative correlation of the λ indices.

As a final note, we point out a gap between the theory and the practice of generalized cross-validation in the early literature. As we see it now, the original theoretical result presented by Craven and Wahba (1979), that the minimizer of $EV(\lambda|\boldsymbol{Y})$ approximately minimizes $EL(\lambda|\boldsymbol{Y})$, has no logical bearing on the practical performance of generalized cross-validation as simulated by the original authors and many others, including the current one. Despite the oversight in their presentation of the theory, however, the general arguments of Craven and Wahba (1979) (with a more careful interpretation) can be seen to lead to the more relevant result as quoted above, proved later by Li (1986). The derivative result of Gu (1990) in the context of non-Gaussian regression suffers from the same logical defect, which can be similarly fixed by applying the result of Li (1986).

4. Cross-Validation versus Plug-in Method

Perceived as superior to the cross-validation method, the so-called plug-in procedures have been mushrooming in the recent literature on smoothing parameter selection; see, e.g., Jones, Marron, and Sheather (1992) for a review and a list of related references. In the light of our previous discussion, however, we see a possible conceptual defect in the philosophy behind the plug-in procedures, to be noted below. To address related issues including the relative merits of the cross-validation method versus the plug-in method, simulations and discussion will be presented in the context of kernel density estimation, where the development of the plug-in procedures appears mature enough to warrant a review.

Observing X_i, $i = 1, \ldots, n$ from a probability density $f(x)$, one is to estimate $f(x)$ by a function of the form

$$f_h(x|\boldsymbol{X}) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}), \tag{6}$$

where the kernel $K(x)$ is some known smooth function satisfying $\int K(x)dx = 1$, $\boldsymbol{X}$ is the data vector $(X_1, \ldots, X_n)$, and the so-called bandwidth h acts as the smoothing parameter. The bandwidth h appears to be the only model index available for one to work with, and an explicit description of the effective constraint seems nowhere in sight.

One commonly referenced cross-validation procedure for the choice of h is based on the least squares cross-validation score

$$V_{LS}(h|\boldsymbol{X}) = \int f_h^2(x|\boldsymbol{X})dx - (2/n)\sum_{i=1}^{n} f_h(X_i|\boldsymbol{X}_{(i)}),$$

which, up to a term independent of h, estimates the integrated square error

$$L_{ISE}(h|\boldsymbol{X}) = \int (f_h(x|\boldsymbol{X}) - f(x))^2 dx.$$

$\boldsymbol{X}_{(i)}$ in V_{LS} denotes the data vector $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Another popular criterion is the Kullback-Leibler cross-validation score

$$V_{KL}(h|\boldsymbol{X}) = -\sum_{i=1}^{n} \log f_h(X_i|\boldsymbol{X}_{(i)})$$

which targets the Kullback-Leibler discrepancy

$$L_{KL}(h|\boldsymbol{X}) = \int \log\{f(x)/f_h(x|\boldsymbol{X})\} f(x)dx.$$

See, e.g., Rudemo (1982) for the derivation of these scores. The plug-in procedures are based on the asymptotic expansion of $EL_{ISE}(h|\boldsymbol{X})$, where estimates are plugged in for the leading terms of the expansion which usually contain functionals of derivatives of $f(x)$, and the resulting estimated truncated asymptotic expansion of $EL_{ISE}(h|\boldsymbol{X})$ is minimized with respect to h. Among the many plug-in procedures, that of Sheather and Jones (1991) is considered to be the most reliable; see Jones, Marron and Sheather (1992).

Among the criticisms against the cross-validation procedures, which served as part of the motivations for the development of the plug-in method, were the large magnitude of sample variability of the cross-validation bandwidths and the negative correlation similar to what we saw in the center frame of Figure 2. The relevance of these criticisms, and the very conceptual validity of the plug-in method itself, however, hinges on the presumed across-replicate interpretability of the h index.

We now carry out a simple simulation to check on the interpretability of the h index. The test density $f_1(x)$ was taken as the half-half mixture of $N(.35, (.1)^2)$ and $N(.65, (.1)^2)$, the same as the $f_1(x)$ test density of Sheather and Johns (1991) but shifted and scaled into $[0, 1]$. We generated 100 replicates of data of size $n = 100$, and used the standard normal density for the kernel function $K(x)$. The estimates $f_h(x|\boldsymbol{X})$ were calculated for h on a fine grid $\log_{10} h = (-2.3)(.02)(-.7)$, and the losses $L_{ISE}(h|\boldsymbol{X})$ and $L_{KL}(h|\boldsymbol{X})$ for each of the estimates were calculated by summation over an equally spaced mesh of 500 points in $[0, 1]$. Following a suggestion by an anonymous reviewer, a $\rho^* = \int \ddot{f}_h^2(x|\boldsymbol{X})dx$ index was also calculated, although it played no direct role in the estimation procedure.

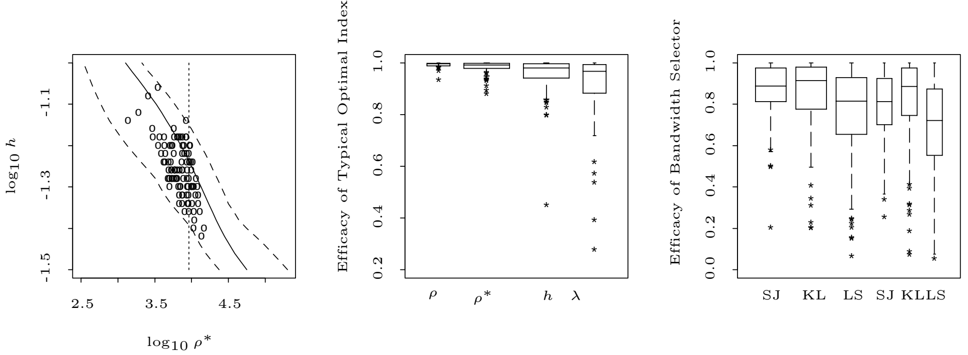

Figure 4. The h and ρ^* indices of kernel estimates and the performance of bandwidth selectors. Left: Empirical relation between h and ρ^* and the optimal indices. Center: L_{ISE} efficacies of typical optimal h and ρ^* indices. Right: L_{ISE} efficacies (fatter boxes) and L_{KL} efficacies (thinner boxes) of the Sheather-Jones, V_{KL}, and V_{LS} bandwidths.

The left frame of Figure 4 parallels the center frame of Figure 1, where $L_{ISE}(h|\boldsymbol{X})$ is used as the loss in defining the optimal estimates; the vertical dotted line marks the "true ρ^*" $\int \ddot{f}_1^2(x)dx = 10^{3.963}$. We pick the medians of the optimal indices as the fixed typical optimal indices, which give $h = 10^{-1.26}$ and $\rho^* = 10^{3.871}$. The L_{ISE} efficacies of these typical optimal indices, defined in the same manner as those in Section 2, are summarized in the center frame of Figure 4 as the fatter boxes. For comparison the boxes of the right frame of Figure 1 are superimposed as thinner boxes. Despite the fact that the indices come from two different problem settings, a rough ordering of the interpretability of these indices according to Heuristic 2 seems in place. The good news is that the h index appears more interpretable than the λ index for smoothing splines, but the bad news is that it runs only second to the ρ^* index under the same problem setting. The tighter scatter of the optimal ρ^* indices is curious, but we do not yet understand why it is the case.

The Associate Editor suggested that one might run the risk of comparing apples with oranges by superimposing the thinner boxes in the center frame of Figure 4, as regression and density estimation are different problem settings.

We followed up on that note by replacing the smoothing spline regression simulation results by some parallel smoothing spline density estimation simulation results, and the resulting plot is visually almost identical to the one presented. As noted above, the comparison is based on Heuristic 2, which seems somewhat independent of specific problem settings.

Conceptual validity aside, the plug-in method does provide us with some data-driven bandwidths, and we now compare its practical performance with that of the cross-validation method. In the course of the above simulation, we calculated $V_{LS}(h|\boldsymbol{X})$ and $V_{KL}(h|\boldsymbol{X})$ from which the cross-validation bandwidths were obtained. The Sheather-Jones bandwidths for the replicates were calculated using a FORTRAN routine kindly provided by Professor Sheather at `sish@agsm.unsw.edu.au`. The L_{ISE} efficacies of the Sheather-Jones (SJ) bandwidth, the V_{KL} (KL) bandwidth, and the V_{LS} (LS) bandwidth are summarized in the right frame of Figure 4 as the fatter boxes. For this test density, the losses $L_{ISE}(h|\boldsymbol{X})$ and $L_{KL}(h|\boldsymbol{X})$ were reasonably "parallel" to each other as functions of h given the data $\boldsymbol{X}$. In spite of its being designed primarily after the L_{KL} loss, V_{KL} has fared well with the Sheather-Jones procedure on the ground of L_{ISE} in the simulation. The performance of V_{LS} is clearly inferior, indicating that it is not too reliable an estimate of L_{ISE} for the purpose. The L_{KL} efficacies of the three procedures are summarized in the same frame via the thinner boxes.

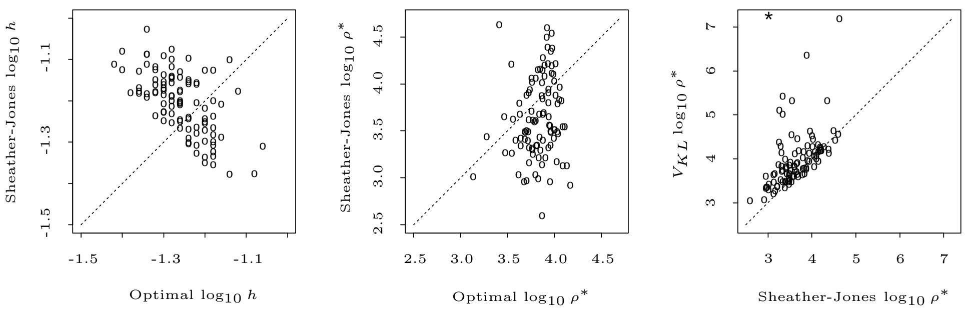

Figure 5. Some features of Sheather-Jones procedure. Left: Sheather-Jones h versus optimal h. Center: Sheather-Jones ρ^* versus optimal ρ^*. Right: V_{KL} ρ^* versus Sheather-Jones ρ^*.

Some features of the Sheather-Jones procedure are depicted in Figure 5. From the left frame, one can see that the famous negative correlation in the h index is no monopoly of the cross-validation method. In the center frame, no obvious correlation is found between the optimal ρ^* and the Sheather-Jones ρ^*. Besides the negative correlation, another prominent feature of the left frame is that the scatter of the optimal indices across-replicate is almost the same as that of the empirical indices across-replicate. In view of Heuristic 2, either the scatter is within reasonable natural fluctuation (but then the empirical procedure would be performing too well) or the indices of different replicates do not compare with each other. Although very good indeed, the practical performance of the Sheather-Jones procedure is still far from perfect, so the plot seems to serve as evidence against the across-replicate interpretability of the h index. The right frame compares the Sheather-Jones ρ^* with the V_{KL} ρ^*, where Sheather-Jones' preference for smoother estimates is evident.

Caution 5. *The kernel bandwidth h does not seem to be interpretable across-replicate.*

Caution 6. *Negative correlation exists between the optimal bandwidth and the plug-in bandwidth.*

To perceive how far the Sheather-Jones procedure and V_{KL} can depart from each other, we chose to take a closer look at the replicate plotted as the star in the right frame of Figure 5. Plotted in Figure 6 are the data in a finely binned histogram, the test density as a smooth dotted line, the Sheather-Jones estimate as a dashed line, and the V_{KL} estimate as a solid line. It is clear that V_{KL} was trying too hard to adapt to the data, perhaps somehow "deceived" by the data, whereas Sheather-Jones played rather conservatively. This replicate happened to record the lowest L_{ISE} efficacy of .182 for the V_{KL} procedure.

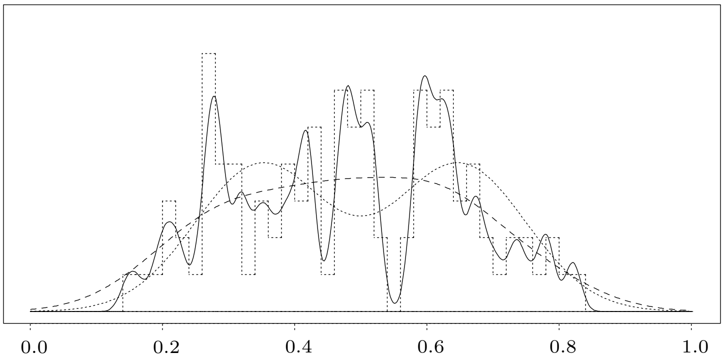

Figure 6. An extreme contrast between the Sheather-Jones estimate and the V_{KL} estimate. The histogram is finely binned data, the dotted line the test density, the dashed line the Sheather-Jones estimate, and the solid line the V_{KL} estimate.

Besides $f_1(x)$, we also carried out parallel simulations using three other test densities: $f_2(x)$ as the half-half mixture of $N(.5,(1/6)^2)$ and $N(.5,(1/6\sqrt{10})^2)$; $f_3(x)$ as the half-half mixture of $N(.5,(1/6)^2)$ and $N(.5,(1/60)^2)$; and $f_4(x)$ as the third-third-third mixture of $N(.3,(1/14)^2)$, $N(.7,(1/14)^2)$, and $N(.5,(1/7)^2)$; $f_2(x)$ and $f_3(x)$ are actually the shifted and scaled versions of the $f_2(x)$ and $f_3(x)$ test densities of Sheather and Jones (1991). The counterparts of the right frame of Figure 4 in the parallel simulations are included in Figure 7. The minimizers of $L_{ISE}(h|\boldsymbol{X})$ and $L_{KL}(h|\boldsymbol{X})$ were far apart for $f_2(x)$ and $f_3(x)$ but close to each other for $f_4(x)$, as reflected in Figure 7.

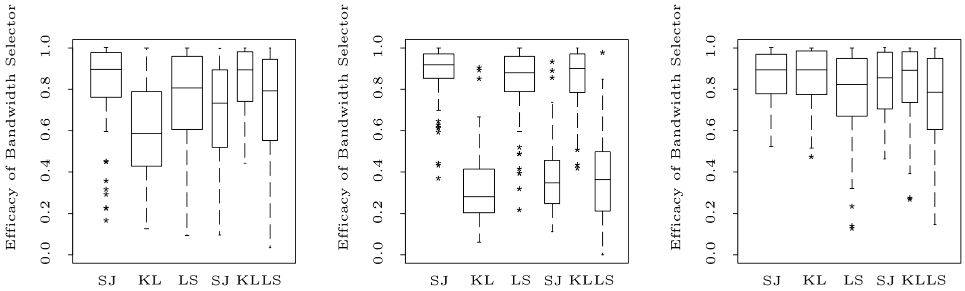

Figure 7. L_{ISE} efficacies (fatter boxes) and L_{KL} efficacies (thinner boxes) of the Sheather-Jones, V_{KL}, and V_{LS} bandwidths in parallel simulations. From left to right: $f_2(x)$, $f_3(x)$, and $f_4(x)$.

V_{LS} again appears inferior, but the unsatisfactory performance of a single score does not necessarily discredit the whole methodology of cross-validation, just as a case in which the plug-in method performed poorly would not be sufficient grounds for the rejection of the whole plug-in methodology. Comparing V_{KL} against Sheather-Jones, we see that they are winners respectively in their own games. The qualitative performance of V_{KL} in terms of L_{KL} and that of Sheather-Jones in terms of L_{ISE} appear comparable.

Following a suggestion by the Associate Editor, the relative efficacy of V_{KL} over Sheather-Jones, defined by the ratio of the loss (L_{ISE} or L_{KL}) of Sheather-Jones over that of V_{KL}, is summarized in Figure 8 for the reported simulations involving test densities $f_1(x)$, $f_2(x)$, $f_3(x)$, and $f_4(x)$. The findings are consistent with those from Figures 4 and 7 summarized above.

As loss function for density estimation, I find little to like about in L_{ISE} other than its tractability in kernel estimation: it is not invariant to base measure substitution, which is external to a probability distribution, and it assigns unduly heavy weight where information from data is scarce. In contrast, L_{KL} is among intrinsic measures free of these problems. Of course this is more a matter of personal preference, but L_{KL} is at least as defensible as L_{ISE}, if not more. Given the adequate performance of V_{KL} in terms of L_{KL}, I see no reason to rate V_{KL} as inferior to Sheather-Jones operationally.

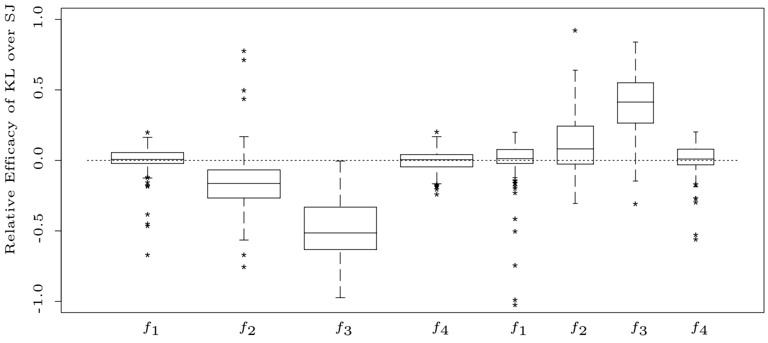

Figure 8. Relative efficacy of V_{KL} over Sheather-Jones in L_{ISE} (fatter boxes) and L_{KL} (thinner boxes). Vertical axis is in $\log_{10}$ scale

From the empirical evidences we have seen so far, it seems fair to say that the h index is not that interpretable across-replicate as is stated in Caution 5. Given this, the validity of concepts based on $EL_{ISE}(h|\boldsymbol{X})$ appears questionable. Since the plug-in method is derived from the asymptotic expansion of $EL_{ISE}(h|\boldsymbol{X})$, and the minimizer of $EL_{ISE}(h|\boldsymbol{X})$ plays an important role in the published theory, the relevance of the derivation and the theory of the plug-in method to the practical performance of the Sheather-Jones procedure is not clear. One possible way to check such relevance in simulation is to try to use *different* replicates to estimate different terms in the asymptotic expansion of $EL_{ISE}(h|\boldsymbol{X})$: if the performance of the method changes significantly after such a substitution, then the original procedure is in fact data-specific.

Caution 7. *The published theory for plug-in method may not have much to do with its practical performance.*

Reflecting different philosophies of smoothing parameter selection are the different performance measures used in our simulations and in the simulations of Sheather and Jones (1991) and Jones, Marron and Sheather (1992). We calculate $\min_h L(h|\boldsymbol{X})/L(h_*|\boldsymbol{X})$ where h_* is the empirical bandwidth under evaluation, which compares the *actual* performance of the estimate one will be using in practice with that of the best possible estimate given the data. Sheather and Jones (1991) and Jones, Marron and Sheather (1992) calculate $\min_h EL(h|\boldsymbol{X})/EL(h_*|\boldsymbol{X})$, which compares the *would-be* average performance of h_* if it were used for all possible but unobserved replicates from the same stochastic structure with the best possible average performance of all h. Even if the h index were perfectly interpretable across-replicate, the data-specific version would still carry more practical meaning than the expected version.

The discussion in the previous paragraph is related to the ISE versus MISE controversy, or loss versus risk, in the literature; related references may be found in Jones, Marron and Sheather (1992). We note that even people advocating risk (MISE) do not deny the conceptual appeal of working with the loss (ISE). They argue however that a purely data-based procedure does not have the needed information to behave well in terms of the loss. With results such as that of Li (1986) in mind (see (5) in Section 3), we have to disagree. On a more fundamental level, with the across-replicate interpretability of h in jeopardy, the very conceptual validity of the h-indexed risk is in question.

With comparable practical performance (between V_{KL} and Sheather-Jones in different terms), the cross-validation method enjoys the conceptual clarity, the computational simplicity, and the ready extensibility to more complicated settings such as multivariate smoothing. Multivariate nonparametric regression with cross-validation smoothing parameters has been in practical use for several years (cf. Wahba (1990)). Parallel developments in density and hazard estimation can be found in some recent work of the author (cf. Gu (1993, 1998)). For the plug-in method, conceptual validity aside, the choice of the asymptotic expansion and the estimation of functionals of the derivatives seem challenging enough beyond a dimension of one or two, yet its potential benefit remains questionable.

5. Summary

In this article, we have attempted to discuss a few conceptual issues related to smoothing parameter selection in nonparametric function estimation. The central idea is to identify, if possible, the *model* behind the estimate, so that the estimate can be viewed as a compromise between the model and the data.

For a nonparametric procedure yielding a continuum of possible estimates, there may or may not exist the luxury of explicit model characterization as in (3). Nevertheless, the nature of the model index used in smoothing parameter selection may imply some dos and don’ts in its theory and practice. It is advisable to carefully examine a working index before loading too much on it.

By empirically examining the interpretability of the λ index for smoothing splines and the h index for the kernel estimates, we find that these default indices are not so interpretable across-replicate. Consequently, a few popular concepts and intuition seem to be in jeopardy: the risk functions indexed by λ or h are somewhat mixtures of apples and oranges. Linear smoothers, or smoothing splines with a fixed λ or kernel estimates with a fixed h, are not quite estimators in the classical sense. The famous negative correlation seems to have little to do with the statistical performance of the bandwidth selector involved.

Working with indices not interpretable across-replicate, one needs to exercise caution to avoid possible conceptual pitfalls in smoothing parameter selection.

Our advice is to use a data-specific criterion to target a data-specific loss function. Among the available methods with comparable practical performance, we favor the cross-validation method for its clarity, simplicity, and extensibility.

Acknowledgements

This research was supported by National Science Foundation under Grant DMS-9301511. The author is grateful to Julian Faraway, Xiaotong Shen, Grace Wahba, and Yuedong Wang for comments on various drafts.

References

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17**, 453-555.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377-403.

Gu, C. (1990). Adaptive spline smoothing in non Gaussian regression models. *J. Amer. Statist. Assoc.* **85**, 801-807.

Gu, C. (1992). Cross-validating non Gaussian data. *J. Comput. Graph. Statist.* **1**, 169-179.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm. *J. Amer. Statist. Assoc.* **88**, 495-504.

Gu, C. (1998). Structural multivariate function estimation: Some automatic density and hazard estimates. *Statist. Sinica* **8**, 317-335.

Hall, P. and Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection (with discussion). *J. Roy. Statist. Soc. Ser. B* **54**, 475-530.

Jones, M. C., Marron, J. S. and Sheather, S. J. (1992). Progress in data-based bandwidth selection for kernel density estimation. Mimeo Series #2088, University of North Carolina, Dept. of Statistics.

Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14**, 1101-1112.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9**, 65-78.

Schoenberg, I. J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci.* **52**, 947-950.

Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation, *J. Amer. Statist. Assoc.* **82**, 1131-1146.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation, *J. Roy. Statist. Soc. Ser. B* **53**, 683-690.

Wahba, G. (1990). *Spline Models for Observational Data.* CBMS–NSF Regional Conference Series in Applied Mathematics, Vol. 59, SIAM.

Department of Statistics, Purdue University, West Lafayette, IN 47907, U.S.A.

E-mail: chong@stat.purdue.edu

(Received April 1996; accepted March 1998)

COMMENT

M. C. Jones

The Open University, U.K.

Overview

There is a novel and intriguing observation at the heart of this paper. In a spline smoothing context, the smoothing parameter λ and the constraint parameter ρ have a data–dependent connection; the summed squared error optimal smoothing parameter varies considerably with the data (this is known) but the "effective constraint" associated with that smoothing parameter is very much more stable (this is new). So what novel consequences does this apparently exciting observation have for practice? Little or none, I fear. Neither the author nor I have managed to find anything new and worthwhile to do with it (perhaps other discussants have). Instead, the paper really pontificates on two topics. The first is the old debate about whether one should choose smoothing parameters to minimise data–dependent losses (such as integrated squared error, ISE) or averaged risks (such as mean integrated squared error, MISE). The author seems not to have appreciated a number of papers on this topic in the kernel density estimation literature; the current discussion seems to me to add little. Secondly, the paper descends into a polemic which I paraphrase as "cross–validation at all costs". I shall discuss each of these at greater length below, but let me try to start on the more positive side by describing an (unsuccessful) attempt at developing the author's spline smoothing insight for kernel density estimation.

Effective Constraint for Kernel Density Estimation?

I have tried hard, but ultimately failed, to emulate the explicit effective constraint work on spline smoothing in the general kernel density estimation context. The key, it seems to me, should lie in the excellent but little known work of Terrell (1990) who identifies a "penalised least squares" problem to which the kernel density estimate — at least for many kernels — is the exact solution. In general, this takes the form

$$f_h(x|\mathbf{X}) = \operatorname{argmin}_f \Big\{ \int f^2 - 2n^{-1} \sum_{i=1}^{n} f(X_i) + R_h(f) \Big\},$$

where $R_h(f)$ is a roughness penalty (with smoothing parameter, the bandwidth h, subsumed).

Different roughness penalties are associated with different kernels. The normal kernel is associated with the roughness penalty

$$\sum_{j=1}^{\infty} \frac{(-1)^j}{j!2^j} \int (f^{(j)})^2(x)dx = \frac{1}{2\pi} \int |\phi(t)|^2 (e^{-h^2t^2/2} - 1)dt,$$

where ϕ is the Fourier transform of f. However, I could not get the behaviour of $\log h$ as a function of the log of this penalty to emulate that in the left–hand frame of Figure 1. I also tried to match the integrated squared second derivative penalty function with use of the usual spline–equivalent kernel in density estimation (e.g. Silverman (1984)), but got no more than the merest hint of the behaviour observed in the (exact) spline smoothing case. This is all rather disappointing. Perhaps my programming is at fault.

What Practical Help Could This Be?

In the spline smoothing context (and in the kernel density estimation context if a general solution to the above could be found), the question remains, what practical use can be made of this? My feeling is little or nothing. It is attractive to think that all one has to do is to direct initial estimation efforts at the roughness penalty rather than λ. Remember that one is interested in the penalty function at the optimal value of the smoothing parameter for the given dataset. The reasonable hope is that this is approximately the same as the penalty function at the "true model" (I shall use the latter nomenclature without spelling out the usual caveats). One might then think of reframing existing work on such estimation problems (e.g. Hall and Marron (1987), Jones and Sheather (1991)) in terms of a further effective constraint instead of a bandwidth, or in the first instance apply such theory as it is. But I am not confident of a positive outcome. In practice, one has just the one dataset, with its own idiosyncracies relative to the true model (see my next section), so disregarding these idiosyncracies to get a handle on ρ must be just as hard as getting a good handle on λ (if not harder, because the penalty functionals are harder to estimate than the original curve). This, like much of the paper, leads me inexorably back to a number of well known properties and arguments of and about smoothing which I come to next, starting with that very problem of estimating λ appropriate to a particular dataset.

"Data–Specific Optimal λ" and ISE versus MISE

"The objective of model selection is thus to locate a *data–specific* optimal λ". This view is at first sight very laudable, but leads to considerable conceptual and practical difficulties, as is already well known in the context of kernel density estimation (and which is too briefly dismissed by Dr. Gu towards the end of Section 4).

The big point to make here is that achieving a "data–specific optimal λ" is an inherently difficult problem. There can be no *very* successful practical solution, at least of a purely data–based type. Consider, for example, the density estimation case in which your data truly come from a normal distribution. Few if any such data samples will be "very normal" in appearance; most will exhibit some kind of (usually small) apparent deviation from normality. Suppose one particular sample "looks a little bimodal". Then for that sample a bandwidth selector is likely to choose a fairly small h, yet if the information were available that the normal density were the true model, the best h choice would be large to smooth out the inappropriate modes. Conversely, if the truth were (not especially well separated) bimodal, many samples would not be especially clearly bimodal. Data–based selectors might well then choose large h signifying a unimodal distribution, while knowledge of the truth would suggest a small best h choice, to retain any bimodality that is there. Scott (1988) gives a nice example involving under/overdispersion relative to the truth. And it's easy to think of other situations like this, in regression as well.

So, a data–specific optimal λ or h requires knowledge of the right answer to be drawn away from reflecting what the data looks like! This is, of course, the source of the well known negative correlation between the "optimal" and cross–validation smoothing parameters. It is also the source of negative correlation between most data–based bandwidth selectors and the "optimal" (Caution 6 is no surprise). While one or two authors may in the past have "charged cross–validation for performing counter–intuitively" because of this, it is no longer a big negative issue: this kind of behaviour, common to most approaches, and the reasons for it, are by now well appreciated.

The same phenomenon shows up in further well understood smoothing theory. Hall and Marron (1991) investigated how well you can possibly do when estimating $\hat{h}_0$, the minimiser of the ISE; this is precisely the kind of target Dr. Gu is arguing for. The answer is a best relative rate of convergence (of any $\hat{h}$ to $\hat{h}_0$) of $O(n^{-1/10})$. This awfully slow rate — the inherent difficulty of the problem — contrasts with the order $n^{-1/2}$ rate available for estimating h_0, the minimiser of the MISE. (Note that these results are about how quickly $\hat{h}$'s approach $\hat{h}_0$ (assuming a certain amount of smoothness) and not about the lesser question of whether they tend to $\hat{h}_0$ at all (though perhaps with less smoothness assumed), which is the realm of Li (1986) and other papers.) Reasoning and results like these fed various papers on the "ISE/MISE controversy", e.g. Jones (1991), Hall and Johnstone (1992) and Grund, Hall and Marron (1994).

My current view remains that estimating $\hat{h}_0$ is not really on. I guess that what one feels one might be doing instead is choosing an h that is appropriate for datasets that are typical of the true model, and using it also for the atypical datasets, when they should at least still show some sign of under/oversmoothing as indicated by the true model. At present, our best surrogate for this would still seem to be to target h_0. Whether Dr. Gu's work can give us any new handle on this, I have my doubts.

I should add that if you do continue to try to track $\hat{h}_0$, how well various kernel density estimation bandwidth selectors perform in this regard is also already well known. See Hall and Johnstone (1992) and, particularly, Jones and Kappenman (1992). In the latter paper, we show (least squares) cross–validation performing poorly relative to a number of other choices, including those methods, such as plug–in, attuned to estimating h_0. Hall and Johnstone (1992) then provided the first method to (theoretically) achieve the best possible performance, and in Jones (1998), I believe I've stumbled across another (but not developed it for practice).

Kullback–Leibler and All That

Dr. Gu also recognises the inadequacy of least squares cross–validation in his Section 4, but ends up advocating likelihood cross–validation (KL) instead. It is interesting to note that in the kernel density estimation literature it is more accurately characterised (see above the definition of V_{KL}) as an *un*popular criterion! The question raised is, should KL possibly be rehabilitated?

KL is unpopular because of its poor ISE performance. Concentrating on ISE assessment for the moment, the reason is clear (and well known). KL takes much more notice of the tails than do (M)ISE–based methods. This is best seen by Taylor series approximation of L_{KL} by the weighted ISE

$$\int (1/f(x))(f_h(x|\mathbf{X}) - f(x))^2 dx.$$

Hall (1987) theoretically developed some of the consequences of this, and was influential in tolling a (partial?) death knell for KL. Sure enough, it is for the heavy tailed densities f_2 and f_3 that KL fails badly in ISE terms. Given the emphasis on tail behaviour, I suspect that KL (badly) oversmooths in these cases. Its undersmoothing in Figure 5 for the bimodal test density would appear to reflect (relatively) short tails here; I wonder whether, even when in this case its ISE performance is not terribly bad, most of the actual density estimates are a little like that in Figure 6 but less extremely so (i.e. 'wobbly' around something like the true model).

I do agree with Dr. Gu that ISE is not a particularly attractive loss measure except for its analytical tractability. Others have worried about this, motivated by things like the end of the previous paragraph: ISE will tend to score wobbles about the true model as better, for example, than excellently reproduced shapes whose location is just a little bit out. Of course L_1 error is often proposed, but this makes little difference in these terms. The only serious alternative of which I'm aware is some interesting work of Marron and Tsybakov (1995).

KL's tail emphasis does not appeal to me, except one has in the back of one's mind that this is the very thing that drives maximum likelihood's full efficiency in parametric fitting! But full efficiency comes at the expense of zero robustness. In Basu, Harris, Hjort and Jones (1998), we explore, in a parametric setting, the minimisation of a class of divergences bridging the gap between KL and LS. For $0 < \alpha \leq 1$, this distance is

$$\alpha^{-1} \int \Big\{ \alpha f_h^{1+\alpha}(x|\mathbf{X}) - (1+\alpha) f(x) f_h^{\alpha}(x|\mathbf{X}) + f^{1+\alpha}(x) \Big\} dx;$$

it is easily seen that $\alpha = 1$ yields LS and $\alpha \to 0$ yields KL. We find that small values of α afford good robustness and high efficiency.

If one wished to pursue this for density estimation, it is immediately clear how a cross–validatory version would work: minimise

$$\int f_h^{1+\alpha}(x|\mathbf{X}) dx - (1+\alpha^{-1}) n^{-1} \sum_{i=1}^{n} f_h^{\alpha}(X_i|\mathbf{X}_{(i)}),$$

which covers V_{LS} when $\alpha = 1$ and (essentially) V_{KL} when $\alpha \to 0$.

I agree with Dr. Gu that a very attractive property of cross–validation is its "ready extensibility to more complicated settings" (I am not sure I rate its "conceptual clarity" above that of other approaches, and "computational simplicity" is overstressed when you need to optimise a quite possibly multimodal function). Others have stressed this too. But this doesn't mean that one has to stick with existing forms of cross–validation. For a better, extensible, cross–validation–like method (in another direction again), for example, see Hurvich, Simonoff and Tsai (1998).

Conclusions

The point starting the paper about the stability of the effective constraint in spline smoothing is novel and intriguing. However, my initial enthusiasm for this has waned as I (and the author) have failed to get anything novel out of it, and we have fallen back on what turns out actually to be largely a rehash of known properties and arguments (at least in the kernel density estimation setting). The author's enthusiasm for cross–validation drives the remainder of the paper, but I feel that the best path is not taken. I am not convinced that I should reinstate likelihood cross–validation in my list of good(ish) bandwidth selection methods, but have suggested some alternative avenues for exploration. I am grateful for the opportunity to contribute to this discussion.

The Open University, Department of Statistics, Walton Hall, Milton Keynes MK7 6AA, United Kingdom.

E-mail: M.C.Jones@open.ac.uk

COMMENT

David W. Scott

Rice University

Professor Gu has written a thought-provoking article about some basic issues in smoothing. The author's experience and opinions are encapsulated in two heuristics and seven cautions. The concept of a "natural model" behind nonparametric estimation schemes is attractive, and the author argues that the smoothing parameterization in a natural model will "largely remain invariant over replicated data". An alternative model, even though in one-to-one transformation of smoothing parameters, is claimed to be conceptually "incorrect" and not "interpretable". Such a surprising state of affairs deserves closer scrutiny.

The article uses nonparametric spline regression to motivate the heuristics, and density estimation to evaluate them. For penalized least squares regression, one may specify the penalty weight, λ, or a bound, ρ, on the penalty functional $\int_0^1 f''(x)^2 dx$. For a particular data set, there is a one-to-one relationship between λ and ρ, and one may plot $\lambda = \lambda(\rho)$, as in the author's Figure 1. Also in that figure, the author shows that the values of λ and ρ minimizing the actual penalized least squares criterion,

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \lambda \int_0^1 f''(x)^2 dx, \tag{1}$$

across replicates are much more variable in the λ-space than in the ρ-space. How should we compare λ and ρ? A dimensional analysis is instructive. Suppose we are predicting household potato consumption (in pounds or #'s) as a function of household income (in $'s). Then the units of $f''(x)$ are measured in $\#/\$^2$; hence, the units of $\int_0^1 f''(x)^2 dx$ (as well as those of ρ) are $\#^2/\$^3$. Similarly, since the units of the first term in (1) are $\$^2$, then the units of λ must be $\$^5/\#^2$, so that the two terms in equation (1) are commensurate. Even if the vertical units were also in $, then λ and ρ are measured in $\$^3$ and $1/\$$, respectively. Thus λ and ρ measure quite different quantities, and comparisons between them may be as apples to oranges. In particular, other aspect ratios may be more appropriate than the one chosen in the middle frame of the author's Figure 1.

In my own work on biased cross-validation (BCV, an early example of a plug-in method), I was impressed that the variation of $\hat{h}$ was reduced from that of least-squares cross-validation (LSCV), but never drew strong conclusions of uniform superiority (Scott and Terrell (1987)). BCV is just very different than LSCV and lives at a different bias/variance point. I was more interested in cases where the two algorithms differed and why. The data presented in the final frame of the author's Figure 1 can be understood in less profound terms. Certainly ρ appears less variable than λ in this example, but given the difference in units more caution is warranted with any interpretation. If a cross-validation method exists that targets ρ directly, might it not be more susceptible to systematic bias than methods targeting λ?

Along the way, Professor Gu criticizes squared error, a popular thing to do. A number of misleading conclusions can be drawn because of particular properties of L_2 error. I call attention to my favorite "hard" density shown in Figure 1, which is the mixture

$$\frac{3}{4}\,\phi(x|0,1)\;+\;\frac{1}{4}\,\phi(x|3,(\frac{1}{3})^2). \tag{2}$$

The integrated squared bias of a positive kernel estimator is asymptotically controlled by $\int f''(x)^2dx$. Recall that $\int \phi''(x|\mu,\sigma^2)^2dx = 3/8(\pi)^{1/2}\sigma^5$. If we treat the two components in Equation (2) as nonoverlapping for simplicity, then the relative contributions to the total integrated squared bias are easily seen to be in the ratio 1:27, respectively! Thus any "good" (fixed-bandwidth) estimator of data from this mixture should essentially ignore the left component in favor of the narrower component on the right. This common but largely ignored property of the L_2 criterion means that in any data set with a tight cluster (say by chance alone), $\hat{h}$ will be too small and $\hat{f}$ likely to be significantly undersmoothed over the entire domain of f. This is not so much a deficiency of L_2-error as of fixed-bandwidth estimators. Thus most multi-component simulations are not illuminating because the contributions to error from non-dominant features only affect the second or third significant digit in the L_2 error. Locally adaptive estimates can significantly improve the visual appearance, but usually provide apparently insignificant decrease in the total L_2 error because of this dominance feature. (It would be better to compute and analyze the local spatial components of the L_2 error separately.) My concern is that many of the author's simulations are less illuminating than hoped, especially at these small sample sizes.

Many workers who visually prefer nonparametric estimates of the density in Figure 1 that pay more attention to the left mode are essentially rejecting L_2 error. Plug-in bandwidth algorithms that favor such estimates are indeed "oversmoothing" with respect to L_2 error and are ignoring the correct MISE criterion. The author is correct in pointing out the deficiencies of plug-in methods, but that is due to the difficulty of doing locally adaptive plug-in estimation, not the

L_2 error itself. Switching to KL or L_1 *per se* does not really change complaints that should be traced to the use of fixed rather than locally adaptive estimators.

Professor Gu also discounts the importance of the negative correlation between $\hat{h}$ and ISE-minimizing bandwidths in the density estimation setting as observed in Scott and Terrell (1987) — c.f. Figure 2 in the current paper. I wish to argue that ISE-minimizing bandwidths are not the most appropriate target, and that the negative correlation observation supports this claim. Given obvious problems with multi-component densities, let us stick with samples from a single normal density and examine the ISE of a kernel estimator as the bandwidth h varies. Clearly the ISE is a function of the sample moments of the data. If we focus on the first two moments, we see that if $\bar{x}$ is quite different from the true μ then no particular choice of h can help, as $\int x\,\hat{f}(x)\,dx = \bar{x}$ for all h (Scott (1992)). However, the bandwidth h does directly affect the variance of the kernel estimator $\hat{f}$. Suppose the sample variance is much less than σ^2. Then the ISE for the particular sample can usually be reduced by inflating the smoothing parameter h and flattening the estimator. Conversely, reducing h in samples where the sample variance is too large can reduce ISE (although the effect is less pronounced). To the extent that ISE-minimizing smoothing parameters improve error by "going against the sample variance," I find them to be an inappropriate target. The knowledge that the sample variance is too large or too small cannot be known from the data alone. That is why many workers prefer MISE to ISE, although the author certainly makes a case that MISE is not perfect. The negative correlation between bandwidths is a consequence of the negative correlation between the ISE-minimizing bandwidths and the sample standard deviation. (In mixture densities, this holds if we focus locally on the dominating component.) Finally, I wonder if there isn't a negative correlation in the ρ plot in the final frame of the author's Figure 2, where scaling choices obscure the scatter?

In summary, while I think Heuristic 1 is useful to study, I think the evidence leading to Heuristic 2 is still weak. Further simulations with very carefully constructed test cases might indeed bolster the author's beliefs. If the author wishes to pursue his empirical investigation further, he might consider using the penalized least-squares linear density estimators devised by Terrell (1993). These parallel the penalized regression formulation very well. Terrell shows that many nonparametric density estimators, including the kernel, can be realized by appropriate choice of penalty functional.

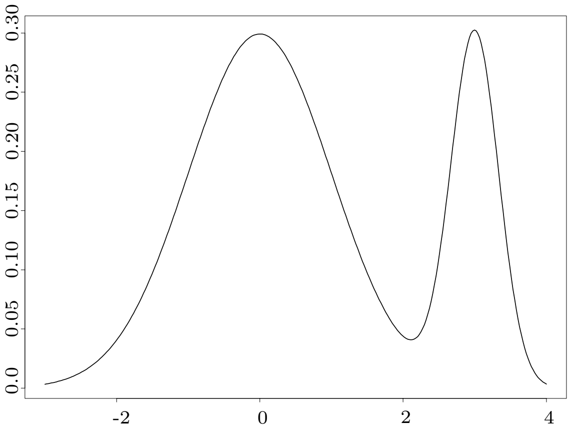

Figure 1. Mixture of two normals: $0.75\,\phi(x|0,1)+0.25\,\phi(x|3,3^{-2})$.

Acknowledgement

This research was supported in part by NSF grant DMS-9626187.

Department of Statistics MS-138, Rice University, 6100 Main Street, Houston, TX 77005-1892, U.S.A.
E-mail: scottdw@stat.rice.edu

COMMENT

Grace Wahba

University of Wisconsin-Madison

1. Thanks

We owe a strong vote of thanks for the author's careful examination of the 'well publicized' negative correlation between the optimal λ (λ^{opt}) and the GCV λ ($\hat{\lambda}$), as observed in the middle panel of Gu's Figure 2. (I am used to using $\hat{\lambda}$ for the GCV estimate of λ and will continue that here, this is $\lambda*$ in Gu.) The argument that it is an inherent property of the choice of the model index (as shown, for example in the left panel of his Figure 5), and not the cross validation, is an important piece of information.

2. Related Work

Gu notes that with sample sizes of $n = 100$, GCV occasionally one gets undersmoothed estimates. If the GCV function has a positive derivative at $\lambda = 0$, then it will have a minimizer there and then the fitted curve will interpolate the data. The probability of this event has been studied in Wahba and Wang (1995) and is shown theoretically to decay exponentially rapidly with the sample size, under general conditions. The GCV function for a smoothing spline with distinct data points behaves as λ^2/λ^2 as λ tends to 0 so special search procedures are required in this case. However, the probability of this event is generally too small for it to be observed in 100 replications of the fitting which are based on a smooth curve with white Gaussian errors of reasonable variance and n as small as 100. A few percent of somewhat (as opposed to extremely) undersmoothed cases similar to those in Figure 3 of Gu appear to be typical in $n = 100$ simulations, at least from what I have observed in similar student simulations in my classes on Spline Models. However this effect also appears to shrink rapidly with sample size, as will be illustrated below, although no theory is offered here (see Li (1986)).

As Gu notes, GCV is being used with multiple smoothing parameters and complex multivariate models. One recent reference is Gong, Wahba, Johnson and Tribbia (1998) (preprint available via my home page, `http://www.stat.wisc.edu/~wahba -> TRLIST.`), where the randomized trace technique (see Girard (1995)) is used to avoid matrix decompositions in the calculation of $traceA(\lambda)$ in large data sets. There five smoothing, weighting and physical parameters are chosen by GCV in a nonparametric regression problem where a dynamical systems equation is included as a weak constraint. It is easy to unearth applications of GCV in various fields via web searches.

3. A Historical Note

Let $\log \lambda^{opt}$ be the Optimal $\log_{10} n * \lambda$ of the center panel of Gu's Figure 2 and let $\log \hat{\lambda}$ be the CV $\log_{10} n * \lambda$ of the same plot. In 1987 I gave a series of CBMS lectures at Ohio State University, which were expanded into Wahba (1990). At the lectures I announced that $\log \hat{\lambda}$ and $\log \lambda^{opt}$, exhibit a strong positive correlation. Iain Johnstone and David Scott immediately assured me that I had got it backwards, or, rather, upside down. The negative correlation, of which the center panel of Figure 2 is an example, was obviously known to them and others, and has been a subject of a lot of discussion since.

Upon being asked to comment on the present very insightful paper, I dug up Wahba and Wold (1975), upon which my comments back in 1987 had been based, to try and figure out where I had gone wrong. Figure I of that paper contains (among other things) four pairs of curves, each pair consisting of one `CV` plot and one `TR` plot. Each `CV` plot is a leaving out 10% cross validation plot as a function of the indexing parameter. GCV hadn't been invented yet, but the results could be expected to be similar on this example. Each `TR` plot is the true mean square error averaged over the data points also as a function of the indexing parameter. Just as I had remembered, the four minima of `CV` *vs* the corresponding minima of `TR` followed each other, and a recent crude hand plot of the four pairs of minima, as read off the plots in that paper, fell almost on a straight line through the origin with slope +1. But the 'indexing' parameter for plotting purposes was in fact not λ but k, where $k = k(\lambda, \sigma^2, Y_1, \ldots, Y_n)$ was defined as

$$k = \frac{1}{n}\sum_{i=1}^{n}(f_\lambda(x_i) - Y_i))^2/\sigma^2. \tag{1}$$

k is not of course a true indexing parameter because σ^2 is unknown, but the mean square residual $r(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(f_\lambda(x_i) - Y_i))^2$ is, since it is (under some assumptions of course) (along with ρ) a strictly monotone function of λ. We may examine the correlation properties of k for the purposes of studying the mean square residual as an indexing parameter, since σ^2 will be the same for all replications, and it will be convenient below to do so.

4. The Mean Square Residual as the Indexing Parameter

We decided to examine the behavior of k (equivalently, the mean square residual) as an indexing parameter, by running a few simulations and plotting $\hat{k}$ *vs* k^{opt}. For Figure 1 we simulated observations $Y_i = f(x_i) + \epsilon_i$ at equally spaced points $x_i \in [0, 1]$ using

$$f(x) = w_1\beta_{p_1,q_1}(x) + w_2\beta_{p_2,q_2}(x), \tag{2}$$

where $\beta_{p,q}$ is the β density with parameters p and q and $w_1 = .4, p_1 = 12, q_1 = 7$; $w_2 = .6, p_2 = 4, q_2 = 11$. f was Example II in Craven and Wahba (1979). The ϵ_i are pseudo-random i.i d. $\mathcal{N}(0, \sigma^2)$. In Panels (a)-(c) of Figure 1 we used $n = 100$, with $\sigma = .01, .1$, and 1.0 respectively. In Panel (d) we used $n = 400$ and $\sigma = .1$ We used the code `smooth.spline()` of Splus with `cv = F` (meaning that GCV is used to choose the smoothing parameter), and `allknots = T` (meaning that there is a knot at every data point, rather than an approximation), to obtain $\hat{\lambda}$. Once $\hat{\lambda}$ was found $\frac{1}{n}\sum_{i=1}^{n}(f_\lambda(x_i) - f(x_i))^2$ was searched as a function of $\log_{10}\lambda$ on a grid from $\log_{10}\hat{\lambda} - 4$ to $\log_{10}\hat{\lambda} + 4$ in increments of .05 to find the minimizer λ^{opt}. We were not able to ascertain if the Splus code has a special procedure for searching the GCV function as $\lambda \to 0$. However, it does have an error message ("smoothing parameter value too small or too large"). No error messages of this type were obtained in our simulations, and, since our n was large enough to assume that 0 was not the minimizer, we assume that it is not an issue in the present simulations. We also ran the example in Gu, with $n = 100, \sigma = 1$ and the plot strongly resembles Panel (c).

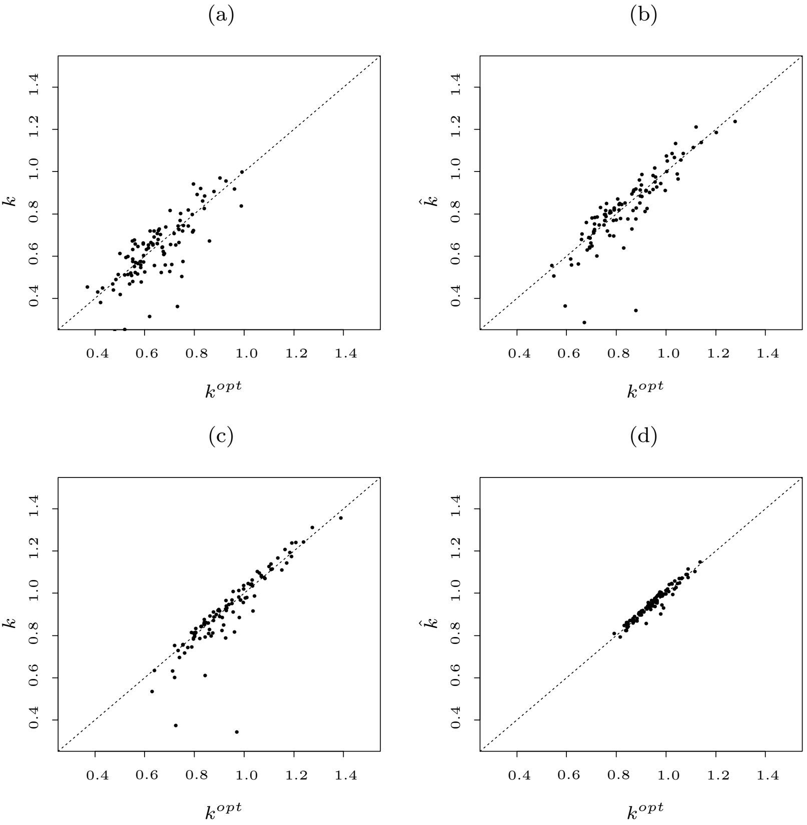

Figure 1. $\hat{k}$ *vs* k^{opt}. Panel (a): $n = 100, \sigma = .01$, Panel (b): $n = 100, \sigma = .1$, Panel (c): $n = 100, \sigma = 1.0$, Panel (d): $n = 400, \sigma = .1$.

The correlation properties of $\hat{k}$ and k^{opt} are fairly obvious and do not need a lot of comment, nor does the fact that the apparent undersmoothing of a few percent of the cases in the $n = 100$ examples is no longer evident in the $n = 400$ case. We note that the 'spread' in k^{opt} in Panels (a)-(c) is about .6, while the 'spread' in k^{opt} in Panel (d) has been halved, to about about .3. We think this is related to the fact that k^{opt} has been normalized by σ^2 whereas $\sum_{i=1}^{n} \epsilon_i^2/\sigma^2$ has mean 1 and standard deviation $(2/n)^{1/2}$. Thus the reduction in the 'spread' of k in Panel (d) could be explained by the fact that $(\frac{100}{400})^{1/2} = 1/2$. A propos of comparing simulations to theory, we note that the distribution of k^{opt} in Panels (a)-(c) appear to have the same 'spread' but there is a slight shift upward along the diagonal as σ^2 goes from .01 to 1.0. This is consistent with the old (Wahba 1975b) theoretical result (roughly, under some assumptions) that

k^{opt} is centered about some number that is (asymptotically) bounded above by $1-(c_m\frac{\|f^{(2m)}\|^2}{\sigma^2 n^{4m}})^{\frac{1}{4m+1}}$, where m refers to splines of degree $2m-1$, here $m=2$. To the extent that this bound is tight, it suggests that as either σ^2 or n increases, the center of the distribution of k^{opt} rises towards 1.

5. A Conjecture and Challenge

We begin this section by noting that $\log\rho(\lambda)$, the mean square residual $r(\lambda)$, the degrees of freedom for signal (defined as $traceA(\lambda)$ here, see Wahba(1983)), as well as any other strictly monotone function of λ (including some other definitions of degrees of freedom for signal) are all equivalent indexing parameters in the smoothing spline case.

In the smoothing spline case it is typically numerically easiest to compute with λ, whereas other indices may provide more intuition for examination and plotting purposes. Hence the correlation properties of the $\hat{}$ and *opt* versions of different indices are irrelevant, we certainly agree with Gu here. Comparing the right panel of Gu's Figure 2 and Figure 1 of these comments, it is clear that between the mean square residual r, $\log\lambda$, and $\log\rho$, $\log\rho$ is the index that comes closest to Gu's Heuristic 2, at least with respect to GCV. These plots show that the scatter in $\log\rho^{opt}$ is very small compared with the scatter in $\log\hat{\rho}$, while the scatter in $r^{opt}=r(\lambda^{opt})$ is similar to the scatter in $\hat{r}=r(\hat{\lambda})$ since they are highly positively correlated.

Nevertheless, if you want to see a plot like that in Panel (d), or even like those in the other panels, we will make a conjecture that this is a fairly general phenomenon.

We state our conjecture in very general terms, which include penalized log likelihood estimates and other regularized estimates widely used in applications. We let $x\in\mathcal{T}$, $f(x)$ a real valued link function on $\mathcal{T}$. Y_i is an observation from a distribution $\mathcal{F}_{f(x_i)}=\mathcal{F}_{f_{true}(x_i)}$ which depends on the parameter of interest $f_{true}(x_i)$, and possibly nuisance parameters independent of x_i. The Y_i are assumed independent. f_λ is the solution of a variational problem of the form: Find $f\in\mathcal{X}$, an appropriate class of functions, to minimize

$$L(Y,f)+\lambda J(f). \tag{3}$$

Here

$$L(Y,f)=\sum_{i=1}^{n}\ell(Y_i,f(x_i)), \tag{4}$$

where $\ell(y_i,f(x_i))$ is some pseudo distance or discrepancy measure between Y_i and the distribution $\mathcal{F}_{f(x_i)}$, and $J(f)$ penalizes complexity as defined in some reasonable manner. We assume that the variational problem (3) has a unique solution for all data sets of possible interest. We assume that $J(f_\lambda)$ is a strictly monotone decreasing function of λ and $L(Y, f_\lambda)$ is strictly monotone increasing in λ. Then λ, $J(\lambda)$ and $L(Y, f_\lambda)$ are equivalent indexing parameters. The 'target' is to choose λ to minimize

$$L(\mu_{true}, f_\lambda), \tag{5}$$

where $\mu_{true}(x_i)$ is a suitable 'prediction' of Y_i given $f_{true}(x_i)$. We let λ^{opt} be the minimizer of $L(\mu_{true}, f_\lambda)$. (Uniqueness is not actually guaranteed.) The usual examples are: $\mathcal{X}$ is a reproducing kernel Hilbert space and $J(f)$ is a squared norm or seminorm on $\mathcal{X}$. But it would be interesting to include cases where the penalty involves a Banach Space or Besov Space norm (see, for example, Chen, Donoho and Saunders (1995), Donoho, Johnstone, Kerkyacharian and Picard (1995), DeVore and Lucier (1992), Rudin, Osher and Fatemi (1992), Wahba (1975a)). We just need that point evaluations in $\mathcal{X}$ are appropriately bounded. The usual suspect for $L(Y, f)$ is the negative log likelihood, but robust functionals (Cox (1983), for example), quantile functionals (Koenker, Ng and Portnoy (1994), for example), and convex support vector machine functionals (Wahba (1997), available via my home page) would be of interest. If $L(Y, f)$ is the negative log likelihood for an exponential family then it is natural to take $\mu_{true} = EY|\mathcal{F}_{f_{true}}$ and then the 'target' will be the comparative Kullback Liebler distance between f_{true} and f_λ, taken over the x_i. With regard to support vector machines used for classification with binary data, if f is the log odds ratio for $Y_i = +1$ *vs* $Y_i = -1$, one example of potential interest (see Wahba (1998)) is $\ell(Y_i, f(x_i)) = [1 - Y_i f(x_i)]_+$ where $[\tau]_+ = \tau, \tau > 0, = 0$ otherwise.

For the conjecture about to be proposed, we probably need the existence of a 'Leaving Out One Lemma' for the variational problem (3). This Lemma says that if you leave out the kth data point, solve the variational problem (λ fixed), obtain $\mu_\lambda(x_k)$ as the prediction of Y_k given $f_\lambda(x_k)$, and then solve the original variational problem with Y_k replaced by $\mu_\lambda(x_k)$, you will get f_λ back again. A Leaving Out One Lemma for different versions of (3) may be found (aside from the original spline version in Craven and Wahba (1979)) in Xiang and Wahba (1996) and Wahba (1997). In the support vector machine case just given a definition of $\mu = \mu(f)$ which 'works' is $\mu(x) = 1$ if $f(x) > 1$, $\mu(x) = -1$ if $f(x) < -1$, and $\mu(x) = 0$ if $f(x) \in [-1, 1]$.

Here is the conjecture: Let $\hat{\lambda}$ be a suitable cross validation based estimate of λ. (Candidates aside from the GCV in the non-Gaussian case include the GACV in Xiang and Wahba (1996) and Wahba (1997)). Let λ^{opt} be the minimizer of $L(\mu_{true}, f_\lambda)$. Let $\hat{r} = r(\hat{\lambda}) = L(Y, f_{\hat{\lambda}})$, $r^{opt} = r(\lambda^{opt}) = L(Y, f_{\lambda^{opt}})$. The conjecture is that under some general circumstances (to be found), $\hat{r}$ and r^{opt} are (strongly) positively correlated. The challenge is: What are the most general circumstances?

Acknowledgement

I would like to thank Alan Chiang for carrying out the simulations. Research sponsored in part by NSF under Grant DMS-9704758 and in part by NEI under Grant R01 EY09946.

University of Wisconsin-Madison, 1210 W. Dayton Street, Madison, WI 53706-1685, U.S.A.
E-mail: wahba@stat.wisc.edu

REJOINDER

Chong Gu

I am grateful to Professors Jones, Scott, and Wahba for an insightful and stimulating discussion. The article itself consists of merely an interesting observation plus some of my readings of its logical consequences, but the ramifications appear to extend further as attested to by the many important issues raised by the discussants. As the comments from the discussants are more or less "orthogonal" to one another, I shall try to attend to the concerns of each discussant individually in the sections to follow.

Jones

I agree with Professor Jones that the observation does not seem to lead directly to the development of new tools for smoothing parameter selection. To be precise, I am *not* advocating any attempt to estimate the "true ρ", which, as Professor Jones pointed out, is "just as hard as getting a handle on λ, if not harder". Nevertheless, the observation does help some, this author included, to better understand the existing tools, and to avoid possible conceptual pitfalls in the use or development of new tools. For example, working with a model index that is not interpretable across-replicate, I would avoid resampling methods for smoothing parameter selection. The practical help of the observation is to caution people that "*the nature of the model index used in smoothing parameter selection may imply some dos and don'ts in its theory and practice*", to quote from Section 5 of the article.

I am sorry that Professor Jones' effort in emulating my simulations was not very successful. For the record, let me report that my observation illustrated in Figure 1 of the article was originally made in a density estimation setting, where the density $f(x) = e^{g(x)}/\int_0^1 e^g$ on $[0,1]$ was estimated by the minimizer of a penalized likelihood functional

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{g(X_i) - \log\int_0^1 e^g\Big\} + \frac{\lambda}{2}\int_0^1 \ddot{g}^2;$$

see Gu and Qiu (1993) and Gu (1993) for further details concerning this estimate. See also the Wahba section below for some empirical results. In general one can establish the equivalence of a penalized minimization problem

$$\min\ A(g) + \lambda J(g),$$

and a constrained minimization problem

$$\min\ A(g) \quad s.t.\ J(g) \le \rho,$$

when $A(g)$ is convex and Frechet differentiable and $J(g)$ is quadratic, for g *in a linear space*. I am not sure that such mathematical equivalence necessarily extends to the penalized solution of Terrell (1990), as cited by Professor Jones, where the constraints $f > 0$ and $\int f = 1$ are lurking in the background. I probably would double check the theory before questioning the programming.

Concerning the data-specific optimal bandwidth, I would like to clarify that there are two ways to measure how close one can get to it. One way is to calculate the difference $\hat{h} - \hat{h}_0$, and the other way is to calculate the ratio $L(\hat{h})/L(\hat{h}_0)$. The bandwidth only assumes its meaning through the estimation procedure but typically has no place in the underlying stochastic structure, so the estimation of $\hat{h}_0$ has little direct statistical meaning except through the minimum loss $L(\hat{h}_0)$. An "awful" estimate $\hat{h}$ of $\hat{h}_0$ in terms of $\hat{h} - \hat{h}_0$ thus can in fact be a very good one if $L(\hat{h})/L(\hat{h}_0)$ is close to 1, depending on how flat the bottom of the loss curve is. In light of this, I am not sure that the work of Hall and Marron (1991) necessarily established the "inherent difficulty" of the problem. On the contrary, the result of Li (1986), as quoted in equation (5) of the article, suggests that the situation may not be as pessimistic as Professor Jones presumes, at least not in regression problems when generalized cross validation is used to select the bandwidth. Note that Li's (1986) result applies to all methods admitting an expression $\hat{\mathbf{Y}} = A(\lambda)\mathbf{Y}$, which include kernel regression and smoothing splines as special cases. Similar results hold in kernel density estimation when the kernel is chosen properly, as shown in Hall's (1987) work cited by Professor Jones, see below. With the "inherent difficulty" out of the way, it appears that Professor Jones may also endorse loss over risk.

Professor Jones correctly points out that the data may not look at all like the generating density. To guard against such circumstance, however, he suggests using bandwidth “appropriate for data sets that are typical of the true model”. This no doubt is ideal in a simulation setting. In real life, however, all one has is the data set at hand, and there often is little clue about what the true model should look like. Instead of second guessing whether the data are “typical” or “atypical”, my personal preference is to stick with a purely data-based method, even if it may not be “*very* successful” in simulations. Of course if reliable information outside of the data does exist, then one should certainly try to make use of such information. On the same line, one should by all means use an appropriate parametric model if one can be identified.

As for the discussion concerning Kullback-Leibler, I wish Professor Jones had made it clearer whether it was the loss or the bandwidth selector that he was referring to by the abbreviation “KL”. Assuming it was the loss, then I am glad that KL is unpopular only in the kernel density estimation literature, but not in the statistical community or information theoretical community at large. I am also glad to know that the principal reason behind this was KL’s “poor ISE performance”, which simply reads that *KL and ISE can be very different*, as is well-known. As metrics for the discrepancy between two probability measures, KL is invariant of a base measure substitution but ISE is sensitive to it, so I am not sure being different from ISE is necessarily a bad thing.

As for Hall’s (1987) work, it might be appropriate to let the author himself tell us what he actually proved. To quote from the abstract:

> “... Kullback-Leibler loss is an appropriate measure of distance in problems of discrimination. We examine it in the context of nonparametric kernel density estimation and show that its asymptotic properties are profoundly influenced by tail properties of the kernel and of the unknown density. We suggest ways of choosing the kernel so as to reduce loss, and describe the extent to which likelihood cross-validation asymptotically minimises loss. ...... if the kernel is chosen appropriately, then likelihood cross-validation does result in asymptotic minimisation of Kullback-Leibler loss.”

Now to this reader, the message simply says that if the model (read kernel) is wrong then the performance of V_{KL} may not be ideal, but if the model is appropriate then V_{KL} does indeed track the KL loss effectively in kernel density estimation, similar to the behavior of generalized cross validation in regression as shown by Li (1986); so much “inherent difficulty” there is. Further, there seems no question at all as to whether the KL loss is appropriate. I would need help to comprehend how Hall’s (1987) work “tolls a death knell for KL”, whether it is the loss or the bandwidth selector. To make sure that the partial quotation does not distort the message intended by the author, the interested reader is strongly urged to read the full text of the original paper.

The work of Basu Harris, Hòort and Doneg (1997) bridging KL and ISE is interesting, and it is good to know that small values of α (meaning a loss closer to KL than to ISE) "afford good robustness and high efficiency". It remains unclear to me, however, in what sense the robustness and the efficiency are defined. One usually talks about robustness and efficiency *of methods* in terms of some criterion, but robustness and efficiency *of criteria* are new to me. I am glad that Professor Jones appreciates the full efficiency of maximum likelihood, but again I would need assistance to be sure about the meaning of robustness in the context. In case the robustness is defined with respect to deviations from a parametric model, then a penalized likelihood estimate as mentioned above, which is the maximum likelihood estimate under a "soft model" of the form $\int \ddot{g} \leq \rho$, may afford both good robustness and high efficiency.

Finally, I have to admit that I could not follow Professor Jones' arguments suggesting that the source of the negative correlation is that "a data-specific optimal λ or h requires knowledge of the right answer to be drawn away from reflecting what the data looks like". I am lost. The original motivation of this work was to understand the negative correlation, and I believe the source is found to be in the across-replicate interpretability of the model index used. If meaningful negative correlation shows up with a model index that is interpretable across-replicate, I submit that one has every reason to worry about the behavior of the bandwidth selector involved. By shrugging off the issue lightly, one may be really in for the likes of "cross validation at all costs".

Scott

Professor Scott makes it crystal clear that the scales of λ and ρ are not to be compared. I wholeheartedly agree. It was for this reason, that I designed the right frame of Figure 1 in the article to illustrate the scatter through the comparative loss, where it is shown that moving ρ away from the data-specific optimum but within the across-replicate "optimal range" yields hardly any performance degradation, whereas doing the same to λ results in appreciable performance changes. As for the middle frame of Figure 1, I doubt any conclusion could be drawn from it with any aspect ratio.

I do not quite follow the discussion in the paragraph involving BCV, LSCV, and the right frame of my Figure 1. To be specific, I do not understand why cross validation has anything to do with that particular plot. It is true that later in the article we use the finding there, that λ is not interpretable across-replicate, to explain the mysterious negative correlation. But at this point, we are still investigating the properties of the two model indices, not yet charged to select either of them for practical applications.

Indeed I do not like the particular square error ISE for density estimation, but I would use the usual mean square error for regression, and I would suggest the square error $\int(\log\hat{f} - \log f)^2 f$ if one really needs a normed distance for density estimation. I am not sure how popular ISE is outside of the kernel density estimation literature. The reason for my lack of enthusiasm for ISE is its sensitivity to a base measure substitution, i.e., a transformation of the data, which is external to probability measures.

I agree with Professor Scott that when something goes wrong in estimation, it is usually the model, but not the loss, that deserves closer scrutiny. After all, the loss presumably should reflect the nature of the underlying problem, and the performace of a model is to be assessed via the loss. That said, I do believe that all losses are not created equal. The imbalance in the bias contribution from the two components of Professor Scott's favorite "hard" density may not be any indication of the deficiency of ISE, but the following fact does make me feel uncomfortable. Suppose that on an interval of positive length the true density $f = .9$ is estimated by $\hat{f} = .8$, and on another interval of the same length the true density is $f = .1$ and the estimate is $\hat{f} = .2$. The ISE takes equal contributions from the two segments, whereas the quality of the estimate in the two segments varies from very good to terrible.

Professor Scott makes a strong case for the use of locally adaptive estimates when some systematic scale imbalance is present, as in his favorite "hard" density. Such scale imbalance is also common in highly skewed data that resemble the likes of a log normal distribution. As a simple alternative to locally adaptive estimates that are inherently difficult to implement, one may instead apply some transformation to spread out points that are jammed together and to bring in the outliers. This is usually done when histograms are drawn for highly skewed data. Scaling is external to a probability measure, but when done properly it can make the resulting density easier to estimate. As a matter of fact, the estimation of a probability density with respect to a given base measure is equivalent to the estimation of the base measure that yields a uniform density. From this perspective, an appropriate front end transformation serves to bring the density to the neighborhood of a uniform one on a "macro" scale, and a subsequent density estimation using smoothing methods finishes the job on a "micro" scale. By the end of the day, however, one should transform the estimate back to the scale that best suits the application. The bottom line is that a scale more appropriate for interpretation is not necessarily the same as a scale more appropriate for estimation, but one does not need to stick to a single scale for both purposes if that makes life unnecessarily difficult. Naturally it is more appealing to have a loss that scores the same discrepancy between two probability measures no matter which scale is used, but ISE does not fit that bill.

The observation Professor Scott makes concerning the relation between the sample moments and the ISE-minimizing bandwidth is certainly very interesting. Nevertheless, I am not sure that the largely intra-replicate observation offers much explanation for the negative correlation, which is an across-replicate phenomenon. As pointed out in the article, the negative correlation is meaningful only when the h values are comparable across-replicate. Although much more subtle here, the situation is actually similar to that in the middle frame of Figure 1 involving the scales of λ and ρ, as discussed earlier. The bandwidth h conveniently indexes the effective models, but the same h value may correspond to different effective models for different samples, as the empirical evidences seem to suggest. When it is unclear whether one can compare one h with the next, I would not rush to draw any conclusion from the negative correlation, whether it is to dismiss the bandwidth selector or to dismiss the minimum loss. As for the correlation in the ρ plot in the right frame of Figure 2, it is -0.0626, just for the record. This correlation however is a nonissue because there is virtually no spread on the horizontal axis of that plot, *not measured by the scale used in the plot but by the relative loss illustrated in the right frame of Figure 1.*

As for the possibility of exploring the issues using Terrell's penalized estimates, another discussant, Professor Jones, had already made some attempt. Please see Professor Jones' comments for how successful the attempt was, and my earlier reply for my reading of the results.

Wahba

First of all, I would like to thank Professor Wahba for her appreciation of the work.

The fact that the failure rate of GCV decreases as $n \to \infty$ comes as no surprise, it is simply asymptotics at work. It is very nice that Wahba and Wang (1985) actually worked out its rapid decay rate, providing important assurance for practitioners. As for the Monte-Carlo GCV with randomized trace$A(\lambda)$, I would like to add that the same asymptotics as quoted in equation (5) of the article still hold, as was shown by Girard (1991).

As for Professor Wahba's k index, I observe that

$$\begin{aligned}
&\sigma^2 k(\lambda)\\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_\lambda(x_i))^2\\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \frac{1}{n}\sum_{i=1}^{n}(f_\lambda(x_i) - f(x_i))^2 - \frac{2}{n}\sum_{i=1}^{n}(Y_i - f(x_i))(f_\lambda(x_i) - f(x_i))
\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + L(\lambda) - 2T(\lambda) - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i(b_i + v_i),$$

where $T(\lambda) = n^{-1}\sum_{i=1}^{n} a_{i,i}\epsilon_i^2$, $b_i = \sum_{j=1}^{n} a_{i,j}f(x_j) - f(x_i)$, and $v_i = \sum_{j\neq i} a_{i,j}\epsilon_j$, for $a_{i,j}$ the (i,j)th entry of $A(\lambda)$. Following arguments developed by Li (1986), it can be shown that $n^{-1}\sum_{i=1}^{n}\epsilon_i b_i = o_p(L(\lambda))$ and $n^{-1}\sum_{i=1}^{n}\epsilon_i v_i = o_p(T(\lambda))$. Now as $n^{-1}\sum_{i=1}^{n}\epsilon_i^2 = \sigma^2(1+O_p(n^{-1}))$ and $O_p(n^{-1}) = o_p(L(\lambda))$, $O_p(n^{-1}) = o_p(T(\lambda))$, one has

$$\sigma^2 k(\lambda) = \sigma^2 + L(\lambda)(1 + o_p(1)) - 2T(\lambda)(1 + o_p(1)).$$

Hence, Professor Wahba's plots show that $L(\hat\lambda) - 2T(\hat\lambda)$ and $L(\lambda^{opt}) - 2T(\lambda^{opt})$ are positively correlated. To go much further beyond this to explain the positive correlation, one seems to need a big assumption that the ratio $L(\lambda)/T(\lambda)$ remains roughly a constant near λ^{opt} *across-replicate*, which however may not hold. In any case the positive correlation in k is just as inconsequential as the negative correlation in λ, as Professor Wahba pointed out, but the very fact that equivalent model indices may yield completely different correlation patterns clearly indicates that correlation patterns are nothing but artifacts due to model indexing.

As for Professor Wahba's conjecture, that the index $L(f_\lambda|\mathbf{Y})$ affords a positive correlation between the optimal index and some cross validation index for f_λ, the minimizer of a penalized likelihood functional $L(f|Y) + \lambda J(\lambda)$, all I am able to do at this point is to add further empirical evidence to its credit. A simulation study was conducted for the penalized likelihood density estimates mentioned in the reply to Jones above. Generated were 100 replicates of samples of size $n = 100$ from the test density $f_1(x)$, the half-half mixture of $N(.35, (.1)^2)$ and $N(.65, (.1)^2)$ truncated to $[0, 1]$. For each of the replicates, an automatic estimate was computed using the performance-oriented iteration of Gu (1993), which employs a λ selector that contains a cross validation component. Fixed λ solutions were also calculated on a fine grid $\log_{10}\lambda = (-7)(.05)(-3)$. Recorded on each estimate $\hat g$ are λ, $\rho = \int_0^1 \ddot{\hat g}$, $L(f) = -(1/n)\sum_{i=1}^{n}\hat g(Y_i) + \log\int_0^1 e^{\hat g}$, and the symmetrized Kullback-Leibler loss $L(\lambda) = \mu_{\hat g}(\hat g - g) + \mu_g(g - \hat g)$, where $\mu_g(h) = \int_0^1 he^g/\int_0^1 e^g$. The counterparts of Figure 1 in the article are qualitatively indistinguishable from the one presented for regression, and are omitted here. The counterparts of Figure 2 in the article are plotted in Figure 9, where the left frame is replaced by the optimal versus the cross validation $L(f)$ index. Note that in this setting it is not clear what makes a "prediction" for Y_i, nor how a leaving-out-one lemma might look like, yet the conjecture still holds. This certainly makes no case for the most general circumstances, but nevertheless appears to be a bit beyond the scope originally anticipated by Professor Wahba.

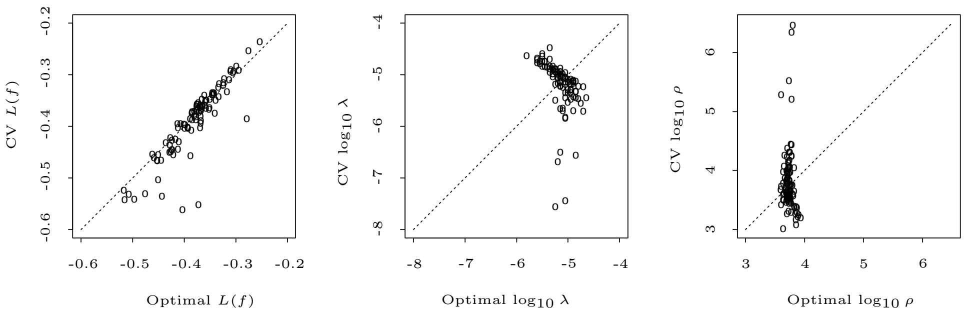

Figure 9. The correlation patterns in penalized likelihood density estimation. Left: Cross-validation $L(f)$ versus optimal $L(f)$. Center: Cross-validation λ versus optimal λ. Right: Cross-validation ρ versus optimal ρ.

To conclude, I would like to thank the discussants again for their thought-provoking comments. The views of the parties are apparently very different at places, which is natural given the subtleness of the issues involved. I hope we all learned something from this discussion. At least I did.

I am indebted to the editors of the journal: Professor Jeff Wu who invited the article, and Professor Ching-Shui Cheng who handled the article with care and organized this discussion. Without their encouragement and understanding, this work could not have made its way into a printed journal.

Additional References

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, to appear.

Chen, S., Donoho, D. and Saunders, M. (1995). Atomic decomposition by basis pursuit. Technical Report 479, Dept of Statistics, Stanford University, Stanford CA.

Cox, D. (1983). Asymptotics for M type smoothing splines. *Ann. Math. Statist.* **11**, 530-551.

DeVore, R. and Lucier, B. (1992). Fast wavelet techniques for near-optimal image processing. In 1992 *IEEE Military Communications Conference, IEE Communications Society*, 1129-1135.

Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B* **57**, 301-370.

Girard, D. A. (1991). Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization. *Ann. Statist.* **19**, 1950-1963.

Girard, D. (1995). The fast Monte-carlo cross-validation and C_L procedures: Comments, new results and application to image recovery problems. *Comput. Statist.* **10**, 205-231.

Gong, J., Wahba, G., Johnson, D. and Tribbia, J. (1998). Adaptive tuning of numerical weather prediction models: simultaneous estimation of weighting, smoothing and physical parameters. *Monthly Weather Review* **125**, 210-231.

Grund, B., Hall, P. and Marron, J. S. (1994). Loss and risk in smoothing parameter selection. *J. Nonparametr. Statist.* **4**, 107-132.

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *Ann. Statist.* **21**, 217-234.

Hall, P. (1987). On Kullback-Leibler loss and density estimation. *Ann. Statist.* **15**, 1491-1519.

Hall, P. and Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6**, 109-115.

Hall, P. and Marron, J. S. (1991). Lower bounds for bandwidth selection in density estimation. *Probab. Theory Rel. Fields* **90**, 149-173.

Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved AIC criterion. *J. Roy. Statist. Soc. Ser. B* **60**, 271-293.

Jones, M. C. (1991). The role of ISE and MISE in density estimation. *Statist. Probab. Lett.* **12**, 51-56.

Jones, M. C. (1998). On some kernel density estimation bandwidth selectors related to the double kernel method. *Sankhyā Ser. A*, to appear.

Jones, M. C. and Kappenman, R. F. (1992). On a class of kernel density estimate bandwidth selectors. *Scand. J. Statist.* **19**, 337-349.

Jones, M. C. and Sheather, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **11**, 511-514.

Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika* **81**, 673-680.

Marron, J. S. and Tsybakov, A. B. (1995). Visual error criteria for qualitative smoothing. *J. Amer. Statist. Assoc.* **90**, 499-507.

Rudin, L., Osher, S. and Fatemi, E. (1992). Nonlinear total-variation-based noise removal algorithms. *Physica D* **60**, 259-268.

Scott, D. W. (1988). Comment on "How far are automatically chosen regression smoothing parameters from their optimum?" by W. Härdle, P. Hall and J. S. Marron. *J. Amer. Statist. Assoc.* **83**, 96-98.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12**, 898-916.

Terrell, G. R. (1990). Linear density estimates. *Proc. Statist. Comput. Sec., Amer. Statist. Assoc.*, 297-302.

Terrell, G. R. (1993). Spline density estimators. *ASA Proc. Statist. Comput. Section* 255-260.

Wahba, G. (1975*a*). Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation. *Ann. Statist.* **3**, 15-29.

Wahba, G. (1975*b*). Smoothing noisy data by spline functions. *Numer. Math.* **24**, 383-393.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45**, 133-150.

Wahba, G. (1998). Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. Technical Report 984r, Department of Statistics, University of Wisconsin, Madison WI. In *Advances in Kernel Methods-Support Vector Learning* (Edited by B. Schölkopf, C. Burges and A. Smola), MIT Press.

Wahba, G. and Wang, Y. (1995). Behavior near zero of the distribution of GCV smoothing parameter estimates. *Statist. Probab. Lett.* **25**, 105-111.

Wahba, G. and Wold, S. (1975). A completely automatic French curve. *Commun. Statist.* **4**, 1-17.

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statist. Sinica* **6**, 675-692.