Model Indexing and Model Selection in Nonparametric Function Estimation

by

Chong Gu
Purdue University

Technical Report #93-55

Department of Statistics
Purdue University

December 1993

△

Model Indexing and Model Selection in Nonparametric Function Estimation

CHONG GU*

Abstract

The role of statistical model seems to be largely neglected in the existing literature on nonparametric function estimation. As a consequence, a few popular working concepts in nonparametric estimation, such as the expected mean square error and the "degrees of freedom", appear vulnerable under close scrutiny. Through heuristic arguments and simple simulations, we try to illustrate that the model indexing via the usual smoothing parameter may lead to conceptual pitfalls if care is not taken. Due to technical constraint, the arguments are mainly developed in the penalized likelihood setting, but we shall discuss the ramifications in other settings as well. This note results from an effort to understand the well-publicized negative correlation between optimal and cross-validation smoothing parameters.

KEY WORDS: Constraint; Cross-validation; Model indexing; Model selection; Penalized likelihood; Smoothing parameter.

1 Introduction

Nonparametric function estimation has been one of the most active research areas in contemporary statistics. In spite of the ever growing number of procedures being proposed, and theorems being proved, however, there remain a few basic concepts to be clarified, and a few mysterious phenomena to be understood. Stemming from an attempt to understand the counter-intuitive negative correlation between the optimal and cross-validation smoothing parameters, to be reproduced in Section 3, we present a set of heuristic arguments and numerical simulations, to offer our views on the concepts, the intuitions, and the explanations of the mystery.

*Chong Gu is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, Indiana 47907. The author is grateful to Xiao-Tong Shen, Grace Wahba, and Yue-Dong Wang for discussion, comments and suggestions. This research was supported by National Science Foundation under Grant DMS-9301511.

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

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

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 model selection. Our arguments are based on two heuristics, to follow.

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 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 model, and the assumptions actively in force may not be the ones explicitly stated. It seems always possible, however, to perceive conceptually some effective constraint which the data make compromise with 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 the estimate has been subject to.*

For some nonparametric procedure, as we shall see shortly, an explicit description of the effective constraint is available, which lends insights for us to understand the mystery. For other procedures, the effective constraint remains 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 functions, on which estimates may be compared. 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 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, and hence there should be nearly 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 largely remain 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.

Our arguments are developed under the setting of penalized likelihood estimation. Assume $f(x)$ to be smooth, in the sense that its second derivative exists and is small. A popular approach to the estimation of $f(x)$ is via 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 interpolator. The practicability of the method hinges on a good choice of λ, the selection of a good model from a continuum of available models.

In Section 2, we shall discuss the proper indexing of models behind the minimizers of (2) via an explicit characterization of the effective constraint; it turns out that the smoothing parameter λ has little statistical meaning across-replicate, and hence any across-replicate concepts directly indexed by λ are likely to mislead. In section 3, we discuss the ramifications of model indexing in model selection, and demonstrate that the mysterious negative correlation is actually an illusion due to improper model indexing. Under settings other than penalized likelihood, the effective constraint behind an estimate is usually very difficult to describe if at all possible, but one may still assess the across-replicate interpretability of an index via simulation, possibly with the help of Heuristic 2; an example is presented in Section 4.

2 Model Indexing

Consider the constrained least squares problem of minimizing

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

for the estimation of $f(x)$ in (1). The solution of such a problem usually falls on the sphere $\int_0^1 \ddot{f}^2(x)dx = \rho$, 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, we see that 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 form $\int_0^1 \ddot{f}^2(x)dx \leq \rho$. See, e.g., Schoenberg (1964).

The models behind (2) are to be characterized by $\int_0^1 \ddot{f}^2(x)dx \leq \rho$, with a natural index ρ. 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 $\int_0^1 \ddot{f}^2(x)dx \leq \rho$, 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 the understanding of further discussion.

Now consider a simple simulation. On $x_i = (i-.5)/50, i = 1, \cdots, 50$, 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 $(1/50)\sum_{i=1}^{50}(\hat{f}(x_i) - f(x_i))^2$. The grid was broad enough to bracket the best-performing estimates for all the 100 replicates.

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 envelop 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

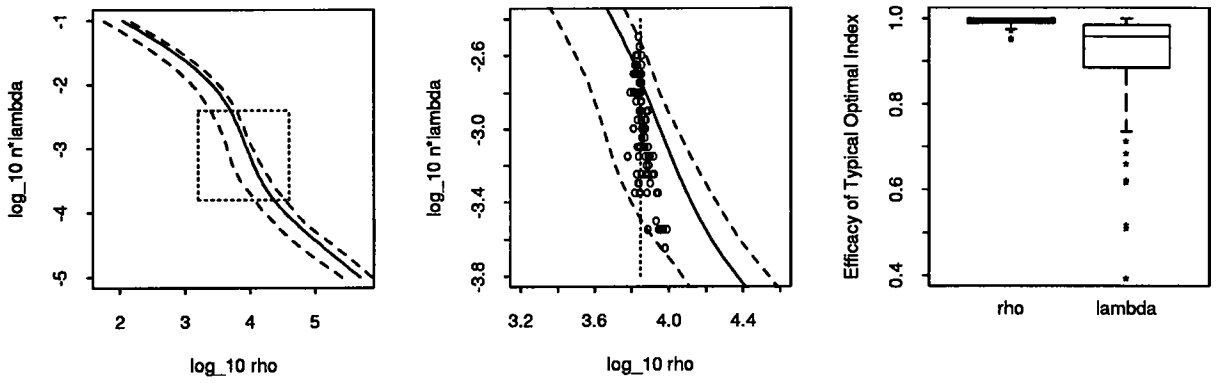

Figure 1: The ρ index and the λ index of models in simulation.

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 $\log_{10} n\lambda = -3$ 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 over 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 order-of-magnitudely 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 the estimates have been subject to; 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) with fixed λ, and by f_ρ the solution of (3) with fixed ρ. A tempting loss function for the study 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 loss function is unfortunately defective: Because a fixed λ implies different models for different realizations of ϵ_i, the expectation is effectively mixing apples with oranges. Concept based on the exact quantification of (4) such as the minimizer of $R(\lambda)$ as the across-replicate "optimal" λ is 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 the analysis of constrained problem is less tractable. Despite the conceptual defect $R(\lambda)$ suffers, however, 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 of $E(1/n)\sum_{i=1}^n (f_\lambda(x_i) - f(x_i))^2 = O(K)$ 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.

Define $\hat{Y}_i = f_\lambda(x_i)$. 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). An ever 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 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, but 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 degrees-of-freedom.

3 Model 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 no strategy by any standard, 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

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

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 any good for the observed data.

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

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

for use in (2), where the matrix $A(\lambda)$ is as defined in Section 2. The score $V(\lambda|\boldsymbol{Y})$ is data-specific, whose minimizer λ_* can be shown to approximately minimize the data-specific loss function $R(\lambda|\boldsymbol{Y})$, in the sense that $1 - \min_\lambda R(\lambda|\boldsymbol{Y})/R(\lambda_*|\boldsymbol{Y}) = o_p(1)$, of course under conditions; see Li (1986). Note that the data in $V(\lambda|\boldsymbol{Y})$ and $R(\lambda|\boldsymbol{Y})$ have to be the same to make this work. One may perceive $V(\lambda|\boldsymbol{Y})$ as a computable proxy of the data-specific loss function $R(\lambda|\boldsymbol{Y})$, based on which the procedure seeks to approximately locate the data-specific optimal 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 $R(\lambda_*|\boldsymbol{Y})$ versus the loss of the best-performing estimates $\min_\lambda R(\lambda|\boldsymbol{Y})$ for each of the replicates. A point on the dotted line indicates a perfect performance of the procedure. Statistical estimation has to be employed in $V(\lambda|\boldsymbol{Y})$ for information carried by the unknown truth $f(x)$ in $R(\lambda|\boldsymbol{Y})$, which is subject to error, so the procedure is not guaranteed to work on every data set, and indeed it worked rather poorly on a few of the replicates. The general performance however appears satisfactory.

In the course of the above simulation, we have collected sufficient information to reproduce in the center frame of Figure 2 the well-publicized negative correlation between the optimal and cross-validation smoothing parameters, 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

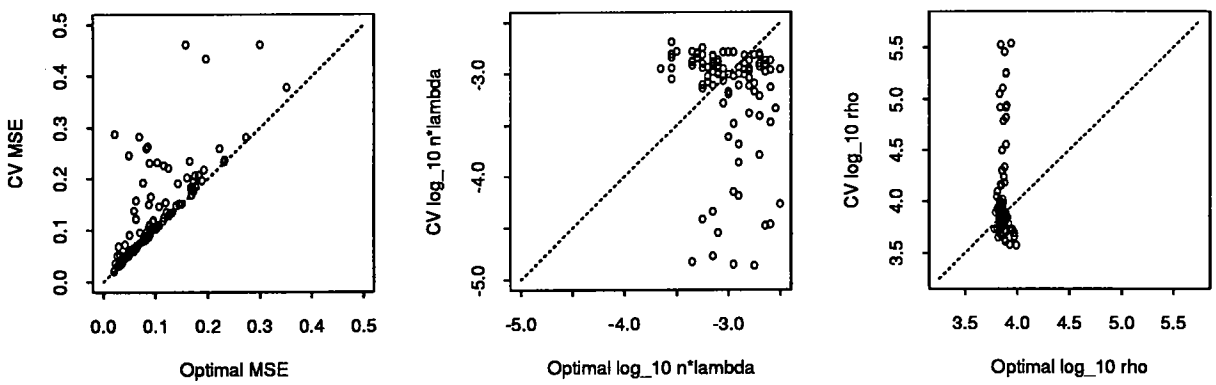

Figure 2: Performance of generalized cross-validation in simulation.

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 errors in the "estimation" of $R(\lambda|\boldsymbol{Y})$ by $V(\lambda|\boldsymbol{Y})$, the actually adopted models are scattered nearby except for a few wild failures. Cross-validation may not be the final word for model 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, 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.

4 Further ramifications

For nonparametric methods other than penalized likelihood, there doesn't seem to exist the luxury of explicit model characterization as in (3). Nevertheless, given a continuum, or almost a continuum, of possible estimates, the nature of the "natural" model index may imply some dos and don'ts in the theory and practice of the methods, where "natural" often means only. It is advisable to carefully examine a working index before loading too much on it.

As an example, let us look at the kernel method for density estimation, for which simulations similar to what we report here are readily available in the literature. Observing X_i, $i = 1, \cdots, n$, from a probability density $f(x)$, one is to estimate $f(x)$ by a function of the form

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}), \tag{5}$$

where $K(x)$ is some known smooth function satisfying $\int K(x)dx = 1$, and the so-called bandwidth h acts as the smoothing parameter. The bandwidth h appears to be the only index one can work with in this setting, and an explicit description of the effective constraint seems nowhere in sight.

A natural question to ask is whether the h index is interpretable across-replicate. In lack of a ρ-like index, Heuristic 1 offers little help in this regard. Besides the negative correlation, however, there is more to learn from plots resembling the center frame of Figure 2. An eminent feature demonstrated in the plot is that the scatter of the optimal indices across-replicate is comparable to 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 don't compare with each other. By examining plots similar to the right frame of Figure 1, or by assessing the performance of the estimates over the scatter for a particular replicate, one should be able to confirm that the former is not the case. Applying this argument to the simulations reported by Scott and Terrell (1987) and Hall and Johnstone (1992), we conclude that the h index of (5) bears little statistical meaning across-replicate.

The lack of a ρ-like index doesn't necessarily affect the theory and practice of the method, but the limited interpretability of the h index certainly has its ramifications. The cautions for the λ index of (2) all seem to apply to the h index of (5). Because the optimal bandwidth is data-specific, an effective model selection procedure has to be data-specific. This disqualifies naive resampling methods for h selection. Similar to (4), an expected mean square error of f_h indexed by h may not make much practical sense, and an across-replicate "optimal" h is of little practical meaning: Assuming such an "optimal" bandwidth is known, what good does it do when the optimal h for the observed data is known to be somewhere else?

The message should be simple and clear, albeit subtle. Using whatever model index, there seems little to lose if one strives to locate a data-specific optimal index. When one needs to borrow information external to the observed data, however, it is important to make sure that the working index is meaningful across-replicate; a carefully interpreted simple simulation often helps

in this regard. By all means, one should avoid across-replicate operation on any index which is not interpretable across-replicate.

References

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear smoothers and additive models" (with discussion), *The Annals of Statistics*, 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," *Numerische Mathematik*, 31, 377 – 403.

Gu, C. (1992), "Cross-validating non Gaussian data," *Journal of Computational and Graphical Statistics*, 1, 169 – 179.

—— (1993), "Smoothing spline density estimation: A dimensionless automatic algorithm," *Journal of the American Statistical Association*, 87, 1051 – 1058.

Hall, P. and Johnstone, I. (1992), "Empirical functionals and efficient smoothing parameter selection" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 54, 475 – 530.

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

Schoenberg, I. J. (1964), "Spline functions and the problem of graduation," *Proceedings of the National Academy of Science*, 52, 947 – 950.

Scott, D. W. and Terrell, G. R. (1987), "Biased and unbiased cross-validation in density estimation," *Journal of the American Statistical Association*, 82, 1131 – 1146.

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