Table 1. Bias and MSE of $\widehat{Err} - Err$ Evaluated at $\hat{\mu}(\hat{M})$ With $\hat{M}$ Selected via AIC and the Corresponding Standard Errors (in parentheses) of the Two Approaches Based on 500 Replications

5k	Methods	Bias	MSE
5	DP	$-4.065_{(.956)}$	$472.66_{(31.00)}$
	PB	$-21.037_{(.934)}$	$877.42_{(52.83)}$
15	DP	$-4.467_{(.911)}$	$434.47_{(27.39)}$
	PB	$-17.575_{(.892)}$	$706.20_{(40.53)}$
25	DP	$-4.389_{(.959)}$	$477.78_{(29.57)}$
	PB	$-14.090_{(.942)}$	$641.26_{(38.92)}$
35	DP	$-1.644_{(1.053)}$	$556.51_{(35.73)}$
	PB	$-7.104_{(1.054)}$	$604.90_{(38.70)}$
45	DP	$-1.620_{(1.098)}$	$604.46_{(37.55)}$
	PB	$-.349_{(1.085)}$	$587.19_{(36.16)}$

are computed by averaging over 100 replications and are reported in Table 1.

Clearly, the PB method performs well and less well for large k and small k values, respectively, because of the choice of the "moderately big" model. Evidently, the estimator $\hat{\mu}_{\text{full}}$ estimates μ well for small k values but poorly for large k values, depending on the true model. In terms of the accuracy of prediction, $\widehat{\text{Err}}$ estimates Err poorly for small k values, yielding bias against candidate models of small size, and vice verse for large k values. Generally, it is impossible to eliminate this problem if any model-dependent $\hat{\mu}$ is used for μ_s in sampling. By comparison, the "model-free" DP method estimates Err consistently well across all situations.

ADDITIONAL REFERENCES

Freedman, D. A., Navidi, W., and Peters, S. C. (1988), "On the Impact of Variable Selection in Fitting Regression Equations," in *On Model Uncertainty and Its Statistical Implications*, ed. T. K. Dijkstra, New York: Springer-Verlag, pp. 1–16.

George, E. I., and Foster, D. P. (2000), "Calibration and Empirical Bayes Variable Selection," *Biometrika*, 87, 731–747.

Rissanen, J. (1996), "Fisher Information and Stochastic Complexity," *IEEE Transactions on Information Theory*, 42, 40–47.

Comment

Chunming ZHANG

A fundamental issue in statistics is to quantify the degree to which a model captures an underlying reality and predicts future cases. With the growing flood of increasingly complex data in real-world applications, it has become pressingly important for statisticians to develop theory and methods that allow dual use of data in making effective assessment of model fitting and critical evaluation of model prediction. The central problem studied in Professor Efron's article is that of estimating the true prediction error. Efron's article has substantially enhanced our understanding of this important problem. I appreciate the opportunity to comment further on this neat and stimulating article.

Efron revisits a well-known model-free method for estimating the prediction error based on cross-validation (CV). This procedure, beginning with the delete-one-out fitted value $\widetilde{\mu}_i$ for outcome y_i, directly estimates the coordinatewise true predictive error, Err_i, by $\widehat{\text{Err}}_i^{\text{CV}} = Q(y_i, \widetilde{\mu}_i)$, with respect to a Q-error measure, and as such adjusts the apparent error, $\text{err}_i = Q(y_i, \widehat{\mu}_i)$, for the full data-based fitted value $\widehat{\mu}_i$, by an amount $\widetilde{O}_i = Q(y_i, \widetilde{\mu}_i) - Q(y_i, \widehat{\mu}_i)$, yielding an equivalent form of CV,

$$\widehat{\text{Err}}_i^{\text{CV}} = \text{err}_i + \widetilde{O}_i, \qquad i = 1, \ldots, n. \tag{1}$$

In many applications, the original cross-validated methods have known to suffer from large variations.

With the introduction of optimism theorem and Rao–Blackwell type of results, Efron not only provides valuable theoretical tools, but also brings new insights into what has been learned before about CV and opens up new vistas in exploration and learning. Among many other contributions, Efron

1. Derives an optimism theorem to represent the expected optimism, $\Omega_i = E(\text{Err}_i - \text{err}_i)$, as the *covariance penalty*, $\Omega_i = 2\,\text{cov}(\widehat{\lambda}_i, y_i)$, with $\widehat{\lambda}_i$ some well-defined mapping of $\widehat{\mu}_i$. In this spirit, the covariance penalty (CP) method, $\widehat{\text{Err}}_i^{\text{CP}}$, estimates Err_i, via estimating the covariance penalty, $\text{cov}_i = \text{cov}(\widehat{\lambda}_i, y_i)$, by some data-driven rule, $\widehat{\text{cov}}_i$, leading to an additive form,

$$\widehat{\text{Err}}_i^{\text{CP}} = \text{err}_i + 2\,\widehat{\text{cov}}_i, \qquad i = 1, \ldots, n. \tag{2}$$

   The covariance penalty theory goes beyond the squared error to a q class of error measures Q, and thus generalizes the work of Mallow's C_p, Akaike's information criterion, and Stein's unbiased risk estimate to a wide range of statistical models. He also develops model-based bootstrap methods to estimate the covariance term.
2. Characterizes Rao–Blackwell type of results to demonstrate that the covariance penalty method enjoys substantially increased efficiency than the conventional CV method for estimating prediction error. These theoretical results offer a very appealing and easily understandable interpretation of two prediction error estimation schemes, which, as can be seen from (1) and (2), operate in very distinct ways.
3. Suggests methods to improve the original CV estimates and the nonparametric bootstrap estimates for prediction error.

Chunming Zhang is Assistant Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: *cmzhang@stat.wisc.edu*). The research was supported in part by National Science Foundation grant DMS-03-53941.

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Theory and Methods
DOI 10.1198/016214504000000890

1. CONDITIONAL MONOTONICITY: NONNEGATIVITY OF COVARIANCE PENALTIES

As pointed out by Efron, one problem arising from the use of the apparent error, err_i, is that it tends to be biased *downward* for the true predictive error, Err_i. Is err_i always biased *downward*? From the viewpoint of optimism theorem, this seems particularly relevant to the question of whether or not the covariance penalty, cov_i, is nonnegative. For the usual squared error measure Q, applied to a linear fitting rule $\widehat{\mu}_i$ (such as smoothing splines, regression splines, wavelet estimators, kernel and local polynomial regression estimators), it is conceivable that the resulting covariance, $\text{cov}_i = \text{cov}(\widehat{\mu}_i, y_i)$, is indeed positive. How can one better understand this implicit feature of the covariance penalty under more general error measures Q in accordance with possibly nonlinear fitting rules?

In what follows, I try to provide some simple arguments for the conditional monotonicity of $\widehat{\lambda}_i$ to illustrate when the desired inequality, $\text{cov}(\widehat{\lambda}_i, y_i) \geq 0$, holds for the generalized q class of error measures Q and when it does not. Let $\mathbf{y}_{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. Note that $\widehat{\mu}_i = \widehat{\mu}_i(\mathbf{y}_{(i)}, y_i)$ and $\widehat{\lambda}_i = \widehat{\lambda}_i(\mathbf{y}_{(i)}, y_i) = -q'(\widehat{\mu}_i)/2$ (defined in Section 3 of Efron's article). Then $\text{cov}(\widehat{\lambda}_i, y_i)$ can be rewritten as

$$E\{\widehat{\lambda}_i \cdot (y_i - \mu_i)\} = E\big[E\big\{\widehat{\lambda}_i\big(\mathbf{y}_{(i)}, y_i\big) \cdot (y_i - \mu_i)\big|\mathbf{y}_{(i)}\big\}\big]. \tag{3}$$

To facilitate discussion, assume that the second derivative of $q(\mu)$ exists. When examining the conditional expectation in (3), it is seen that, for fixed $\mathbf{y}_{(i)}$,

$$\begin{aligned}\frac{\partial \widehat{\lambda}_i(\mathbf{y}_{(i)}, y_i)}{\partial y_i} &= \frac{\partial \widehat{\lambda}_i(\mathbf{y}_{(i)}, y_i)}{\partial \widehat{\mu}_i(\mathbf{y}_{(i)}, y_i)} \frac{\partial \widehat{\mu}_i(\mathbf{y}_{(i)}, y_i)}{\partial y_i} \\ &= -\frac{1}{2} q''\big(\widehat{\mu}_i\big(\mathbf{y}_{(i)}, y_i\big)\big) \frac{\partial \widehat{\mu}_i(\mathbf{y}_{(i)}, y_i)}{\partial y_i}. \end{aligned} \tag{4}$$

On the right side of (4), the choice of a concave function q, as introduced in Efron's article to define Q (and ensure $Q \geq 0$), entails $-q''(\widehat{\mu}_i(\mathbf{y}_{(i)}, y_i)) \geq 0$. Meanwhile, the other term in (4), $\partial \widehat{\mu}_i(\mathbf{y}_{(i)}, y_i)/\partial y_i$, measures the sensitivity of a fitted value to perturbation in the corresponding observed value (Ye 1998). These two considerations lead to the following conclusions:

1. If $\partial \widehat{\mu}_i(\mathbf{y}_{(i)}, y_i)/\partial y_i \geq 0$, (4) indicates that $\partial \widehat{\lambda}_i(\mathbf{y}_{(i)}, y_i)/\partial y_i \geq 0$. The implication is that, given $\mathbf{y}_{(i)}$, $\widehat{\lambda}_i(\mathbf{y}_{(i)}, y_i)$ is a nondecreasing function of y_i and that $\widehat{\lambda}_i(\mathbf{y}_{(i)}, y_i)$ and $y_i - \mu_i$ are monotone in the same directions. An appealing to some expanded version of Chebyshev's inequality (see, e.g., Gurland 1967, p. 25) yields $E\{\widehat{\lambda}_i(\mathbf{y}_{(i)}, y_i) \cdot (y_i - \mu_i)|\mathbf{y}_{(i)}\} \geq 0$, which, applied to (3), in turn induces $\text{cov}(\widehat{\lambda}_i, y_i) \geq 0$.
2. On the contrary, if $\partial \widehat{\mu}_i(\mathbf{y}_{(i)}, y_i)/\partial y_i \leq 0$, then $\text{cov}(\widehat{\lambda}_i, y_i) \leq 0$, revealing that err_i tends to be an *upward* biased estimator of Err_i.

2. RAO–BLACKWELL THEOREM: VARIANCE REDUCTION OF COVARIANCE PENALTY METHOD

A key quantity of interest in the conclusion of Theorem 1 is the Rao–Blackwell type of relation established between the covariance penalty method and the CV counterpart. Some remarkable aspect of the proof rests on a careful construction of the bootstrap data $(\mathbf{y}_{(i)}, y_i^*)$, in which $\mathbf{y}_{(i)}$ is kept fixed and, given $\mathbf{y}_{(i)}$, the probability mechanism of y_i^* dictates its conditional distribution $\widetilde{f}_i$, with the conditional mean $E_{\widetilde{f}_i}\{y_i^*|\mathbf{y}_{(i)}\} = \widetilde{\mu}_i$. Based on the same data $(\mathbf{y}_{(i)}, y_i^*)$, the associated CV estimate, $\widetilde{O}_i^* = Q(y_i^*, \widetilde{\mu}_i) - Q(y_i^*, \widehat{\mu}_i(\mathbf{y}_{(i)}, y_i^*))$, is compared with the conditional version of the covariance penalty estimate, $2\widehat{\text{cov}}_{(i)} = 2\,\text{cov}^*_{\widetilde{f}_i}\{(\widehat{\lambda}_i(\mathbf{y}_{(i)}, y_i^*), y_i^*)|\mathbf{y}_{(i)}\}$. Efron shows that $E_{\widetilde{f}_i}\{\widetilde{O}_i^*|\mathbf{y}_{(i)}\} \doteq 2\widehat{\text{cov}}_{(i)}$.

I find this result attractive because it integrates the classical theory of point estimation with the prediction error estimation techniques, and therefore enables one to further comprehend the stochastic way that distinguishes the covariance penalty method from the CV method. Meanwhile, I discuss some additional questions regarding how to compare these two methods.

1. From the preceding data construction, the reader can clearly observe that $\widetilde{O}_i^*$ is introduced to mimic (or predict) an observable random variable, namely, the term $\widetilde{O}_i$ in (1), whereas $2\widehat{\text{cov}}_{(i)}$, similar to the term $2\widehat{\text{cov}}_i$ in (2), aims to estimate an unknown deterministic quantity, $2\,\text{cov}_i$. Henceforth, it may not strike the reader as particularly surprising that the variance of $\widetilde{O}_i^*$ exceeds that of $2\widehat{\text{cov}}_{(i)}$.
2. To better appreciate the value of the covariance penalty method, it would be natural to quantify how much variance reduction is achieved by $2\widehat{\text{cov}}_{(i)}$ relative to $\widetilde{O}_i^*$. In addition to carrying out the simulation studies, some theoretical calculations in certain concrete examples will be particularly interesting and enlightening.
3. A homoscedastic model, assumed for data points displayed in figure 1, facilitates the parametric bootstrap computations. Had this type of deviation from model assumptions existed, would the model-based covariance penalty estimates have been affected?
4. More precisely speaking, the Rao–Blackwell type of result compares the relative performance of the CV and covariance penalty methods in estimating the expected optimism; this thoughtful result, when placed back into (1)–(2), gives an indirect way of comparing the prediction error estimation. In practical settings, a direct way of assessing the two methods is to compare $\text{var}(\widehat{\text{Err}}_i^{\text{CV}})$ versus $\text{var}(\widehat{\text{Err}}_i^{\text{CP}})$. Generally, the original CV estimate, $\widehat{\text{Err}}_i^{\text{CV}}$, becomes less noisy as the sample size increases.

3. DEGREES OF FREEDOM: DIRECT ESTIMATION OF COVARIANCE PENALTIES

Ideally, the covariance penalty would be known, or could easily be estimated by a data-oriented procedure. The parametric bootstrap method suggested in Efron's article provides a useful device in general situations. This approach consists of generating bootstrap resamples $\mathbf{y}^{*b}$, $b = 1, \ldots, B$, at the ith individual data point, from a "bootstrap model" assumed to be "believable," and obtaining the replicated estimates $\widehat{\mu}_i^{*b}$ and $\widehat{\lambda}_i^{*b}$. While producing the bootstrapped estimates of covariance at the entire collection of sample points is suitable for samples of small or medium size, it can potentially become a problem for large and huge sample sizes that one may face nowadays in data-mining tasks. Typical examples include processing functional data (Ramsay and Silverman 1997) and longitudinal data (Diggle, Heagerty, Liang, and Zeger 2002), in which each data element is associated with a high-dimensional

curve, other than a univariate number. The computational burden of the bootstrap procedure will continue to grow as demand increases for a more complicated model-fitting technique. Moreover, there is no unique way of building a "bootstrap model." On the other hand, care needs to be taken to reduce biases caused by an inadequate choice of the "bootstrap model." This is particularly important when the data structure is complex; see further examples in Section 4.1.

For practical purposes, some alternative methods for estimating covariance penalty within the different contexts of its use deserve further exploration. Below I will focus on the situations in which some nonparametric modeling techniques are employed. In these cases, the covariance penalty either is fully known or can be approximated by its asymptotic expression in large samples.

Case I. Consider $\mathbf{y} \sim (\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$. Recall that for a squared error measure combined with any linear fitting rule, $\mathrm{cov}_i = \sigma^2 M(i,i)$ and $\sum_{i=1}^n \mathrm{cov}_i = \sigma^2 \operatorname{tr}(M)$. Under a nonparametric regression model, if the mean response is fitted by a linear nonparametric smoother, such as the local polynomial regression estimator (see, e.g., Fan and Gijbels 1996), then $M_h(i,i)$ has a closed-form expression and thus the exact values of the total degrees of freedom, $\operatorname{tr}(M_h)$ and $\operatorname{tr}(M_h^T M_h)$, can be directly computed, in which M_h is used to denote its dependence on a bandwidth parameter h. The unknown parameter σ^2 can be estimated by a nonparametric variance estimator, $\widehat{\sigma}^2 = \sum_{i=1}^n (y_i - \widehat{\mu}_i)^2 / \{n - \operatorname{tr}(2M_h - M_h^T M_h)\}$ (Buckley, Eagleson, and Silverman 1988; Cleveland and Devlin 1988). Hence, the total covariance penalties can be directly estimated whenever the sample size keeps the computational cost affordable. Furthermore, Zhang (2003a) showed that $\operatorname{tr}(M_h) \doteq d\{(p+1-a) + \mathcal{C}n/(n-1)\mathcal{K}(0)|\Omega|/h\}$ and $\operatorname{tr}(M_h^T M_h) \doteq d\{(p+1-a) + \mathcal{C}n/(n-1)\mathcal{K} * \mathcal{K}(0)|\Omega|/h\}$ inform the asymptotic total degrees of freedom in a univariate nonparametric regression model and a varying-coefficient regression model, where all of the constants involved in the expressions are known. These empirical formulas suggest a second way of directly estimating the total covariance penalties, by

$$\sum_{i=1}^n \widehat{\mathrm{cov}}_i = \widehat{\sigma}^2 d\big\{(p+1-a) + \mathcal{C}n/(n-1)\mathcal{K}(0)|\Omega|/h\big\}. \quad (5)$$

Case II. Consider response observations from the exponential family with a density (or probability) function, $\exp[\{y_i\theta_i - b(\theta_i)\}/a(\psi) + c(y_i, \psi)]$. For likelihood-based models, the local-likelihood regression estimation, introduced by Tibshirani and Hastie (1987), is a nonparametric analogue of the parametric generalized linear model regression estimation. For this nonlinear fitting rule, numerically obtained via the Newton–Raphson iterative algorithm, the covariance penalty does not necessarily have an explicit form of expression. Nonetheless, $\widehat{\theta}_i$, the local polynomial likelihood estimate of the canonical parameter, satisfies $\widehat{\theta}_i \doteq \sum_{j=1}^n \mathcal{M}_h(i,j)\{g(\widehat{\mu}_j) + (y_j - \widehat{\mu}_j) g'(\widehat{\mu}_j)\}$, for a link function g and a smoother matrix $\mathcal{M}_h$. As I learned from Efron's article, the choice $q(\mu) = 2\{b(\theta) - \mu\theta\}$ gives $\widehat{\lambda}_i = \widehat{\theta}_i$. With this convenient result, it is readily seen that

$$\begin{aligned} \mathrm{cov}_i = \mathrm{cov}(\widehat{\theta}_i, y_i) &\doteq \mathcal{M}_h(i,i) \operatorname{var}(y_i) g'(\widehat{\mu}_i) \\ &= \mathcal{M}_h(i,i) a(\psi) b''(\widehat{\theta}_i) g'(\widehat{\mu}_i). \end{aligned}$$

For the commonly used canonical link function g, $\sum_{i=1}^n \mathrm{cov}_i \doteq a(\psi) \operatorname{tr}(\mathcal{M}_h)$. Again, Zhang (2003b) showed that $\operatorname{tr}(\mathcal{M}_h) \doteq d\{(p+1-a) + \mathcal{C}n/(n-1)\mathcal{K}(0)|\Omega|/h\}$ in a generalized smooth model and a generalized varying-coefficient model, implying the direct estimation method for the total covariance penalties by

$$\sum_{i=1}^n \widehat{\mathrm{cov}}_i = a(\widehat{\psi})\big\{(p+1-a) + \mathcal{C}n/(n-1)\mathcal{K}(0)|\Omega|/h\big\}. \quad (6)$$

For a Gaussian family, the empirical formula (6) reduces to (5). Among non-Gaussian outcomes, the Bernoulli-distributed binary responses and the Poisson-distributed count responses no longer carry in (6) the estimate, $a(\widehat{\psi})$, for the nuisance parameter. This makes the direct estimation further simplified.

4. NONPARAMETRIC MODEL SELECTION: APPLICATION OF COVARIANCE PENALTY METHOD

An important research problem in applications of nonparametric modeling techniques is the automatic selection of smoothing parameters. Essentially, this issue can be formulated as a nonparametric model selection problem: Choose the amount of smoothing that produces a nonparametric model with the minimum prediction error. Indeed, the arrival of Efron's article provides the theoretical basis for evaluating a wide variety of existing selection methods in the literature and broadens the scope of the covariance penalty method to more application fields in which nonparametric techniques have been under developed.

For illustration, I consider the bandwidth parameter h in the context of local polynomial model-fitting method. Hereafter, $\widehat{\mu}_{h,i}$ and $\widehat{\lambda}_{h,i}$ are used for $\widehat{\mu}_i$ and $\widehat{\lambda}_i$, respectively. According to (2), the optimal data-driven bandwidth selector $\widehat{h}^{\mathrm{CP}}$, based on the covariance penalty method, minimizes with respect to $h > 0$ the total prediction error estimates,

$$\widehat{\mathrm{Err}}^{\mathrm{CP}}(h) = \sum_{i=1}^n Q(y_i, \widehat{\mu}_{h,i}) + 2 \sum_{i=1}^n \widehat{\mathrm{cov}}(\widehat{\lambda}_{h,i}, y_i). \quad (7)$$

1. For Gaussian responses, with the squared loss function, the bandwidth selector studied in Hurvich, Simonoff, and Tsai (1998) is asymptotically equivalent to the above $\widehat{h}^{\mathrm{CP}}$.
2. Currently, most of the existing methods for the optimal smoothing deal with metrical responses and there is a clear lack of methodology and scheme for smoothing non-Gaussian responses. With the flexible choice of error measures Q, Efron's article makes the optimal bandwidth selector, $\widehat{h}^{\mathrm{CP}}$, continue to be applicable to responses in the exponential families. For Q chosen to be deviance of the local polynomial likelihood estimates, it can also be shown that the EGCV-minimizing bandwidth selector (Zhang 2003b) is asymptotically equivalent to $\widehat{h}^{\mathrm{CP}}$. Further research along the line of (7) will be fruitful.
3. The covariance penalty method has an added advantage: A *locally* optimal bandwidth selector can easily be obtained via minimizing the sum of neighboring coordinate-wise prediction error estimates. The resulting selector is *spatially adaptive* and outperforms the *globally* optimal bandwidth selector, $\widehat{h}^{\mathrm{CP}}$, at locations of fitting points requiring varying amount of smoothing.

4.1 Correlated Data

Technological invention and information advancement have revolutionized scientific research and technological development. Many sophisticated datasets have recently been collected. Data types range from the brain functional magnetic resonance imaging data in biomedical study and neuroscience, traffic time series data in transportation management, to financial time series data in econometrics and finance. All these data share a common characteristic: The measurements are highly correlated time series data. Compared with the traditional parametric modeling techniques, statistical nonparametric modeling techniques for complex observational data will lead to considerable reduction of modeling bias and false positive rates.

However, compared with uncorrelated data, the likely presence of correlation effects poses more challenges to estimating the covariance penalties, in addition to developing nonparametric model-fitting techniques. The bootstrap estimation method needs to be used with care; similarly, the validity of the direct estimation method based on the total degrees of freedom may also call for reexamination. Regarding the nonparametric model selection problem, most smoothing parameter selection methods do not perform well to be adaptive to correlated errors (see Hart 1994; Opsomer, Wang, and Yang 2001). For the preceding bandwidth selector $\widehat{h}^{\mathrm{CP}}$, based on the covariance penalty method, the criterion function (7) may need to be modified to take into full account data dependencies.

ADDITIONAL REFERENCES

Buckley, M. J., Eagleson, G. K., and Silverman, B. W. (1988), "The Estimation of Residual Variance in Nonparametric Regression," *Biometrika*, 75, 189–200.

Cleveland, W., and Devlin, S. (1988), "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, 83, 596–610.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., and Zeger, S. (2002), *Analysis of Longitudinal Data* (2nd ed.), Oxford, U.K.: Oxford University Press.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, London: Chapman & Hall.

Gurland, J. (1967), "An Inequality Satisfied by the Expectations of the Reciprocal of a Random Variable," *The American Statistician*, 21, 24–25.

Hart, J. D. (1994), "Automated Kernel Smoothing of Dependent Data by Using Time Series Cross-Validation," *Journal of the Royal Statistical Society*, Ser. B, 56, 529–542.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of the Royal Statistical Society*, Ser. B, 60, 271–293.

Opsomer, J. D., Wang, Y., and Yang, Y. (2001), "Nonparametric Regression With Correlated Errors," *Statistical Science*, 16, 134–153.

Ramsay, J. O., and Silverman, B. W. (1997), *Functional Data Analysis*, New York: Springer-Verlag.

Tibshirani, R., and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559–567.

Ye, J. M. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Association*, 93, 120–131.

Zhang, C. M. (2003a), "Calibrating the Degrees of Freedom for Automatic Data Smoothing and Effective Curve Checking," *Journal of the American Statistical Association*, 98, 609–628.

——— (2003b), "Cross-Validated Local Likelihood Estimates in the Exponential Family," Technical Report 1082, University of Wisconsin, Dept. of Statistics.

Rejoinder

Bradley EFRON

Classical statistics as developed in the first half of the 20th century has two obvious deficiencies from the point of view of practical applications: an overreliance on the normal distribution and failure to account for model selection. The first of these was dealt with in the century's second half by nonparametrics, generalized linear models, and computer-intensive techniques such as the jackknife and bootstrap.

Model selection, the data-based choice among structural models of different dimensions, remains mostly *terra incognita* as far as statistical inference is concerned. This article aims at a small corner of the model selection problem, the assessment of predictive accuracy. Its main result is a Rao–Blackwell type of relationship between cross-validation and what I called "covariance penalties." The latter are shown to have better estimation properties at the expense of increased assumptions.

The assessment of predictive accuracy is a form of bias estimation: "err," the apparent error (1.1), is downward biased for the true predictive error. As usual the bias is of order only $O(1/n)$ compared to err. This makes for difficult and often unrealistic asymptotics, the $O(1/n)$ term disappearing too quickly for easy extrapolation from large-sample behavior. The Rao–Blackwell result (4.6) relies on just a simple algebraic identity, providing at least heuristic grounds for believing its small-sample applicability.

The discussants' comments brought home some defects in the article's presentation. My numerical examples, with the exception of remark B, failed to include model selection. Reasonably enough, Burman and also Denby, Landwehr, and Mallows question the efficacy of parametric bootstrap covariance estimates in a model selection situation. Numerical experimentation, admittedly of limited scope, is reassuring on this point.

Figure 12 concerns a cholesterol-lowering experiment described in figure 4 of Efron and Tibshirani (1998): 201 men in the experiment's control arm have been measured for drug-taking compliance and cholesterol decrease. Even though the "drug" is placebo, there is evidence of a positive regression, perhaps because the better compliers were also better dieters or exercisers. Polynomial predictors, of degrees 0 through 7, were fit to the data by ordinary least squares, with the quadratic regression, the solid curve in the left panel, being the clear C_p minimizer. The dashed curve is the ordinary least squares (OLS) seventh-degree polynomial fit.

The right panel displays coordinatewise degree-of-freedom estimates $\widehat{\mathrm{df}}_i = \widehat{\mathrm{cov}}_i/\widehat{\sigma}^2$ for the rule $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$ that selects among polynomial fits of degree 0 through 7 according to minimum C_p value. Parametric bootstrapping from $\widehat{\mathbf{f}} \sim \mathrm{N}(\widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2\mathbf{I})$ was used as in (2.14)–(2.15), with $\widehat{\sigma}^2$ obtained from the

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Theory and Methods
DOI 10.1198/016214504000000917