The Estimation of Prediction Error: Covariance Penalties and Cross-Validation

Bradley EFRON

Having constructed a data-based estimation rule, perhaps a logistic regression or a classification tree, the statistician would like to know its performance as a predictor of future cases. There are two main theories concerning prediction error: (1) penalty methods such as C_p, Akaike's information criterion, and Stein's unbiased risk estimate that depend on the covariance between data points and their corresponding predictions; and (2) cross-validation and related nonparametric bootstrap techniques. This article concerns the connection between the two theories. A Rao–Blackwell type of relation is derived in which nonparametric methods such as cross-validation are seen to be randomized versions of their covariance penalty counterparts. The model-based penalty methods offer substantially better accuracy, assuming that the model is believable.

KEY WORDS: C_p; Degrees of freedom; Nonparametric estimates; Parametric bootstrap; Rao–Blackwellization; SURE.

1. INTRODUCTION

Prediction problems arise in the following way: A model $m(\cdot)$, for example, an ordinary linear regression, is fit to some data $\mathbf{y}$ producing an estimate $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$; we wonder how well $\widehat{\boldsymbol{\mu}}$ will predict a future dataset independently generated from the same mechanism that produced $\mathbf{y}$. Two quite separate statistical theories are used to answer this question, *cross-validation* and what we will call *covariance penalties*, the latter including Mallow's C_p, Akaike's information criterion (AIC), and Stein's unbiased risk estimate (SURE). This article concerns the relationship between the two theories.

Figure 1 illustrates a simple prediction problem. Data (x_i, y_i) have been observed for 157 healthy volunteers, with x_i age and y_i a measure of total kidney function. The original goal was to study the decline in function over time, an important factor in kidney transplantation. The response variable y is a composite of several standard kidney function indices. A robust locally linear smoother "lowess($\mathbf{x}, \mathbf{y}, f = 1/3$)" ($f$ controlling the local window width) produces $\widehat{\boldsymbol{\mu}}$, the indicated regression curve, with sum of squared residuals

$$\mathrm{err} \equiv \sum_{i=1}^{157} (y_i - \widehat{\mu}_i)^2 = 495.1. \tag{1.1}$$

However err, the *apparent error*, is an optimistic assessment of how well the curve in Figure 1 would predict future y values because lowess has fit the curve to this particular dataset. How well can we expect $\widehat{\boldsymbol{\mu}}$ to perform on future data?

In this case the two theories give almost identical estimates of "Err," the true predictive error of $\widehat{\boldsymbol{\mu}}$: $\widehat{\mathrm{Err}} = 538.8$ for cross-validation and 538.3 for the covariance penalty method, 9% larger than (1.1). Sections 2–4 describe these calculations.

Cross-validation and the related bootstrap techniques of Efron (1983) are completely nonparametric. Covariance penalties, on the other hand, are model based, in this case relying on an estimated version of the standard additive homescadastic model $y_i = \mu_i + \epsilon_i$. Nonparametric methods are often preferable, but we will show that cross-validation pays a substantial price in terms of decreased estimating efficiency.

The model used to estimate a covariance penalty can also be employed to improve cross-validation, by averaging the cross-validation estimate of Err over a collection of the model's possible datasets. This is the subject of Section 4, where it is shown that the averaged cross-validation estimate nearly equals the covariance penalty estimate of Err. Roughly speaking, covariance penalties are a Rao–Blackwellized version of cross-validation (and also of the nonparametric bootstrap; Sec. 6) and as such enjoy increased efficiency for estimating prediction error.

Covariance penalties originated in the work of Mallows (1973), Akaike (1973), and Stein (1981). The formula was extended to generalized linear models in Efron (1986). Sections 2 and 3 broaden the penalty formula to include all models, and also develop it in a conditional setting that facilitates comparisons with cross-validation and the nonparametric bootstrap. Versions of the covariance penalty appear in Breiman (1992), Ye (1998), and Tibshirani and Knight (1999), with Ye's article being particularly relevant here.

2. C_p AND SURE

Covariance penalty methods first arose in the context where prediction error, say $Q(y_i, \widehat{\mu}_i)$, is measured by squared error

$$Q(y_i, \widehat{\mu}_i) = (y_i - \widehat{\mu}_i)^2. \tag{2.1}$$

Mallows (1973) considered prediction error for the homoscedastic model

$$\mathbf{y} \sim (\boldsymbol{\mu}, \sigma^2 \mathbf{I}), \tag{2.2}$$

the notation indicating that the components of $\mathbf{y}$ are uncorrelated, y_i having mean μ_i and variance σ^2.

Suppose that we are using a linear estimation rule

$$\widehat{\boldsymbol{\mu}} = M\mathbf{y}, \tag{2.3}$$

where M is an $n \times n$ matrix not depending on $\mathbf{y}$. Define

$$\mathrm{err}_i = (y_i - \widehat{\mu}_i)^2 \qquad \text{and} \qquad \mathrm{Err}_i = E_0(y_i^0 - \widehat{\mu}_i)^2, \tag{2.4}$$

the expectation "E_0" being over $y_i^0 \sim (\mu_i, \sigma^2)$ independent of $\mathbf{y}$, with $\widehat{\mu}_i$ held fixed. Mallows showed that

$$\widehat{\mathrm{Err}}_i \equiv \mathrm{err}_i + 2\sigma^2 M_{ii} \tag{2.5}$$

Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: *brad@stat.stanford.edu*). Author is grateful to Dr. Bryan D. Myers for bringing the kidney function estimation problem and data to author's attention, and for several helpful discussions.

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Theory and Methods
DOI 10.1198/016214504000000692

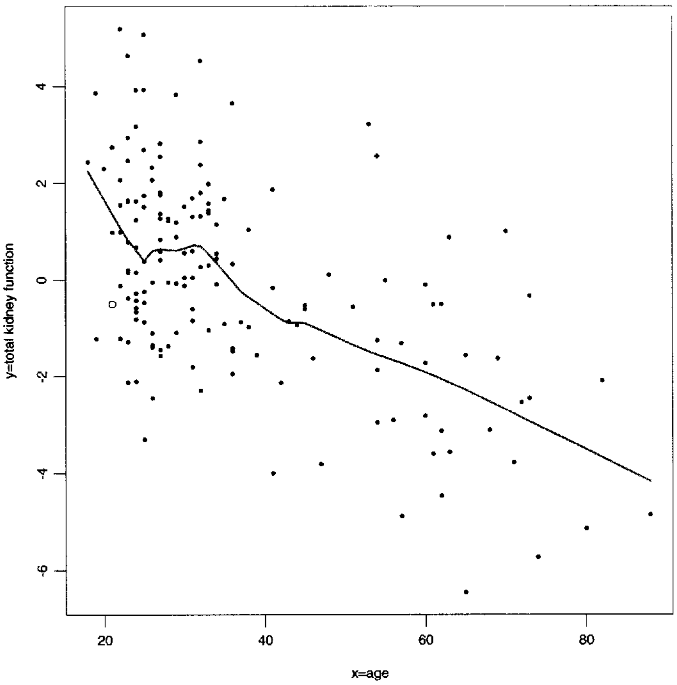

*Figure 1. Kidney Data: An Omnibus Measure of Kidney Function Plotted versus Age for n = 157 Healthy Volunteers. Fitted curve is lowess(**x**, **y**, f = 1/3); sum of squared residuals 495.1. How well can we expect this curve to predict future (x, y) pairs?*

is an unbiased estimator for the expectation of Err_i, leading to the C_p *formula* for estimating $\mathrm{Err} = \sum_{i=1}^n \mathrm{Err}_i$,

$$\widehat{\mathrm{Err}} = \mathrm{err} + 2\sigma^2 \, \mathrm{trace}(M), \qquad \mathrm{err} = \sum_{i=1}^{n} \mathrm{err}_i. \tag{2.6}$$

In practice, we usually need to replace σ^2 with an estimate $\widehat{\sigma}^2$ as in the examples that follow; see section 7 of Efron (1986).

Dropping the linearity assumption, let $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$ be any rule at all for estimating $\boldsymbol{\mu}$ from $\mathbf{y}$. Taking expectations in the identity

$$(y_i - \mu_i)^2 + (\mu_i - \widehat{\mu}_i)^2 \\ = (y_i - \widehat{\mu}_i)^2 + 2(\widehat{\mu}_i - \mu_i)(y_i - \mu_i), \tag{2.7}$$

and using $E(y_i - \mu_i)^2 = E_0(y_i^0 - \mu_i)^2$, gives a convenient expression for the expectation of Err_i, (2.4),

$$E\{\mathrm{Err}_i\} = E\{\mathrm{err}_i + 2\,\mathrm{cov}(\widehat{\mu}_i, y_i)\}. \tag{2.8}$$

Because $\mathrm{cov}(\widehat{\mu}_i, y_i)$ equals $\sigma^2 M_{ii}$ for a linear rule, (2.8) is seen to be a generalization of (2.5). In words, we must add a *covariance penalty* to the apparent error err_i in order to unbiasedly estimate Err_i.

Formula (2.8) is not directly applicable since $\mathrm{cov}(\widehat{\mu}_i, y_i)$ is not an observable statistic. Stein (1981) overcame this impediment in the *Gaussian case*

$$\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}) \tag{2.9}$$

by showing that

$$\mathrm{cov}_i = \sigma^2 E\{\partial \widehat{\mu}_i / \partial y_i\} \tag{2.10}$$

[assuming (2.9) and a differentiability condition on the mapping $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$]. Because $\partial \widehat{\mu}_i / \partial y_i$ *is* observable, this leads to Stein's unbiased risk estimate (SURE) for total prediction error,

$$\widehat{\mathrm{Err}} = \mathrm{err} + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \widehat{\mu}_i}{\partial y_i}. \tag{2.11}$$

In the linear case it is now common, as in Hastie and Tibshirani (1990), to define trace(M) as the *degrees of freedom* (df) of the rule $\widehat{\boldsymbol{\mu}} = M\mathbf{y}$. If we are in the usual regression or analysis of variance (ANOVA) situation, where M is a projection matrix, then trace(M) $= p$, the dimension of the projected space, agreeing with the usual df definition. As in Ye (1998), we can extend this definition to

$$\mathrm{df} = \sum_{i=1}^{n} \frac{\mathrm{cov}(\widehat{\mu}_i, y_i)}{\sigma^2} \tag{2.12}$$

for a general rule $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$.

Traditional applications of linear models try to keep df $\ll n$. Because $\sum_{i=1}^n M_{ii} = \mathrm{df}$, this can be interpreted as $M_{ii} = \mathrm{O}(1/n)$ in reasonable experimental designs. Similarly, the informal order of magnitude calculations that follow assume

$$\mathrm{cov}(\widehat{\mu}_i, y_i) = \mathrm{O}(1/n). \tag{2.13}$$

This might better be stated as "O(df/n)," the crucial ingredient for the asymptotics being a small value of df/n.

The bootstrap, or more exactly the *parametric bootstrap*, suggests a direct way of estimating the covariance penalty $\mathrm{cov}(\widehat{\mu}_i, y_i)$. Let $\widehat{\mathbf{f}}$ be an assumed density for $\mathbf{y}$. In the Gaussian case we might take $\widehat{\mathbf{f}} = \mathrm{N}(\widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2 \mathbf{I})$ with $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$ and $\widehat{\sigma}^2$ obtained from the residuals of some "big" model presumed to have negligible bias. We then generate a large number "B" of simulated observations and estimates from $\widehat{\mathbf{f}}$,

$$\widehat{\mathbf{f}} \rightarrow \mathbf{y}^* \rightarrow \widehat{\boldsymbol{\mu}}^* = m(\mathbf{y}^*), \tag{2.14}$$

and estimate $\mathrm{cov}_i = \mathrm{cov}(\widehat{\mu}_i, y_i)$ from the observed bootstrap covariance, say

$$\widehat{\mathrm{cov}}_i = \sum_{b=1}^{B} \widehat{\mu}_i^{*b} (y_i^{*b} - y_i^{*\cdot})/(B-1), \qquad y_i^{*\cdot} = \sum_{b} \frac{y_i^{*b}}{B}, \tag{2.15}$$

leading to the Err estimate

$$\widehat{\mathrm{Err}} = \mathrm{err} + 2 \sum_{i=1}^{n} \widehat{\mathrm{cov}}_i. \tag{2.16}$$

Both Breiman (1992) and Ye (1998) proposed variations on (2.14) intended to improve the efficiency of the bootstrap estimation procedure; see Remark A.

Figure 2 displays SURE and parametric bootstrap estimates of the coordinatewise degrees of freedom df_i for the kidney data. The two sets of estimates $\partial \widehat{\mu}_i / \partial y_i$ and $\widehat{\mathrm{cov}}_i / \widehat{\sigma}^2$ are plotted versus age_i, vividly demonstrating the decreased stability of the lowess fitting process near the extremes of the age scale. The resampling algorithm (2.14) employed

$$\mathbf{y}^* = \widehat{\boldsymbol{\mu}} + \boldsymbol{\epsilon}^*, \tag{2.17}$$

with the components of $\boldsymbol{\epsilon}^*$ a random sample of size n from the empirical distribution of the observed residuals $\widehat{\epsilon}_j = y_j - \widehat{\mu}_j$

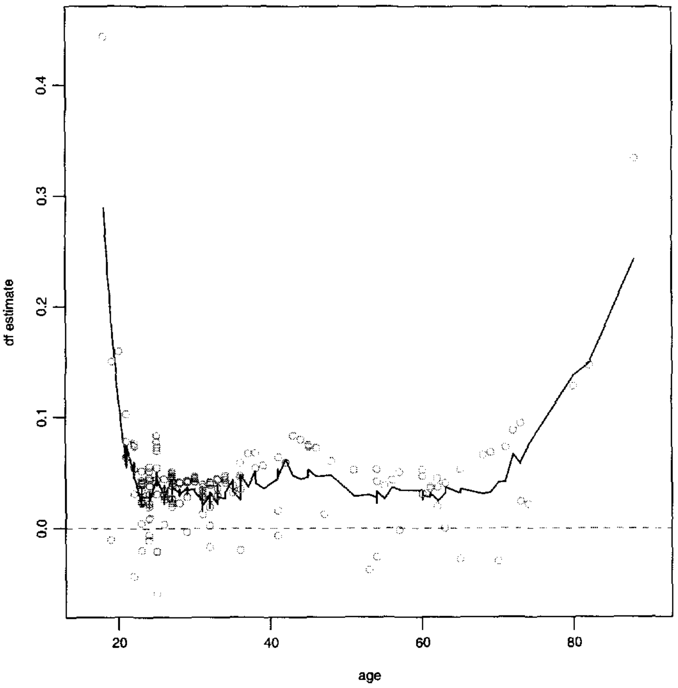

Figure 2. Coordinatewise Degrees of Freedom for Lowess Fit of Figure 1, Plotted versus Age. Open circles, SURE estimate $\widehat{df}_i = \partial\widehat{\mu}_i/\partial y_i$; solid line, parametric bootstrap estimates $\widehat{\mathrm{cov}}_i/\widehat{\sigma}^2$, (2.14)–(2.15), $B = 1{,}000$. Total df estimates 6.85 (SURE) and 6.67 (parametric bootstrap). The coordinatewise bootstrap estimates are noticeably less noisy.

(having $\widehat{\sigma}^2 = 3.17$). Almost identical results were obtained taking $\widehat{\epsilon}_i^* \sim \mathrm{N}(0, \widehat{\sigma}^2)$. The lowess estimator was chosen here because it is nonlinear and unsmooth, making the df calculations more challenging.

The two methods gave similar estimates for the total degrees of freedom $\mathrm{df} = \sum \mathrm{df}_i$: 6.85 using SURE and 6.67 ± .30 with the bootstrap, the ± value indicating simulation error, estimated as

$$\frac{n}{\widehat{\sigma}^2}\left[\frac{\sum(C^{*b} - C^{*\cdot})^2}{B(B-1)}\right]^{1/2},$$

$$C^{*b} = \sum_{i=1}^{n} \frac{\widehat{\mu}_i^{*b}(y_i^{*b} - y_i^{*\cdot})}{n}, \quad C^{*\cdot} = \sum \frac{C^{*b}}{B}. \tag{2.18}$$

However, the componentwise bootstrap estimates are noticeably less noisy, having standard deviation 2.5 times smaller than the SURE values over the range $20 \le \text{age} \le 75$.

Remark A. It is not necessary that the bootstrap model $\widehat{\mathbf{f}}$ in (2.14) be based on $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$. The solid curve in Figure 2 was recomputed starting from the bigger model (more degrees of freedom) $\widehat{\mathbf{f}} = \mathrm{N}(\widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2\mathbf{I})$, with $\widehat{\boldsymbol{\mu}}$ the fit from lowess($\mathbf{x}, \mathbf{y}$, $f = 1/6$), but still using $f = 1/3$ for $m(\mathbf{y}^*)$ at the final step of (2.14). This gave almost the same results as in Figure 2.

The ultimate "bigger model" is

$$\widehat{\mathbf{f}} = \mathrm{N}(\mathbf{y}, \widehat{\sigma}^2\mathbf{I}). \tag{2.19}$$

This choice, which is the one made in Ye (1998), Shen, Huang, and Ye (2002), Shen and Ye (2002), and Breiman (1992), has the advantage of not requiring model assumptions. It pays for this luxury with increased estimation error: the $\widehat{\mathrm{df}}_i$ plot looks more like the open circles than the solid line in Figure 2. The author prefers checking the $\widehat{\mathrm{df}}_i$ estimates against moderately bigger models, such as lowess($\mathbf{x}, \mathbf{y}$, 1/6), rather than going all the way to (2.19); see Remark C.

In fact, the exact choice of $\widehat{\mathbf{f}}$ is often quite unimportant. Notice that $\mathrm{df}_i \equiv \mathrm{cov}(\widehat{\mu}_i, y_i)/\sigma^2$ is the linear regression coefficient of $\widehat{\mu}_i$ on y_i. If the regression function $E\{\widehat{\mu}_i|y_i\}$ is roughly linear in y_i, then its slope can be estimated by a wide variety of devices. Algorithm 1 of Ye (1998) takes $\mathbf{y}^*$ in (2.14) from a shrunken version of (2.19),

$$\mathbf{y}^* \sim \mathrm{N}(\mathbf{y}, c\widehat{\sigma}^2\mathbf{I}), \tag{2.20}$$

with c a constant between .6 and 1, and estimates df_i by the linear regression coefficient of $\widehat{\mu}_i$ on y_i^*. Breiman's "little bootstrap" (1992) employs a related technique, the "little" referring to using $c < 1$ in (2.20), and winds up recommending c between .6 and .8 (though $c = 1$ gave slightly superior accuracy in his simulation experiments). Shen and Ye (2002) used an equivalent form of covariance estimation, with $c = .5$.

Remark B. The parametric bootstrap algorithm (2.14)–(2.15) can also be used to assess the *difference* between fits obtained from two models, say Model A and Model B. We will think of A as the smaller of the two, that is, the one with fewer degrees of freedom, though this is not essential. The estimated difference of prediction error is

$$\Delta\widehat{\mathrm{Err}} = \Delta\mathrm{err} + 2\sum_{i=1}^{n} \widehat{\mathrm{cov}}(\Delta\widehat{\mu}_i^*, y_i^*), \tag{2.21}$$

Δ denoting "Model A minus Model B."

Calculation (2.21) was carried out for the kidney data with lowess($\mathbf{x}, \mathbf{y}, f = 2/3$) for Model A and lowess($\mathbf{x}, \mathbf{y}, f = 1/3$) for Model B; $\Delta\mathrm{err} = 498.5 - 495.1 = 3.4$. With $\widehat{\mathbf{f}}$ in (2.14) estimated from Model A, 1,000 parametric bootstraps (each requiring both model fits) gave -18.4 for the second term in (2.21), so

$$\Delta\widehat{\mathrm{Err}} = 3.4 - 18.4 = -15.0,$$

favoring the smaller Model A.

The 1,000 pairs of bootstrap fits $\widehat{\boldsymbol{\mu}}(A)^*$ and $\widehat{\boldsymbol{\mu}}(B)^*$ contain useful information, beyond evaluating the second term of (2.21). Figure 3 displays the thousand values of

$$\Delta\widehat{\mathrm{Err}}^* = \Delta\mathrm{err}^* - 18.4. \tag{2.22}$$

This can be considered as a null hypothesis distribution for testing "Model B is no improvement over Model A." In this case the observed $\Delta\widehat{\mathrm{Err}}$ falls in the lower part of the distribution, but for a larger observed value, say $\Delta\widehat{\mathrm{Err}} = 20.0$, we might use the histogram to assign the approximate p value $\#\{\Delta\widehat{\mathrm{Err}}^* > \Delta\widehat{\mathrm{Err}}\}/1{,}000$.

This calculation ignores the fact that the penalty -18.4 in (2.22) is itself variable. For linear models the penalty is a constant, obviating concern. In general, the penalty term is an order of magnitude smaller than $\Delta\mathrm{err}$, and not likely to contribute much to the bootstrap variability of $\Delta\widehat{\mathrm{Err}}^*$. This was checked here using a second level of bootstrapping, which made very little difference to Figure 3.

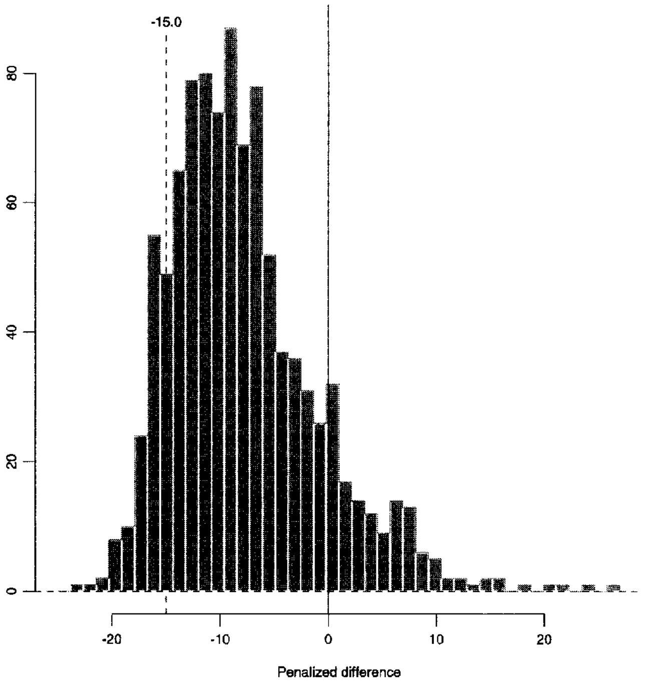

Figure 3. 1,000 Bootstrap Replications of $\Delta\widehat{\mathrm{Err}}^*$ for the Difference Between lowess(**x**, **y**, 2/3) and lowess(**x**, **y**, 1/3), Kidney Data. The point estimate $\Delta\widehat{\mathrm{Err}} = -15.0$ is in the lower part of the histogram.

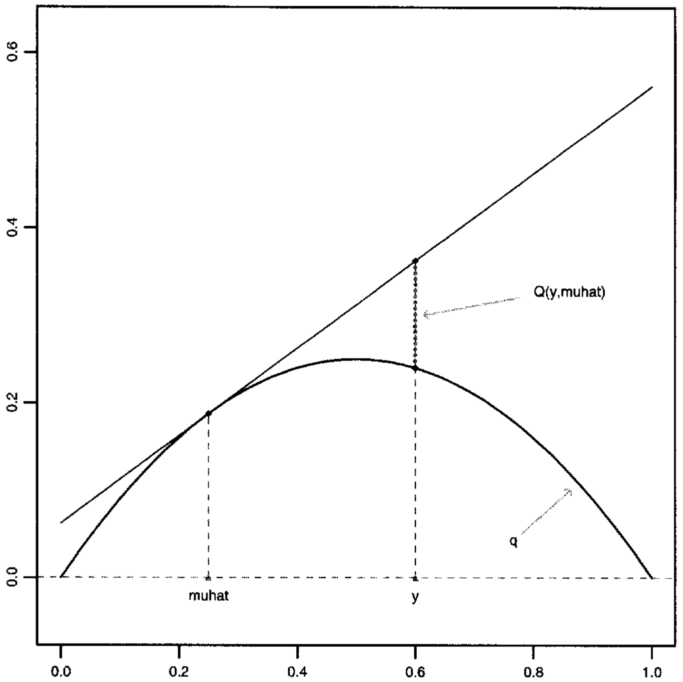

Figure 4. Tangency Construction (3.1) for General Error Measure $Q(y,\widehat{\mu})$; $q(\cdot)$ Is an Arbitrary Concave Function. The illustrated case has $q(\mu) = \mu(1-\mu)$ and $Q(y,\widehat{\mu}) = (y-\widehat{\mu})^2$.

Remark C. The parametric bootstrap estimate (2.14)–(2.15), unlike SURE, does not depend on $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$ being differentiable or even continuous. A simulation experiment was run taking the true model for the diabetes data to be $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, with $\sigma^2 = 3.17$ and $\boldsymbol{\mu}$ the lowess($\mathbf{x}, \mathbf{y}, f = 1/6$) fit, a noticeably rougher curve than that of Figure 1. A discontinuous adaptive estimation rule $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$ was used: Polynomial regressions of y on x for powers of x from 0 to 7 were fit, with the one having the minimum C_p value selected to be $\widehat{\boldsymbol{\mu}}$.

Because this is a simulation experiment, we can estimate the true expected difference between Err and err, (2.8): 1,000 simulations of $\mathbf{y}$ gave

$$E\{\mathrm{Err} - \mathrm{err}\} = 33.1 \pm 2.02. \tag{2.23}$$

The parametric bootstrap estimate (2.14)–(2.16) worked well here, 1,000 replications of $\mathbf{y} \sim N(\widehat{\boldsymbol{\mu}}, \sigma^2\mathbf{I})$, with $\widehat{\boldsymbol{\mu}}$ from lowess($\mathbf{x}, \mathbf{y}, f = 1/3$), yielding

$$\widehat{\mathrm{Err}} - \mathrm{err} = 31.4 \pm 2.85. \tag{2.24}$$

In contrast, bootstrapping from $\mathbf{y}^* \sim N(\mathbf{y}, \sigma^2\mathbf{I})$ as in (2.19) gave 14.6 ± 1.82, badly underestimating the true difference 33.1. Starting with the true $\boldsymbol{\mu}$ equal to the seventh-degree polynomial fit gave nearly the same results as (2.23).

3. GENERAL COVARIANCE PENALTIES

The covariance penalty theory of Section 2 can be generalized beyond squared error to a wide class of error measures. The *q class* of error measures (Efron 1986), begins with any concave function $q(\cdot)$ of a real-valued argument. $Q(y,\widehat{\mu})$, the assessed error for outcome y given prediction $\widehat{\mu}$, is then defined to be

$$Q(y,\widehat{\mu}) = q(\widehat{\mu}) + \dot{q}(\widehat{\mu})(y-\widehat{\mu}) - q(y) \qquad \left[\dot{q}(\widehat{\mu}) = dq/d\mu|_{\widehat{\mu}}\right]. \tag{3.1}$$

$Q(y,\widehat{\mu})$ is the tangency function to $q(\cdot)$, as illustrated in Figure 4; (3.1) is a familiar construct in convex analysis (Rockafellar 1970). The choice $q(\mu) = \mu(1-\mu)$ gives squared error, $Q(y,\widehat{\mu}) = (y-\widehat{\mu})^2$.

Our examples will include the *Bernoulli case* $\mathbf{y} \sim \mathrm{Be}(\boldsymbol{\mu})$, where we have n independent observations y_i,

$$y_i = \begin{cases} 1, & \text{probability } \mu_i \\ 0, & \text{probability } 1-\mu_i \end{cases} \qquad \text{for } \mu_i \in [0,1]. \tag{3.2}$$

Two common error functions used for Bernoulli observations are *counting error*

$$q(\mu) = \min(\mu, 1-\mu)$$
$$\to Q(y,\mu) = \begin{cases} 0 & \text{if } y, \mu \text{ on same side of } 1/2 \\ 1 & \text{if } y, \mu \text{ on different sides of } 1/2 \end{cases} \tag{3.3}$$

(see Remark F) and *binomial deviance*

$$q(\mu) = -2[\mu\log(\mu) + (1-\mu)\log(1-\mu)]$$
$$\to Q(y,\mu) = \begin{cases} -2\log\mu & \text{if } y = 1 \\ -2\log(1-\mu) & \text{if } y = 0. \end{cases} \tag{3.4}$$

By a linear transformation we can always make

$$q(0) = q(1) = 0, \tag{3.5}$$

which is convenient for Bernoulli calculations.

We assume that some unknown probability mechanism $\mathbf{f}$ has given the observed data $\mathbf{y}$, from which we estimate the expectation vector $\boldsymbol{\mu} = E_{\mathbf{f}}\{\mathbf{y}\}$ according to the rule $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$,

$$\mathbf{f} \to \mathbf{y} \to \widehat{\boldsymbol{\mu}} = m(\mathbf{y}). \tag{3.6}$$

Total error will be assessed by summing the component errors,

$$Q(\mathbf{y},\widehat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} Q(y_i,\widehat{\mu}_i). \tag{3.7}$$

The following definitions lead to a general version of the C_p formula (2.8). Letting

$$\text{err}_i = Q(y_i, \widehat{\mu}_i) \quad \text{and} \quad \text{Err}_i = E_0\{Q(y_i^0, \widehat{\mu}_i)\} \tag{3.8}$$

as in (2.4), with $\widehat{\mu}_i$ fixed in the expectation and y_i^0 from an independent copy of $\mathbf{y}$, define the

$$\text{Optimism:} \quad O_i = O_i(\mathbf{f}, \mathbf{y}) = \text{Err}_i - \text{err}_i \tag{3.9}$$

and

$$\text{Expected optimism:} \quad \Omega_i = \Omega(\mathbf{f}) = E_{\mathbf{f}}\{O_i(\mathbf{f}, \mathbf{y})\}. \tag{3.10}$$

Finally, let

$$\widehat{\lambda}_i = -\dot{q}(\widehat{\mu}_i)/2. \tag{3.11}$$

For $q(\mu) = \mu(1-\mu)$, the squared error case, $\widehat{\lambda}_i = \widehat{\mu}_i - 1/2$; for counting error (3.3), $\widehat{\lambda}_i = -1$ or 1 as $\widehat{\mu}_i$ is less or greater than $1/2$; for binomial deviance (3.4),

$$\widehat{\lambda}_i = \log(\widehat{\mu}_i/(1-\widehat{\mu}_i)), \tag{3.12}$$

the logit parameter. [If $Q(y, \widehat{\mu})$ is the deviance function for any exponential family, then $\widehat{\lambda}$ is the corresponding natural parameter; see sec. 6 of Efron 1986.]

Optimism Theorem 1. For error measure $Q(y, \widehat{\mu})$, (3.1), we have

$$E\{\text{Err}_i\} = E\{\text{err}_i + \Omega_i\}, \tag{3.13}$$

where

$$\Omega_i = 2\,\text{cov}(\widehat{\lambda}_i, y_i), \tag{3.14}$$

the expectations and covariance being with respect to $\mathbf{f}$, (3.6).

Proof. $\text{Err}_i = \text{err}_i + O_i$ by definition, immediately giving (3.13). From (3.1) we calculate

$$\begin{aligned} \text{Err}_i &= q(\widehat{\mu}_i) + \dot{q}(\widehat{\mu}_i)(\mu_i - \widehat{\mu}_i) - E\{q(y_i^0)\}, \\ \text{err}_i &= q(\widehat{\mu}_i) + \dot{q}(\widehat{\mu}_i)(y_i - \widehat{\mu}_i) - q(y_i) \end{aligned} \tag{3.15}$$

and so, from (3.9)–(3.11),

$$O_i = 2\widehat{\lambda}_i(y_i - \mu_i) + q(y_i) - E\{q(y_i^0)\}. \tag{3.16}$$

Because $E\{q(y_i^0)\} = E\{q(y_i)\}$, y_i^0 being a copy of y_i, taking expectations in (3.16) verifies (3.14).

The optimism theorem generalizes Stein's result for squared error, (2.8), to the q class of error measures. It was developed by Efron (1986) in a generalized linear model (GLM) context but as verified here it applies to any probability mechanism $\mathbf{f} \rightarrow \mathbf{y}$. Even independence is not required among the components of $\mathbf{y}$, though it is convenient to assume independence in the conditional covariance computations that follow.

Parametric bootstrap computations can be used to estimate the penalty $\Omega_i = 2\,\text{cov}(\widehat{\lambda}_i, y_i)$ as in (2.14), the only change being the substitution of $\widehat{\lambda}_i^* = -\dot{q}(\widehat{\mu}_i^*)/2$ for $\widehat{\mu}_i^*$ in (2.15):

$$\widehat{\text{cov}}_i = \sum_{i=1}^{B} \widehat{\lambda}_i^{*b}(y_i^{*b} - y_i^{*\cdot})/(B-1). \tag{3.17}$$

Method (3.17) was suggested in remark J of Efron (1986). Shen et al. (2002), working with deviance in exponential families, employed a "shrunken" version of (3.17), as in (2.20).

Section 4 relates covariance penalties to cross-validation. In doing so it helps to work with a conditional version of cov_i. Let $\mathbf{y}_{(i)}$ indicate the data vector with y_i deleted,

$$\mathbf{y}_{(i)} = (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n), \tag{3.18}$$

and define the *conditional covariance*

$$\text{cov}_{(i)} = E\{\widehat{\lambda}_i \cdot (y_i - \mu_i) | \mathbf{y}_{(i)}\} \equiv E_{(i)}\{\widehat{\lambda}_i \cdot (y_i - \mu_i)\}, \tag{3.19}$$

$E_{(i)}$ indicating $E\{\cdot|\mathbf{y}_{(i)}\}$; likewise $\Omega_{(i)} = 2\,\text{cov}_{(i)}$. In situation (2.1)–(2.3) $\text{cov}_{(i)} = \text{cov}_i = \sigma^2 M_{ii}$, but, in general, we only have $E\{\text{cov}_{(i)}\} = \text{cov}_i$. The conditional version of (3.13),

$$E_{(i)}\{\text{Err}_i\} = E_{(i)}\{\text{err}_i + \Omega_{(i)}\} \tag{3.20}$$

is a more refined statement of the optimism theorem. The SURE formula (2.10) also applies conditionally, $\text{cov}_{(i)} = \sigma^2 E_{(i)}\{\partial\widehat{\mu}_i/\partial y_i\}$, assuming normality (2.9).

Figure 5 illustrates conditional and unconditional covariance calculations for subject $i = 93$ of the kidney study (the open circle in Fig. 1). Here we have used squared error and the Gaussian model $\mathbf{y}^* \sim N(\widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2\mathbf{I})$, $\widehat{\sigma}^2 = 3.17$, with $\widehat{\boldsymbol{\mu}} = \text{lowess}(\mathbf{x}, \mathbf{y}, 1/3)$. The conditional and unconditional covariances are nearly the same, $\widehat{\text{cov}}_{(i)} = .221$ versus $\widehat{\text{cov}}_i = .218$, but the dependence of $\widehat{\mu}_i^*$ on y_i^* is much clearer conditioning on $\mathbf{y}_{(i)}$.

The conditional approach is computationally expensive: We would need to repeat the conditional resampling procedure of Figure 5 separately for each of the n cases, whereas a single set of unconditional resamples suffices for all n. Here we will use the conditional covariances (3.19) mainly for theoretical purposes. The less expensive unconditional approach performed well in all of our examples.

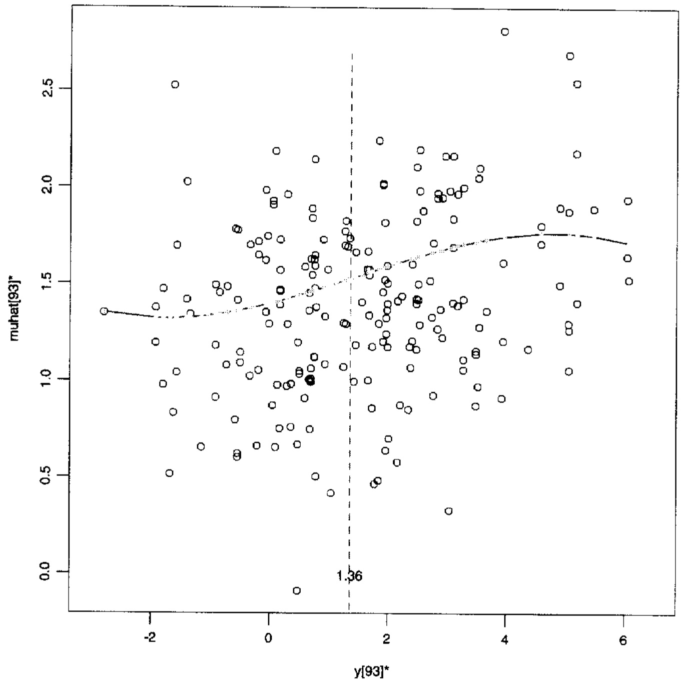

Figure 5. Conditional and Unconditional Covariance Calculations for Subject $i = 93$, Kidney Study. Open circles: 200 pairs $(y_i^*, \widehat{\mu}_i^*)$, unconditional resamples $\mathbf{y}^* \sim N(\widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2\mathbf{I})$; $\widehat{\text{cov}}_i = .218$; Dots: 100 conditional resamples, $y_i^* \sim N(\widehat{\mu}_i, \widehat{\sigma}^2)$, $\mathbf{y}_{(i)}$ fixed; $\widehat{\text{cov}}_{(i)} = .221$. Vertical line at $\widehat{\mu}_{93} = 1.36$.

There is, however, one situation where the conditional covariances are easily computed: the Bernoulli case $\mathbf{y} \sim \mathrm{Be}(\boldsymbol{\mu})$. In this situation it is easy to see that

$$\mathrm{cov}_{(i)} = \mu_i(1-\mu_i)\big[\widehat{\lambda}_i(\mathbf{y}_{(i)}, 1) - \widehat{\lambda}_i(\mathbf{y}_{(i)}, 0)\big], \qquad (3.21)$$

the notation indicating the two possible values of $\widehat{\lambda}_i$ with $\mathbf{y}_{(i)}$ fixed and y_i either 1 or 0. This leads to estimates

$$\widehat{\mathrm{cov}}_{(i)} = \widehat{\mu}_i(1-\widehat{\mu}_i)\big[\widehat{\lambda}_i(\mathbf{y}_{(i)}, 1) - \widehat{\lambda}_i(\mathbf{y}_{(i)}, 0)\big]. \qquad (3.22)$$

Calculating $\widehat{\mathrm{cov}}_{(i)}$ for $i = 1, 2, \ldots, n$, requires only n recomputations of $m(\cdot)$, one for each i, the same number as for cross-validation. For reasons discussed next, (3.22) will be termed the *Steinian.*

There is no general equivalent to the Gaussian SURE formula (2.10), that is, an unbiased estimator for $\mathrm{cov}_{(i)}$. However, a useful approximation can be derived as follows. Let $t_i(y_i^*) = \widehat{\lambda}_i(y_{(i)}, y_i^*)$ indicate $\widehat{\lambda}_i^*$ as a function of y_i^*, with $\mathbf{y}_{(i)}$ fixed, and denote $\dot{t}_i = \partial t_i(y_i^*)/\partial y_i^*|_{\widehat{\mu}_i}$; in Figure 5, $\dot{t}_i$ is the slope of the solid curve as it crosses the vertical line. Suppose y_i^* has bootstrap mean and variance $(\widehat{\mu}_i, \widehat{V}_i)$. Taylor series yield a simple approximation for $\widehat{\mathrm{cov}}_{(i)}$,

$$\begin{aligned}\widehat{\mathrm{cov}}_{(i)} &= E_{(i)}\{\widehat{\lambda}_i \cdot (y_i^* - \widehat{\mu}_i)\}\\ &\doteq E_{(i)}\big\{[t_i(\widehat{\mu}_i) + \dot{t}_i \cdot (y_i^* - \widehat{\mu}_i)](y_i^* - \widehat{\mu}_i)\big\}\\ &= \widehat{V}_i \dot{t}_i, \end{aligned} \qquad (3.23)$$

only y_i^* being random here. The Steinian (3.22) is a discretized version of (3.23), applied to the Bernoulli case, which has $\widehat{V}_i = \widehat{\mu}_i(1-\widehat{\mu}_i)$.

If $Q(y, \widehat{\mu})$ is the deviance function for a one-parameter exponential family, then λ_i is the natural parameter and $d\widehat{\lambda}_i/d\widehat{\mu}_i = 1/\widehat{V}_i$. Therefore

$$\widehat{\mathrm{cov}}_{(i)} \doteq \widehat{V}_i \frac{\partial \widehat{\lambda}_i}{\partial y_i^*}\bigg|_{\widehat{\mu}_i} = \widehat{V}_i \frac{1}{\widehat{V}_i} \frac{\partial \widehat{\mu}_i}{\partial y_i^*}\bigg|_{\widehat{\mu}_i} = \frac{\partial \widehat{\mu}_i}{\partial y_i^*}\bigg|_{\widehat{\mu}_i}. \qquad (3.24)$$

This is a *centralized* version of the SURE estimate, where now $\partial\widehat{\mu}_i/\partial y_i$ is evaluated at $\widehat{\mu}_i$ instead of y_i. [In the exponential family representation of the Gaussian case (2.9), $Q(y, \widehat{\mu}) = (y - \widehat{\mu})^2/\sigma^2$, so the factor σ^2 in (2.10) has been absorbed into Q in (3.24).]

Remark D. The centralized version of SURE in (3.24) gives the correct total degrees of freedom for maximum likelihood estimation in a p-parameter generalized linear model or, more generally, in a p-parameter curved exponential family (Efron 1975):

$$\sum_{i=1}^{n} \frac{\partial \widehat{\mu}_i}{\partial y_i}\bigg|_{\mathbf{y}=\widehat{\boldsymbol{\mu}}} = p. \qquad (3.25)$$

The usual uncentralized version of SURE does not satisfy (3.25) in curved families.

Using deviance error and maximum likelihood estimation in a curved exponential family makes $\mathrm{err}_i = -2\log f_{\widehat{\mu}_i}(y_i) + \text{constant}$. Combining (3.14), (3.24), and (3.25) gives

$$\widehat{\mathrm{Err}} \doteq -2\bigg[\sum_i \log f_{\widehat{\mu}_i}(y_i) - p + \text{constant}\bigg]. \qquad (3.26)$$

Choosing among competing models by minimizing $\widehat{\mathrm{Err}}$ is equivalent to maximizing the penalized likelihood $\sum \log f_{\widehat{\mu}_i}(y_i) - p$, which is *Akaike's information criterion* (AIC). These results generalize those for GLM's in section 6 of Efron (1986) and will not be verified here.

Remark E. It is easy to show that the true prediction error Err_i, (3.8), satisfies

$$\mathrm{Err}_i = Q(\mu_i, \widehat{\mu}_i) + D(\mu_i) \qquad \big[D(\mu_i) \equiv E\{Q(y_i, \mu_i)\}\big]. \quad (3.27)$$

For squared error this reduces to the familiar result $E_0(y_i^0 - \widehat{\mu}_i)^2 = (\widehat{\mu}_i - \mu_i)^2 + \sigma^2$. In the Bernoulli case (3.2), $D(\mu_i) = q(\mu_i)$ and the basic result (3.16) can be simplified to

$$\text{Bernoulli case:} \quad O_i = 2\widehat{\lambda}_i(y_i - \mu_i), \qquad (3.28)$$

using (3.5).

Remark F. The q class includes an *asymmetric* version of counting error (3.3) that allows the decision boundary to be at a point π_1 in (0, 1) other than 1/2. Letting $\pi_0 = 1 - \pi_1$ and $\rho = (\pi_0/\pi_1)^{1/2}$,

$$q(\mu) = \min\left\{\rho\mu, \frac{1}{\rho}(1-\mu)\right\}$$
$$\rightarrow Q(y, \widehat{\mu}) = \begin{cases} 0 & \text{if } y, \widehat{\mu} \text{ same side of } \pi_1 \\ \rho & \text{if } y = 1 \text{ and } \widehat{\mu} \le \pi_1 \\ \dfrac{1}{\rho} & \text{if } y = 0 \text{ and } \widehat{\mu} > \pi_1. \end{cases} \qquad (3.29)$$

Now $Q(1, 0)/Q(0, 1) = \pi_0/\pi_1$. This is the appropriate loss structure for a simple hypothesis-testing situation in which we want to compensate for unequal prior sampling probabilities.

4. THE RELATIONSHIP WITH CROSS-VALIDATION

Cross-validation is the most widely used error prediction technique. This section relates cross-validation to covariance penalties, more exactly to conditional parametric bootstrap covariance penalties. A Rao–Blackwell type of relationship will be developed: If we average cross-validation estimates across the bootstrap datasets used to calculate the conditional covariances, then we get, to a good approximation, the covariance penalty. The implication is that covariance penalties are more accurate than cross-validation, assuming of course that we trust the parametric bootstrap model. A similar conclusion is reached in Shen et al. (2002).

The cross-validation estimate of prediction error for coordinate i is

$$\widetilde{\mathrm{Err}}_i = Q(y_i, \widetilde{\mu}_i), \qquad (4.1)$$

where $\widetilde{\mu}_i$ is the ith coordinate of the estimate of $\boldsymbol{\mu}$ based on the deleted dataset $\mathbf{y}_{(i)} = (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$, say

$$\widetilde{\mu}_i = m(\mathbf{y}_{(i)})_i \qquad (4.2)$$

(see Remark H). Equivalently, cross-validation estimates the optimism $O_i = \mathrm{Err}_i - \mathrm{err}_i$ by

$$\widetilde{O}_i = \widetilde{\mathrm{Err}}_i - \mathrm{err}_i = Q(y_i, \widetilde{\mu}_i) - Q(y_i, \widehat{\mu}_i). \qquad (4.3)$$

Lemma. Letting $\widetilde{\lambda}_i = -\dot{q}(\widetilde{\mu}_i)/2$ and $\widehat{\lambda}_i = -\dot{q}(\widehat{\mu}_i)/2$ as in (3.11),

$$\widetilde{O}_i = 2(\widehat{\lambda}_i - \widetilde{\lambda}_i)(y_i - \mu_i) - Q(\widetilde{\mu}_i, \widehat{\mu}_i) - 2(\widehat{\lambda}_i - \widetilde{\lambda}_i)(\widetilde{\mu}_i - \mu_i). \tag{4.4}$$

This is verified directly from definition (3.1).

The lemma helps connect cross-validation with the conditional covariance penalties of (3.19)–(3.20). Cross-validation itself is conditional in the sense that $\mathbf{y}_{(i)}$ is fixed in the calculation of $\widetilde{O}_i$, so it is reasonable to suspect some sort of connection. Suppose that we estimate $\text{cov}_{(i)}$ by bootstrap sampling as in (3.17) but now with $\mathbf{y}_{(i)}$ fixed and only y_i^* random, say with density $\widetilde{f}_i$. The form of (4.4) makes it especially convenient for y_i^* to have conditional expectation $\widetilde{\mu}_i$ (rather than the obvious choice $\widehat{\mu}_i$) which we denote by

$$\widetilde{E}_{(i)}\{y_i^*\} \equiv E_{\tilde{f}_i}\left\{y_i^* \middle| \mathbf{y}_{(i)}\right\} = \widetilde{\mu}_i. \tag{4.5}$$

In a Bernoulli situation we would take $y_i^* \sim \text{Be}(\widetilde{\mu}_i)$.

Denote the bootstrap versions of μ_i and λ_i as $\widehat{\mu}_i^* = m(\mathbf{y}_{(i)}, y_i^*)$ and $\widehat{\lambda}_i^* = -\dot{q}(\widehat{\mu}_i^*)/2$.

Theorem 1. With $y_i^* \sim \widetilde{f}_i$ satisfying (4.5), and $\mathbf{y}_{(i)}$ fixed,

$$\widetilde{E}_{(i)}\{\widetilde{O}_i^*\} = 2\widehat{\text{cov}}_{(i)} - \widetilde{E}_{(i)}\{Q(\widetilde{\mu}_i, \widehat{\mu}_i^*)\}, \tag{4.6}$$

$\widehat{\text{cov}}_{(i)}$ being the conditional covariance estimate $\widetilde{E}_{(i)}\{\widehat{\lambda}_i^* \cdot (y_i^* - \widetilde{\mu}_i)\}$.

Proof. Applying the lemma with $\mu_i \to \widetilde{\mu}_i$, $y_i \to y_i^*$, $\widehat{\lambda}_i \to \widehat{\lambda}_i^*$, and $\widehat{\mu}_i \to \widehat{\mu}_i^*$ gives

$$O_i^* = 2(\widehat{\lambda}_i^* - \widetilde{\lambda}_i)(y_i^* - \widetilde{\mu}_i) - Q(\widetilde{\mu}_i, \widehat{\mu}_i^*). \tag{4.7}$$

Notice that $\widetilde{\mu}_i$ and $\widetilde{\lambda}_i$ stay fixed in (4.7) because they depend only on $\mathbf{y}_{(i)}$ and that this same fact eliminates the last term in (4.4). Taking conditional expectations $\widetilde{E}_{(i)}$ in (4.7) completes the proof.

In (4.6), $2\widehat{\text{cov}}_{(i)}$ equals $\widehat{\Omega}_{(i)}$, the estimate of the conditional covariance penalty $\Omega_{(i)}$, (3.20). Typically $\widehat{\Omega}_{(i)}$ is of order $\text{O}_p(1/n)$, as in (2.13), while the remainder term $\widetilde{E}_{(i)}\{Q(\widetilde{\mu}_i, \widehat{\mu}_i^*)\}$ is only $\text{O}_p(1/n^2)$. See Remark H. The implication is that

$$\widetilde{E}_{(i)}\{\widetilde{O}_i^*\} \doteq \widehat{\Omega}_{(i)} = 2 \cdot \widehat{\text{cov}}_{(i)}. \tag{4.8}$$

In other words, averaging the cross-validation estimate $\widetilde{O}_i^*$ over $\widetilde{f}_i$, the distribution of y_i^* used to calculate the covariance penalty $\widehat{\Omega}_{(i)}$, gives approximately $\widehat{\Omega}_{(i)}$ itself. If we think of $\widetilde{f}_i$ as summarizing all available information for the unknown distribution of y_i, that is, as a sufficient statistic, then $\widehat{\Omega}_{(i)}$ is a Rao–Blackwellized improvement on $\widetilde{O}_i$.

This same phenomenon occurs beyond the conditional framework of the theorem. Figure 6 concerns cross-validation of the lowess($\mathbf{x}, \mathbf{y}, 1/3$) curve in Figure 1. Using the same unconditional resampling model (2.17) as in Figure 2, $B = 200$ bootstrap replications of the cross-validation estimate (4.3) were generated,

$$\widetilde{O}_i^{*b} = Q(y_i^{*b}, \widetilde{\mu}_i^{*b}) - Q(y_i^{*b}, \widehat{\mu}_i^{*b}),$$
$$i = 1, 2, \ldots, n, \text{ and } b = 1, 2, \ldots, B. \tag{4.9}$$

The small points in Figure 6 indicate individual values $\widetilde{O}_i^{*b}/2\widehat{\sigma}^2$. The triangles show averages over the 200 replications, $\widetilde{O}_i^{*\cdot}/2\widehat{\sigma}^2$. There is striking agreement with the covariance

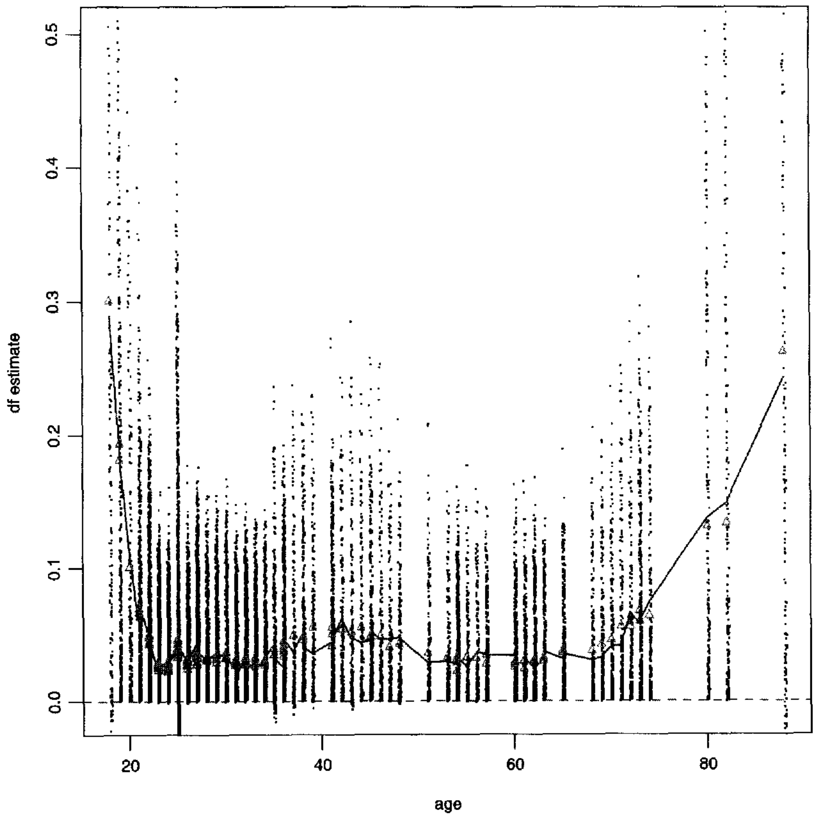

Figure 6. Small Dots Indicate 200 Bootstrap Replications of Cross-Validation Optimism Estimates (4.9); Triangles, Their Averages, Closely Match the Covariance Penalty Curve From Figure 2. (Vertical distance plotted in df units.)

penalty curve $\widehat{\text{cov}}_i/\widehat{\sigma}^2$ from Figure 2, confirming $\widehat{E}\{\widetilde{O}_i^*\} \doteq \widehat{\Omega}_i$ as in (4.8). Nearly the same results were obtained bootstrapping from $\widetilde{\boldsymbol{\mu}}$ rather than $\widehat{\boldsymbol{\mu}}$.

Approximation (4.8) can be made more explicit in the case of squared error loss applied to linear rules $\widehat{\boldsymbol{\mu}} = M\mathbf{y}$ that are "self-stable," that is, where the cross-validation estimate (4.2) satisfies

$$\widetilde{\mu}_i = \sum_{j \neq i} \widetilde{M}_{ij} y_j, \qquad \widetilde{M}_{ij} = \frac{M_{ij}}{(1 - M_{ii})}. \tag{4.10}$$

Self-stable rules include all the usual regression and ANOVA estimate as well as spline methods; see Remark I. Suppose we are resampling from $y_i \sim \bar{f}_i$ with mean and variance

$$y_i^* \sim (\bar{\mu}_i, \bar{\sigma}^2), \tag{4.11}$$

where $\bar{\mu}_i$ might differ from $\widehat{\mu}_i$ or $\widetilde{\mu}_i$. The covariance penalty Ω_i is then estimated by $\widehat{\Omega}_i = 2\bar{\sigma}^2 M_{ii}$.

Using (4.10), it is straightforward to calculate the conditional expectation of the cross-validation estimate $\widetilde{O}_i$,

$$E_{\bar{f}_i}\{\widetilde{O}_i^* | \mathbf{y}_{(i)}\} = \widehat{\Omega}_i \cdot [1 - M_{ii}/2]\left[1 + \left(\frac{\widetilde{\mu}_i - \bar{\mu}_i}{\bar{\sigma}}\right)^2\right]. \tag{4.12}$$

If $\bar{\mu}_i = \widetilde{\mu}_i$ then (4.12) becomes $\widehat{E}_{(i)}\{O_i^*\} = \widehat{\Omega}_i[1 - M_{ii}/2]$, an exact version of (4.8). The choice $\bar{\mu}_i = \widehat{\mu}_i$ results in

$$E_{\bar{f}_i}\{O_i^* | \mathbf{y}_{(i)}\} = \widehat{\Omega}_i[1 - M_{ii}/2]\left[1 + \left(M_{ii}\frac{y_i - \widehat{\mu}_i}{\bar{\sigma}}\right)^2\right]. \tag{4.13}$$

In both cases the conditional expectation of O_i^* is $\widehat{\Omega}_i[1 + \mathrm{O}_p(1/n)]$, where the $\mathrm{O}_p(1/n)$ term tends to be slightly negative.

The unconditional expectation of $\widetilde{O}_i$ with respect to the true distribution $\mathbf{y} \sim (\boldsymbol{\mu}, \sigma^2\mathbf{I})$ looks like (4.12),

$$E\{\widetilde{O}_i\} = \Omega_i[1 - M_{ii}/2]\left[1 + \sum_{j \neq i} \widetilde{M}_{ij}^2 + \Delta_i^2\right], \tag{4.14}$$

Ω_i equaling the covariance penalty $2\sigma^2 M_{ii}$ and

$$\Delta_i^2 = \left[\left(\mu_i - \sum_{j \neq i} \widetilde{M}_{ij}\mu_j\right)\Big/\sigma\right]^2. \tag{4.15}$$

For M a projection matrix, $M^2 = M$, the term $\sum \widetilde{M}_{ij}^2 = M_{ii}/(1 - M_{ii})$; $E\{\widetilde{O}_i\}$ exceeds Ω_i, but only by a factor of $1 + \mathrm{O}(1/n)$ if $\Delta_i^2 = 0$. Notice that

$$\sum_{j \neq i} \widetilde{M}_{ij}\mu_j - \mu_i = E\{\widetilde{\mu}_i - \mu_i\}, \tag{4.16}$$

so that Δ_i^2 will be large if the cross-validation estimator $\widetilde{\mu}_i$ is badly biased.

Finally, suppose $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ and $\widehat{\boldsymbol{\mu}} = M\mathbf{y}$ is a self-stable projection estimator. Then the coefficient of variation of $\widetilde{O}_i$ is

$$\mathrm{CV}\{\widetilde{O}_i\} = \frac{2 + 4(1 - M_{ii})\Delta_i^2}{(1 + 2(1 - M_{ii})\Delta_i^2)^2} \doteq 2, \tag{4.17}$$

the last approximation being quite accurate unless $\widetilde{\mu}_i$ is badly biased. This says that $\widetilde{O}_i$ must always be a highly variable estimate of its expectation (4.14), or, approximately, of $\Omega_i = 2\sigma^2 M_{ii}$. However, it is still possible for the sum $\widetilde{O} = \Sigma_i \widetilde{O}_i$ to estimate $\Sigma_i \Omega_i = 2\sigma^2\mathrm{df}$ with reasonable accuracy.

As an example $\widehat{\boldsymbol{\mu}} = M\mathbf{y}$ was fit to the kidney data, where M represented a natural spline with 8 degrees of freedom (including the intercept). One hundred simulated data vectors $\mathbf{y}^* \sim \mathrm{N}(\widehat{\boldsymbol{\mu}}_j, \widehat{\sigma}^2\mathbf{I})$ were independently generated, $\widehat{\sigma}^2 = 3.17$, each giving a cross-validated df estimate $\widetilde{\mathrm{df}}^* = \widetilde{O}^*/2\widehat{\sigma}^2$. These had empirical mean and standard deviation

$$\widetilde{\mathrm{df}}^* \sim 8.34 \pm 1.64. \tag{4.18}$$

Of course, there is no reason to use cross-validation here because the covariance penalty estimate $\widehat{\mathrm{df}}$ always equals the correct df value 8. This is an extreme example of the Rao–Blackwell type of result in Theorem 1, showing the cross-validation estimate $\widetilde{\mathrm{df}}$ as a randomized version of $\widehat{\mathrm{df}}$.

Remark G. Theorem 1 applies directly to *grouped cross-validation*, in which the observations are removed in groups rather than singly. Suppose group i consists of observations $(y_{i1}, y_{i2}, \ldots, y_{iJ})$, and likewise $\mu_i = (\mu_{i1}, \ldots, \mu_{iJ})$, $\widehat{\mu}_i = (\widehat{\mu}_i, \ldots, \widehat{\mu}_{iJ})$; $\mathbf{y}_{(i)}$ equals $\mathbf{y}$ with group i removed, and $\widetilde{\mu}_i = m(\mathbf{y}_{(i)})_{i1,i2,\ldots,iJ}$. Theorem 1 then holds as stated with $\widetilde{O}_i^* = \sum_j \widetilde{O}_{ij}^*$, $\widehat{\text{cov}}_{(i)} = \sum_j \widehat{\text{cov}}_{(ij)}$, and so forth. Another way to say this is that by additivity the theory of Sections 2–4 can be extended to vector observations y_i.

Remark H. The full dataset for a prediction problem, the "training set," is of the form

$$\mathbf{v} = (v_1, v_2, \ldots, v_n) \quad \text{with } v_i = (x_i, y_i), \tag{4.19}$$

x_i being a p vector of observed covariates, such as age for the kidney data, and y_i a response. Covariance penalties operate in a regression theory framework where the x_i are considered fixed ancillaries whether or not they are random, which is why notation such as $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$ can suppress $\mathbf{x}$. Cross-validation, however, changes $\mathbf{x}$ as well as $\mathbf{y}$. In this framework it is more precise to write the prediction rule as

$$m(x, \mathbf{v}) \quad \text{for } x \in \mathcal{X}, \tag{4.20}$$

indicating that the training set $\mathbf{v}$ determines a rule $m(\cdot, \mathbf{v})$, which then can be evaluated at any x in the predictor space $\mathcal{X}$; (4.2) is better expressed as $\widetilde{\mu}_i = m(x_i, \mathbf{v}_{(i)})$.

In the cross-validation framework we can suppose that $\mathbf{v}$ has been produced by random sampling ("iid") from some $(p+1)$-dimensional distribution F,

$$F \overset{\text{iid}}{\rightarrow} v_1, v_2, \ldots, v_n. \tag{4.21}$$

Standard influence function arguments, as in chapter 2 of Hampel, Ronchetti, Rousseeuw, and Stahel (1986), give the first-order approximation

$$\widehat{\mu}_i - \widetilde{\mu}_i = m(x_i, \mathbf{v}) - m(x_i, \mathbf{v}_{(i)}) \doteq \frac{\mathrm{IF}_i - \overline{\mathrm{IF}}_{(i)}}{n}, \tag{4.22}$$

where $\mathrm{IF}_j = \mathrm{IF}(v_j; m(x_i, \mathbf{v}), F)$ is the influence function for $\widehat{\mu}_i$ evaluated at v_j, and $\overline{\mathrm{IF}}_{(i)} = \sum_{j \neq i} \mathrm{IF}_j/(n-1)$.

The point here is that $\widehat{\mu}_i - \widetilde{\mu}_i$ is $\mathrm{O}_p(1/n)$ in situations where the influence function exists boundedly; see Li (1987) for a more careful discussion. In situation (4.10), $\widehat{\mu}_i - \widetilde{\mu}_i = M_{ii}(y_i - \widetilde{\mu}_i)$ so that $M_{ii} = \mathrm{O}(1/n)$ as in (2.13) implies $\widehat{\mu}_i - \widetilde{\mu}_i = \mathrm{O}_p(1/n)$. Similarly $\widehat{\mu}_i^* - \widetilde{\mu}_i = \mathrm{O}_p(1/n)$ in (4.6). If the function $q(\mu)$ of Figure 4 is locally quadratic near $\widetilde{\mu}_i$, then $Q(\widetilde{\mu}_i, \widehat{\mu}_i^*)$ in (4.6) will be $O_p(1/n^2)$ as claimed in (4.8).

Order of magnitude asymptotics are only a rough guide to practice and are not crucial to the methods discussed here. In any given situation bootstrap calculations such as (3.17) will give valid estimates whether or not (2.13) is meaningful.

Remark I. A prediction rule is "self-stable" if adding a new point (x_i, y_i) that falls exactly on the prediction surface does not change the prediction at x_i; in notation (4.20) if

$$m\big(x_i, \mathbf{v}_{(i)} \cup (x_i, \widetilde{\mu}_i)\big) = \widetilde{\mu}_i. \tag{4.23}$$

This implies $\widetilde{\mu}_i = \sum_{j \neq i} M_{ij} y_j + M_{ii} \widetilde{\mu}_i$ for a linear rule, which is (4.10). Any "least-Q" rule, which chooses $\widehat{\boldsymbol{\mu}}$ by minimizing $\sum Q(y_i, \mu_i)$ over some candidate collection of possible $\boldsymbol{\mu}$'s, must be self-stable, and this class can be extended by adding penalty terms as with smoothing splines. Maximum likelihood estimation in ordinary or curved GLM's belongs to the least-Q class.

5. A SIMULATION

Here is a small simulation study intended to illustrate covariance penalty/cross-validation relationships in a Bernoulli data setting. Figure 7 shows the underlying model used to generate the simulations. There are 30 bivariate vectors x_i and their associated probabilities μ_i,

$$(x_i, \mu_i), \qquad i = 1, 2, \ldots, 30, \tag{5.1}$$

from which we generated 200 30-dimensional Bernoulli response vectors

$$\mathbf{y} \sim \mathrm{Be}(\boldsymbol{\mu}) \tag{5.2}$$

as in (3.2). [The underlying model (5.1) was itself randomly generated by 30 independent replications of

$$Y_i \sim \mathrm{Be}\left(\frac{1}{2}\right) \quad \text{and} \quad x_i \sim \mathrm{N}_2\left(\left(Y_i - \frac{1}{2}, 0\right), \mathbf{I}\right), \tag{5.3}$$

with μ_i the Bayesian posterior $\mathrm{Prob}\{Y_i = 1 | x_i\}$.]

Our prediction rule $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$ was based on the coefficients for Fisher's linear discriminant boundary $\widehat{\alpha} + \widehat{\beta}'x = 0$:

$$\widehat{\mu}_i = 1/\left[1 + e^{-(\widehat{\alpha} + \widehat{\beta}'x_i)}\right]. \tag{5.4}$$

Equation (2.13) of Efron (1983) describes the $(\widehat{\alpha}, \widehat{\beta})$ computations. Rule (5.4) is *not* the logistic regression estimate of $\widehat{\mu}_i$ and in fact will be somewhat more efficient given mechanism (5.3) (Efron 1975).

Binomial deviance error (3.4) was used to assess prediction accuracy. Three estimates of the total expected optimism $\Omega = \sum_{i=1}^{30} \Omega_i$, (3.10), were computed for each of the 200 $\mathbf{y}$ vectors: the cross-validation estimate $\widetilde{O} = \sum \widetilde{O}_i$, (4.3); the parametric bootstrap estimate $2 \sum \widehat{\mathrm{cov}}_i$, (3.17) with $\mathbf{y}^* \sim \mathrm{Be}(\widehat{\boldsymbol{\mu}})$; and the Steinian $2 \sum \widehat{\mathrm{cov}}_{(i)}$, (3.22).

The results appear in Figure 8 as histograms of the 200 df estimates (i.e., estimates of optimism/2). The Steinian and parametric bootstrap gave similar results, correlation .72, with the bootstrap estimates slightly but consistently larger. Strikingly,

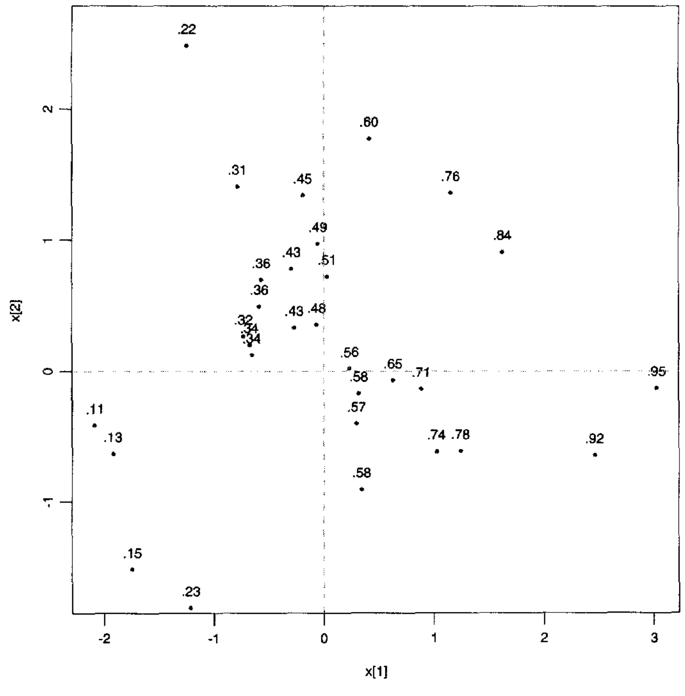

Figure 7. Underlying Model Used for Simulation Study: $n = 30$ Bivariate Vectors x_i and Associated Probabilities μ_i, (5.1).

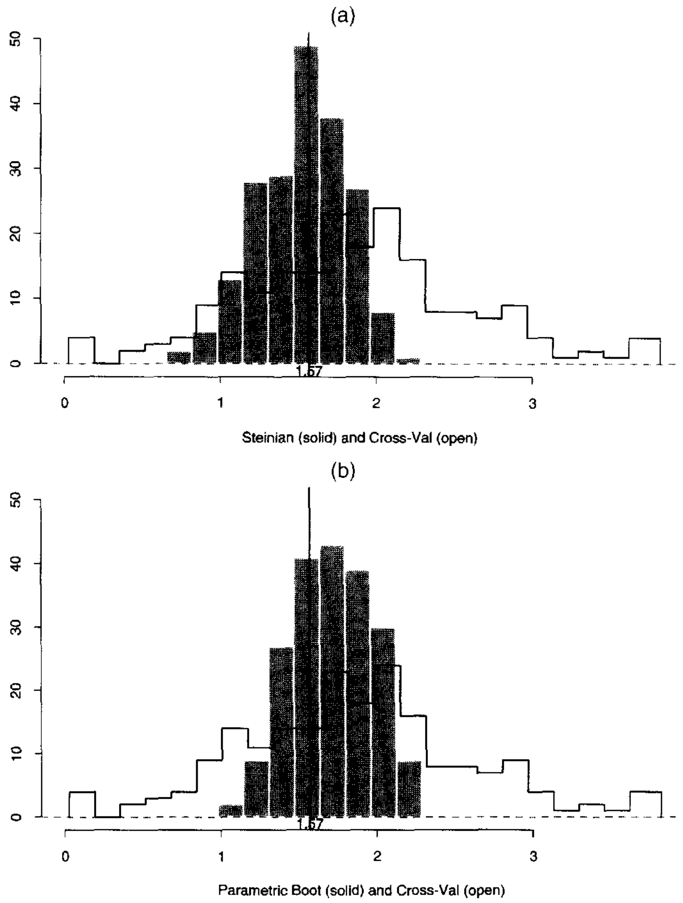

Figure 8. Degree-of-Freedom Estimates (optimism/2); 200 Simulations (5.2) and (5.4). The two covariance penalty estimates, Steinian (a) and parametric bootstrap (b), have about one-third the standard deviation of cross-validation. Error measured by binomial deviance (3.4); true $\Omega/2 = 1.57$.

the cross-validation estimates were much more variable, having about three terms larger standard deviation than either covariance penalty. All three methods were reasonably well centered near the true value $\Omega/2 = 1.57$.

Figure 8 exemplifies the Rao–Blackwell relationship (4.8), which guarantees that cross-validation will be more variable than covariance penalties. The comparison would have been more extreme if we had estimated μ by logistic regression rather than (5.4), in which case the covariance penalties would be nearly constant while cross-validation would still vary.

In our simulation study we can calculate the true total optimism (3.28) for each $\mathbf{y}$,

$$O = 2\sum_{i=1}^{n} \widehat{\lambda}_i \cdot (y_i - \mu_i). \qquad (5.5)$$

Figure 9 plots the Steinian estimates versus $O/2$ for the 200 simulations. The results illustrate an unfortunate phenomenon noted in Efron (1983): Optimism estimates tend to be small when they should be large and vice versa. Cross-validation or the parametric bootstrap exhibited the same inverse relationships. The best we can hope for is to estimate the *expected* optimism Ω.

If we are trying to estimate $\mathrm{Err} = \overline{\mathrm{err}} + O$ with $\widehat{\mathrm{Err}} = \overline{\mathrm{err}} + \widehat{\Omega}$, then

$$\widehat{\mathrm{Err}} - \mathrm{Err} = \widehat{\Omega} - O, \qquad (5.6)$$

so inverse relationships such that those in Figure 9 make $\widehat{\mathrm{Err}}$ less accurate. Table 1 shows estimates of $E\{(\widehat{\mathrm{Err}} - \mathrm{Err})^2\}$ from the simulation experiment.

None of the methods did very much better than simply estimating Err by the apparent error, that is, taking $\widehat{\Omega} = 0$, and cross-validation was actually worse. It is easy to read too much into these numbers. The two points at the extreme right of Figure 9 contribute heavily to the comparison, as do other details of the computation; see Remarks J and L. Perhaps the main point is that the efficiency of covariance penalties helps more in estimating Ω than in estimating Err. Estimating Ω can be important in its own right because it provides df values for the comparison, formal or informal, of different models, as emphasized in Ye (1998). Also, the values of df_i as a function of x_i, as in Figure 2, are a useful diagnostic for the geometry of the fitting process.

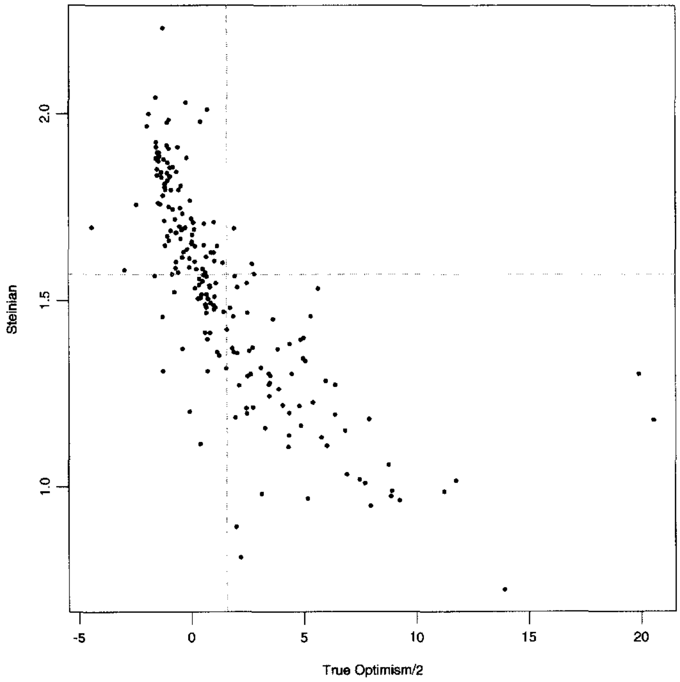

Figure 9. Steinian Estimate versus True Optimism/2, (5.5), for the 200 Simulations. Similar inverse relationships hold for parametric bootstrap or cross-validation.

Table 1. Average $(\widehat{Err} - Err)^2$ for the 200 Simulations; "Apparent" Takes $\widehat{Err} = \overline{err}$ (i.e., $\widehat{\Omega} = 0$). Outsample Averages Discussed in Remark L. All Three Methods Outperformed $\overline{err}$ When Prediction Rule Was Ordinary Logistic Regression

	Steinian	*ParBoot*	*CrossVal*	*Apparent*
$E(\widehat{\mathrm{Err}} - \mathrm{Err})^2$	**53.9**	**52.9**	**63.3**	**57.8**
Outsample	59.4	58.2	68.9	64.1
Logistic regression	36.2	34.6	33.2	53.8

The bottom line of Table 1 reports $E(\widehat{\mathrm{Err}} - \mathrm{Err})^2$ for the prediction rule "ordinary logistic regression," rather than (5.4). Now all three methods handily beat the apparent error. The average prediction $\widehat{\mathrm{Err}}$ was much bigger for logistic regression, 6.15 versus 2.93 for (5.4), but Err was easier to estimate for logistic regression.

Remark J. Four of the cross-validation estimates, corresponding to the rightmost points in Figure 9, were negative (ranging from −9 to −28). These were truncated at 0 in Figure 8 and Table 1. The parametric bootstrap estimates were based on only $B = 100$ replications per case, leaving substantial simulation error. Standard components-of-variance calculations for the 200 cases were used in Figure 8 and Table 1, to approximate the ideal situation $B = \infty$.

Remark K. The asymptotics in Li (1985) imply that in his setting it is possible to estimate the optimism itself rather than its expectation. However, the form of (5.5) strongly suggests that O is unestimable in the Bernoulli case, since it directly involves the unobservable componentwise differences $y_i - \mu_i$.

Remark L. $\mathrm{Err} = \sum \mathrm{Err}_i$, (3.8), is the total prediction error at the n observed covariate points x_i. "Outsample error,"

$$\mathrm{Err}_{\mathrm{out}} = n \cdot E_0\{Q(y^0, m(x^0, \mathbf{v}))\}, \qquad (5.7)$$

where the training set $\mathbf{v}$ is fixed while $v^0 = (x^0, y^0)$ is an independent random test point drawn from F, (4.9), is the natural setting for cross-validation. (The factor n is included for comparison with Err.) See section 7 of Efron (1986). The second line of Table 1 shows that replacing Err with $\mathrm{Err}_{\mathrm{out}}$ did not affect our comparisons. Formula (4.14) suggests that this might be less true if our estimation rule had been badly biased.

Table 2 shows the comparative ranks of $|\widehat{\mathrm{Err}} - \mathrm{Err}|$ for the four methods of Table 1 applied to rule (5.4). For example, the Steinian was best in 14 of the 200 simulations, and worst only once. The corresponding ranks are also shown for $|\widehat{\mathrm{Err}} - \mathrm{Err}_{\mathrm{out}}|$, with very similar results: Cross-validation performed poorly, apparent error tended to be either best or worst, the Steinian was usually second or third, while the parametric bootstrap spread rather evenly across the four ranks.

Table 2. Left: Comparative Ranks of $|\widehat{Err} - Err|$ for the 200 Simulations (5.1)–(5.4). Right: Same for $|\widehat{Err} - Err_{out}|$

	Stein	*ParBoot*	*CrVal*	*App*	*Stein*	*ParBoot*	*CrVal*	*App*
1	14	48	33	105	17	50	32	101
2	106	58	31	5	104	58	35	3
3	79	56	55	10	78	54	55	13
4	1	38	81	80	1	38	78	83
Mean rank	2.34	2.42	2.92	2.33	2.31	2.40	2.90	2.39

6. THE NONPARAMETRIC BOOTSTRAP

Nonparametric bootstrap methods for estimating prediction error depend on simple random resamples $\mathbf{v}^* = (v_1^*, v_2^*, \dots, v_n^*)$ from the training set $\mathbf{v}$, (4.17), rather than parametric resamples as in (2.14). Efron (1983) examined the relationship between the nonparametric bootstrap and cross-validation. This section develops a Rao–Blackwell type of connection between the nonparametric and parametric bootstrap methods, similar to Section 4's cross-validation results.

Suppose we have constructed B nonparametric bootstrap samples $\mathbf{v}^*$, each of which gives a bootstrap estimate $\widehat{\boldsymbol{\mu}}^*$, with $\widehat{\mu}_i^* = m(x_i, \mathbf{v}^*)$ in the notation of (4.18). Let N_i^b indicate the number of times v_i occurs in bootstrap sample $\mathbf{v}^{*b}$, $b = 1, 2, \dots, B$; define the indicator

$$I_i^b(h) = \begin{cases} 1 & \text{if } N_i^b = h \\ 0 & \text{if } N_i^b \neq h, \end{cases} \tag{6.1}$$

$h = 0, 1, \dots, n$; and let $\bar{Q}_i(h)$ be the average error when $N_i^b = h$,

$$\bar{Q}_i(h) = \sum_b I_i^b(h) Q\big(y_i, \widehat{\mu}_i^{*b}\big) \Big/ \sum_b I_i^b(h). \tag{6.2}$$

We expect $\bar{Q}_i(0)$, the average error when v_i not involved in the bootstrap prediction of y_i, to be larger than $\bar{Q}_i(1)$, which will be larger than $\bar{Q}_i(2)$, and so on.

A useful class of nonparametric bootstrap optimism estimates takes the form

$$\widehat{O}_i = \sum_{h=1}^n B(h)\bar{S}_i(h), \qquad \bar{S}_i(h) = \frac{\bar{Q}_i(0) - \bar{Q}(h)}{h}, \tag{6.3}$$

the "S" standing for "slope." Letting $P_n(h)$ be the binomial resampling probability

$$p_n(h) = \text{Prob}\{\text{Bi}(n, 1/n) = h\} = \binom{n}{h} \frac{(n-1)^{n-h}}{n^n}, \tag{6.4}$$

section 8 of Efron (1983) considers two particular choices of $B(h)$:

$$\begin{aligned} \text{“}\widehat{\omega}^{(\text{boot})}\text{”}: &\quad B(h) = h(h-1)p_n(h) \quad \text{and} \\ \text{“}\widehat{\omega}^{(0)}\text{”}: &\quad B(h) = hp_n(h). \end{aligned} \tag{6.5}$$

Here we will concentrate on (6.3) with $B(h) = hp_n(h)$, which is convenient but not crucial to the discussion. Then $B(h)$ is a probability distribution, $\sum_1^n B(h) = 1$, with expectation

$$\sum_1^n B(h) \cdot h = 1 + \frac{n-1}{n}. \tag{6.6}$$

The estimate $\widehat{O}_i = \sum_1^n B(h)\bar{S}_i(h)$ is seen to be a weighted average of the downward slopes $\bar{S}_i(h)$. Most of the weight is on the first few values of h because $B(h)$ rapidly approaches the shifted Poisson(1) density $e^{-1}/(h-1)!$ as $n \to \infty$.

We first consider a conditional version of the nonparametric bootstrap. Define $\mathbf{v}_{(i)}(h)$ to be the augmented training set

$$\mathbf{v}_{(i)}(h) = \mathbf{v}_{(i)} \cup \{h \text{ copies of } (x_i, y_i)\}, \qquad h = 0, 1, \dots, n, \tag{6.7}$$

giving corresponding estimates $\widehat{\mu}_i(h) = m(x_i, \mathbf{v}_{(i)}(h))$ and $\widehat{\lambda}_i(h) = -\dot{q}(\widehat{\mu}_i(h))/2$. For $\mathbf{v}_{(i)}(0) = \mathbf{v}_{(i)}$, the training set with $v_i = (x_i, y_i)$ removed, $\widehat{\mu}_i(0) = \widetilde{\mu}_i$, (4.2), and $\widehat{\lambda}_i(0) = \widetilde{\lambda}_i$. The conditional version of definition (6.3) is

$$\widehat{O}_{(i)} = \sum_{h=1}^n B(h)S_i(h),$$
$$S_i(h) = \big[Q(y_i, \widetilde{\mu}_i) - Q\big(y_i, \widehat{\mu}_i(h)\big)\big]/h. \tag{6.8}$$

This is defined to be the *conditional nonparametric bootstrap* estimate of the conditional optimism $\Omega_{(i)}$, (3.20). Notice that setting $B(h) = (1, 0, 0, \dots, 0)$ would make $\widehat{O}_{(i)}$ equal $\widetilde{O}_i$, the cross-validation estimate (4.3).

As before we can average $\widehat{O}_{(i)}(\mathbf{y})$ over conditional *parametric* resamples $\mathbf{y}^* = (\mathbf{y}_{(i)}, y_i^*)$, (4.5), with $\mathbf{y}_{(i)}$ and $\mathbf{x} = (x_1, x_2, \dots, x_n)$ fixed. That is, we can parametrically average the nonparametric bootstrap. The proof of Theorem 1 applies here, giving a similar result:

Theorem 2. Assuming (4.5), the conditional parametric bootstrap expectation of $\widehat{O}_{(i)}^* = \widehat{O}_{(i)}(\mathbf{y}_{(i)}, y_i^*)$ is

$$\widetilde{E}_{(i)}\big\{\widehat{O}_{(i)}^*\big\} = 2\sum_{h=1}^n B(h)\widehat{\text{cov}}_{(i)}(h)/h$$
$$- \sum_{h=1}^n B(h)\widetilde{E}_{(i)}\big\{Q\big(\widetilde{\mu}_i, \widehat{\mu}_i^*(h)\big)\big\}/h, \tag{6.9}$$

where

$$\widehat{\text{cov}}_{(i)}(h) = \widetilde{E}_{(i)}\{\widehat{\lambda}_i(h)^*(y_i^* - \widetilde{\mu}_i)\}. \tag{6.10}$$

The second term on the right side of (6.9) is $O_p(1/n^2)$ as in (4.8), giving

$$\widetilde{E}_{(i)}\big\{\widehat{O}_{(i)}^*\big\} \doteq 2\sum_{h=1}^n B(h) \frac{\widehat{\text{cov}}_{(i)}(h)}{h}. \tag{6.11}$$

Point v_i has h times more weight in the augmented training set $\mathbf{v}_{(i)}(h)$ than in $\mathbf{v} = \mathbf{v}_i(1)$; so, as in (4.20), influence function calculations suggest

$$\widehat{\mu}_i^*(h) - \widetilde{\mu}_i \doteq h \cdot (\widehat{\mu}_i^* - \widetilde{\mu}_i) \quad \text{and} \quad \widehat{\text{cov}}_{(i)}(h) \doteq h \cdot \widehat{\text{cov}}_{(i)}, \tag{6.12}$$

$\widehat{\mu}_i^* = \widehat{\mu}_i^*(1)$, so that (6.11) becomes

$$\widetilde{E}_{(i)}\big\{\widehat{O}_{(i)}^*\big\} \doteq 2\widehat{\text{cov}}_{(i)} = \widehat{\Omega}_{(i)}. \tag{6.13}$$

Averaging the conditional nonparametric bootstrap estimates over parametric resamples $(\mathbf{y}_{(i)}, y_i^*)$ results in a close approximation to the conditional covariance penalty $\widehat{\Omega}_{(i)}$.

Expression (6.9) can be exactly evaluated for linear projection estimates $\widehat{\boldsymbol{\mu}} = M\mathbf{y}$ (using squared error loss)

$$M = X(X^1X)^{-1}X', \qquad X' = (x_1, x_2, \ldots, x_n). \tag{6.14}$$

Then the projection matrix corresponding to $\mathbf{v}_{(i)}(h)$ has ith diagonal element

$$M_{ii}(h) = \frac{hM_{ii}}{1 + (h-1)M_{ii}}, \qquad M_{ii} = M_{ii}(1) = x_i'(X'X)^{-1}x_i, \tag{6.15}$$

and if $y_i^* \sim (\widetilde{\mu}_i, \widehat{\sigma}^2)$ with $\mathbf{y}_{(i)}$ fixed, then $\widehat{\text{cov}}_{(i)} = \widehat{\sigma}^2 M_{ii}(h)$. Using (6.6) and the self-stable relationship $\widehat{\mu}_i - \widetilde{\mu}_i = M_{ii}(y_i - \widetilde{\mu}_i)$, (6.9) can be evaluated as

$$\widetilde{E}_{(i)}\{\widehat{O}_i^*\} = \widehat{\Omega}_{(i)} \cdot [1 - 4M_{ii}]. \tag{6.16}$$

In this case (6.13) errs by a factor of only $[1 + O(1/n)]$.

Result (6.12) implies an approximate Rao–Blackwell relationship between nonparametric and parametric bootstrap optimism estimates when both are carried out conditionally on $\mathbf{v}_{(i)}$. As with cross-validation, this relationship seems to extend to the more familiar *unconditional* bootstrap estimator. Figure 10 concerns the kidney data and squared error loss, where this time the fitting rule $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$ is "loess(tot ~ age, span = .5)." Loess, unlike lowess, is a linear rule $\widehat{\boldsymbol{\mu}} = M\mathbf{y}$, although it is not self-stable. The solid curve traces the coordinatewise covariance penalty df$_i$ estimates M_{ii} as a function of age$_i$.

The small points in Figure 10 represent individual unconditional nonparametric bootstrap df$_i$ estimates $\widehat{O}_i^*/2\widehat{\sigma}^2$, (6.3), evaluated for 50 parametric bootstrap data vectors $\mathbf{y}^*$ obtained as in (2.17), Remark M provides the details. Their means across the 50 replications, the triangles, follow the M_{ii} curve. As with cross-validation, if we attempt to improve nonparametric bootstrap optimism estimates by averaging across the $\mathbf{y}^*$ vectors giving the covariance penalty $\widehat{\Omega}_i$, we wind up close to $\widehat{\Omega}_i$ itself.

As in Figure 8 we can expect nonparametric bootstrap df estimates to be much more variable than covariance penalties. Various versions of the nonparametric bootstrap, particularly the ".632 rule," outperformed cross-validation in Efron (1983) and Efron and Tibshirani (1997) and may be preferred when nonparametric error predictions are required. However, covariance penalty methods offer increased accuracy whenever their underlying models are believable.

A general verification of the results of Figure 10, linking the unconditional versions of the nonparametric and parametric bootstrap df estimates, is conjectural at this point. Remark N outlines a plausibility argument.

Remark M. Figure 10 involved two resampling levels: Parametric bootstrap samples $\mathbf{y}^{*a} = \widehat{\boldsymbol{\mu}} + \boldsymbol{\epsilon}^{*a}$ were drawn as in (2.17) for $a = 1, 2, \ldots, 50$, with $\widehat{\boldsymbol{\mu}}$ and the residuals $\widehat{\epsilon}_j = y_j - \widehat{\mu}_j$ determined by loess(span = .5); then $B = 200$ nonparametric bootstrap samples were drawn from each set $\mathbf{v}^{*a} = (v_1^{*a}, v_2^{*a}, \ldots, v_n^{*a})$, with, say, N_j^{ab} repetitions of $v_j^{*a} = (x_j, y_j^{*a})$ in the abth nonparametric resample, $b = 1, 2, \ldots, B$. For each "a," the $n \times B$ matrices of counts N_i^{ab} and estimates $\widehat{\mu}_i^{*ab}$ gave $\bar{Q}_i(h)^{*a}$,

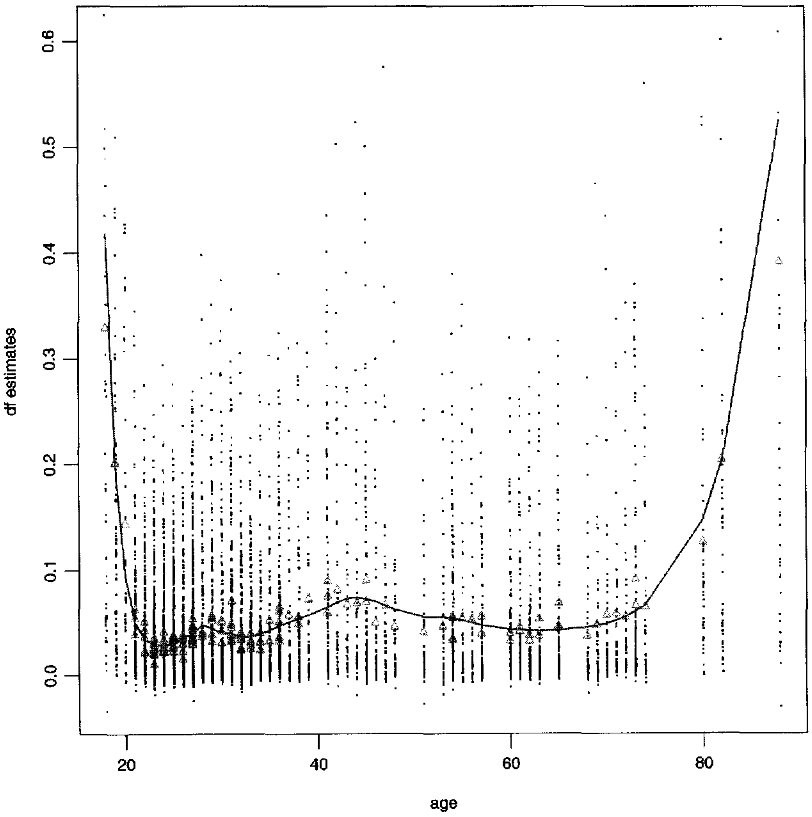

Figure 10. Small Dots Indicate 50 Parametric Bootstrap Replications of Unconditional Nonparametric Optimism Estimates (6.3); Triangles, Their Averages, Closely Follow the Covariance Penalty Estimates (solid curve). Vertical distance plotted in df units. Here the estimation rule is loess(span = .5). See Remark M for details.

$\bar{S}_i(h)^{*a}$, and $\widehat{O}_i^{*a}$, as in (6.2)–(6.3). The points $\widehat{O}_i^{*a}/2\widehat{\sigma}^2$ (with $\widehat{\sigma}^2 = \sum \widehat{\epsilon}_j^2/n$) are the small dots in Figure 10, while the triangles are their averages over $a = 1, 2, \ldots, 50$. Standard t tests accepted the null hypotheses that the averages were centered on the solid curve.

Remark N. Theorem 2 applies to parametric averaging of the *conditional* nonparametric bootstrap. For the usual *unconditional* nonparametric bootstrap, the bootstrap weights N_j on the points $v_j = (x_j, y_j)$ in $\mathbf{v}_{(i)}$ vary so that the last term in (4.4) is no longer negated by assumption (4.5). Instead it adds a remainder term to (6.9):

$$-2 \sum_{h=1} B(h) \widetilde{E}_{(i)} \big\{ \big(\widehat{\lambda}_i^*(h) - \widetilde{\lambda}_i^* \big) (\widetilde{\mu}_i^* - \widetilde{\mu}_i) \big\} / h. \tag{6.17}$$

Here $\widetilde{\mu}_i^* = m(x_i, \mathbf{v}_{(i)}^*)$, where $\mathbf{v}_{(i)}^*$ puts weight N_j on v_j for $j \neq i$, and $\widehat{\lambda}_i^* = -\dot{q}(\widetilde{\mu}_i^*)/2$.

To justify approximation (6.13), we need to show that (6.17) is $o_p(1/n)$. This can be demonstrated explicitly for linear projections (6.14). The result seems plausible in general since $\widehat{\lambda}_i^*(h) - \widetilde{\lambda}_i^*$ is $O_p(1/n)$ while $\widetilde{\mu}_i^* - \widetilde{\mu}_i$, the nonparametric bootstrap deviation of $\widetilde{\mu}_i^*$ from $\widetilde{\mu}_i$, would usually be $O_p(1/\sqrt{n})$.

7. SUMMARY

Figure 11 classifies prediction error estimates on two criteria: Parametric (model-based) versus nonparametric, and conditional versus unconditional. The classification can also be described by which parts of the training set $\{(x_j, y_j),\ j = 1, 2, \ldots, n\}$ are varied in the error rate computations: The Steinian only varies y_i in estimating the ith error rate, keeping all the covariates x_j and also y_j for $j \neq i$ fixed; at the other extreme the nonparametric bootstrap simultaneously varies the entire training set.

Here are some comparisons and comments concerning the four methods.

- The parametric methods require modeling assumptions in order to carry out the covariance penalty calculations.

	CONDITIONAL (local)	UNCONDITIONAL (global)	
PARAMETRIC (model-based covariance penalties	Steinian	Parametric Bootstrap	covariates fixed
NONPARAMETRIC (model-free)	Cross-Validation	Nonparametric Bootstrap	covariates random
	only ith case random	all cases random	

Figure 11. Two-Way Classification of Prediction Error Estimates Discussed in This Article. The conditional methods are local in the sense that only the ith case data are varied in estimating the ith error rate.

When these assumptions are justified, the Rao–Blackwell type of results of Sections 4 and 6 imply that the parametric techniques will be more efficient than their nonparametric counterparts, particularly for estimating degrees of freedom.

- The modeling assumptions need not rely on the estimation rule $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$ under investigation. We can use "bigger" models as in Remark A, that is, ones less likely to be biased.
- Modeling assumptions are less important for rules $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$ that are close to linear. In genuinely linear situations such as those needed for the C_p and AIC criteria, the covariance corrections are constants that do not depend on the model at all. The centralized version of SURE, (3.25), extends this property to maximum likelihood estimation in curved families.
- Local methods extrapolate error estimates from small changes in the training set. Global methods make much larger changes in the training set, of a size commensurate with actual random sampling, which is an advantage in dealing with "rough" rules $m(\mathbf{y})$ such as nearest neighbors or classification trees; see Efron and Tibshirani (1997).
- Stein's SURE criterion (2.11) is local, because it depends on *partial* derivatives, and parametric (2.9) without being model based. It performed more like cross-validation than the parametric bootstrap in the situation of Figure 2.
- The computational burden in our examples was less for global methods. Equation (2.18), with $\widehat{\lambda}_i^{*b}$ replacing $\widehat{\mu}_i^{*b}$ for general error measures, helps determine the number of replications B required for the parametric bootstrap. Grouping, the usual labor-saving tactic in applying cross-validation, can also be applied to covariance penalty methods as in Remark G, though now it is not clear that this is computationally helpful.
- As shown in Remark B, the bootstrap method's computations can also be used for hypothesis tests comparing the efficacy of different models.

Accurate estimation of prediction error tends to be difficult in practice, particularly when applied to the choice between competing rules $\widehat{\boldsymbol{\mu}} = m(\mathbf{y})$. In the author's opinion it will often be worth chancing plausible modeling assumptions for the covariance penalty estimates, rather than relying entirely on nonparametric methods.

[Received December 2002. Revised October 2003.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," *Second International Symposium on Information Theory*, 267–281.

Breiman, L. (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error," *Journal of the American Statistical Association*, 87, 738–754.

Efron, B. (1975), "Defining the Curvature of a Statistical Problem (With Applications to Second Order Efficiency)" (with discussion), *The Annals of Statistics*, 3, 1189–1242.

——— (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analyses," *Journal of the American Statistical Association*, 70, 892–898.

——— (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331.

——— (1986), "How Biased Is the Apparent Error Rate of a Prediction Rule?" *Journal of the American Statistical Association*, 81, 461–470.

Efron, B., and Tibshirani, R. (1997), "Improvements on Cross-Validation: The 632+ Bootstrap Method," *Journal of the American Statistical Association*, 92, 548–560.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986), *Robust Statistics, the Approach Based on Influence Functions*, New York: Wiley.

Hastie, T., and Tibshirani, R. (1990), *Generalized Linear Models*, London: Chapman & Hall.

Li, K. (1985), "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross-Validation," *The Annals of Statistics*, 13, 1352–1377.

——— (1987), "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975.

Mallows, C. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

Rockafellar, R. (1970), *Convex Analysis*, Princeton, NJ: Princeton University Press.

Shen, X., Huang, H.-C., and Ye, J. (2004), "Adaptive Model Selection and Assessment for Exponential Family Models," *Technometrics*, to appear.

Shen, X., and Ye, J. (2002), "Adaptive Model Selection," *Journal of the American Statistical Association*, 97, 210–221.

Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9, 1135–1151.

Tibshirani, R., and Knight, K. (1999), "The Covariance Inflation Criterion for Adaptive Model Selection," *Journal of the Royal Statistical Society*, Ser. B, 61, 529–546.

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

Comment

Prabir Burman

I would like to begin by thanking Professor Efron for writing a paper that sheds new light on cross-validation and related methods along with his proposals for stable model selection procedures. Stability can be an issue for ordinary cross-validation, especially for not-so-smooth procedures such as stepwise regression and other such sequential procedures. If we use the language of learning and test sets, ordinary cross-validation uses a learning set of size $n - 1$ and a test set of size 1. If one can average over a test set of infinite size, then one gets a stable estimator. Professor Efron demonstrates this leads to a Rao–Blackwellization of ordinary cross-validation.

The parametric bootstrap proposed here does require knowing the conditional distribution of an observation given the rest, which in turn requires a knowledge of the unknown parameters. Professor Efron argues that for a near-linear case, this poses no problem. A question naturally arises: What happens to those cases where the methods are considerably more complicated such as stepwise methods?

Another issue that is not entirely clear is the choice between the conditional and unconditional bootstrap methods. The conditional bootstrap seems to be better, but it can be quite expensive computationally. Can the unconditional bootstrap be used as a general method always?

It seems important to point out that ordinary cross-validation is not as inadequate as the present paper seems to suggest. If model selection is the goal, then estimation of the overall prediction error is what one can concentrate on. Even if the componentwise errors are not necessarily small, ordinary cross-validation may still provide reasonable estimates especially if the sample size n is at least moderate and the estimation procedure is reasonably smooth.

In this connection, I would like to point out that methods such as repeated v-fold (or multifold) cross-validation or repeated (or bootstrapped) learning-testing can improve on ordinary cross-validation because the test set sizes are not necessarily small (see, e.g., the CART book by Breiman, Friedman, Olshen, and Stone 1984; Burman 1989; Zhang 1993). In addition, such methods can reduce computational costs substantially. In a repeated v-fold cross-validation, the data are repeatedly randomly split into v groups of roughly equal sizes. For each repeat, there are v learning sets and the corresponding test sets. Each learning set is of size $n(1 - 1/v)$ approximately and each test set is of size n/v. In a repeated learning–testing method, the data are randomly split into a learning set of size $n(1 - p)$ and a test set of size np, where $0 < p < 1$. If one takes a small v, say $v = 3$, in a v-fold cross-validation or a value of $p = 1/3$ in a repeated learning–testing method, then each test set is of size $n/3$. However, a correction term is needed in order to account for the fact that each learning set is of a size that is considerably smaller than n (Burman 1989).

I ran a simulation for the classification case with the model: Y is Bernoulli$(\pi(X))$, where $\pi(X) = 1 - \sin^2(2\pi X)$ and X is Uniform$(0, 1)$. A majority voting scheme was used among the k nearest neighbor neighbors. True misclassification errors (in percent) and their standard errors (in parentheses) are given in Table 1 along with ordinary cross-validation, corrected threefold cross-validation with 10 repeats, and corrected repeated-testing (RLT) methods with $p = .33$ and 30 repeats. The sample size is $n = 100$ and the number of replications is 25,000. It can be seen that the corrected v-fold cross-validation or repeated learning–testing methods can provide some improvement over ordinary cross-validation.

I would like to end my comments with thanks to Professor Efron for providing significant and valuable insights into the subject of model selection and for developing new methods that are improvements over a popular method such as ordinary cross-validation.

Table 1. Classification Error Rates (in percent)

TCH	$k = 7$	$k = 9$	$k = 11$
True	$22.82_{(4.14)}$	$24.55_{(4.87)}$	$27.26_{(5.43)}$
CV	$22.95_{(7.01)}$	$24.82_{(7.18)}$	$27.65_{(7.55)}$
Threefold CV	$22.74_{(5.61)}$	$24.56_{(5.56)}$	$27.04_{(5.82)}$
RLT	$22.70_{(5.70)}$	$24.55_{(5.65)}$	$27.11_{(5.95)}$

Prabir Burman is Professor, Department of Statistics, University of California, Davis, CA 95616 (E-mail: *burman@wald.ucdavis.edu*).

© 2004 American Statistical Association
Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Theory and Methods
DOI 10.1198/016214504000000908

ADDITIONAL REFERENCES

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Burman, P. (1989), "A Comparative Study of Ordinary Cross-Validation, v-Fold Cross-Validation and Repeated Learning–Testing Methods," *Biometrika*, 76, 503–514.

Zhang, P. (1993)," Model Selection via Multifold Cross-Validation," *The Annals of Statistics*, 21, 299–313.

Comment

L. DENBY, J. M. LANDWEHR, and C. L. MALLOWS

We welcome this authoritative review of the field. We would like to point to some areas that seem to need further study. First, consider the model selection problem. It is well known that the "C_p estimate" of prediction error in the subset regression problem [Efron's (2.6)] does not allow for the fact that the subset is data dependent and may badly underestimate the true prediction error. Efron does not refer to this difficulty explicitly, but the general "covariance penalty" formula (2.8) allows the prediction formula $m(y)$ to be completely general, which allows for model selection. Thus we have the problem of estimating the covariance in (2.8), namely $\operatorname{cov}(\hat\mu_i, y_i)$, where y_i is generated by the (unknown) true model and $\hat\mu_i$ is the fitted value using our subset selection rule.

The parametric bootstrap proposed in (2.14)–(2.16) will not do this. Here one generates bootstrap samples with means at the fitted (subset) model. This method will be biased against allowing any other regressors into the fit. The algorithms of (2.19) and (2.20) do not work well, either, because they generate y's with variances that are too large, which makes it harder for regressors to be selected. Of course, all these difficulties arise from the discontinuous nature of the prediction rule and can be avoided by refusing to use a subset least squares predictor; any continuous rule such as ridge or "lasso" will be much easier to deal with.

In the work reported in Denby, Landwehr, and Mallows (DLM) (2001), we came across an extreme example of the phenomenon noted before (2.20), that the exact choice of model may be unimportant for estimating the prediction error. A simplified version of our situation is as follows. We had data $\{X_{ij},\ Y_{ik},\ i=1,\dots,I,\ j=1,\dots,J,\ k=1,\dots,K\}$, where the X's are replicate observations $(j=1,\dots,J)$ on certain devices $(i=1,\dots,I)$ using one kind of equipment and the Y's are replicate observations $(k=1,\dots,K)$ on the same devices using different equipment. Our study of the data led us to propose the prediction formula [for a Y observation on a new device $(i=0)$, based on observations X_{0m} $(m=1,\dots,M)$ on that device] as

$$\hat y_0 = \bar Y_{..} - \bar X_{..} + \bar X_0. \qquad (1)$$

We needed to estimate the precision of these estimates. We did this in two ways. First, we used a model we had fitted to the data, according to which $X_{ij} = \mu + a_i + e_{ij}$ and $Y_{ik} = \nu + b_i + f_{ij}$, where the e's and f's are independent of the a's and b's. Using estimates of various variances, we arrived at the "model-based" (MB) formula

$$\widehat{\mathrm{Err}}_{\mathrm{MB}} = \left(\frac{I+1}{I}\right)\mathrm{BMS} + \left(\frac{1}{L} - \frac{1}{J}\right)\mathrm{XWMS} + \left(1 - \frac{1}{K}\right)\mathrm{YWMS},$$

where BMS is the between-devices sum of squares

$$\mathrm{BMS} = \sum(\bar Y_{i.} - \bar X_{i.} - \bar Y_{..} + \bar X_{..})^2,$$

and XWMS and YWMS are within-devices sums of squares.

Second, we knew that this model did not fit the data perfectly. We had identified several systematic (but small) deviations. [Also, the prediction formula (1) is not optimal for this model.] We performed a cross-validation (CV) computation, dropping out each device in turn and predicting a Y measurement on that device from the rest. In DLM (2001) we were delighted to find that the CV estimate was very close to the "model-based" estimate, because this seemed to validate the model. However, subsequently, as reported in DLM (2002), some algebra led us to realize that the CV calculation had given the estimate

$$\widehat{\mathrm{Err}}_{\mathrm{CV}} = \left(\frac{I}{I-1}\right)\mathrm{BMS} + \left(\frac{1}{L} - \frac{1}{J}\right)\mathrm{XWMS} + \left(1 - \frac{1}{K}\right)\mathrm{YWMS},$$

which is necessarily very close to $\widehat{\mathrm{Err}}_{\mathrm{MB}}$ for any data whatsoever. Thus the model could be completely wrong, and still these two estimates of Err would agree.

The underlying reason for this close agreement is that the prediction formula that is being used is a reasonable choice for this model. The result of the algebraic analysis is that $\widehat{\mathrm{Err}}_{\mathrm{MB}}$, which uses both the prediction formula and the model, is necessarily close to $\widehat{\mathrm{Err}}_{\mathrm{CV}}$, which uses the prediction formula but no model.

We pose two questions. How general is it that $\widehat{\mathrm{Err}}_{\mathrm{MB}}$ and $\widehat{\mathrm{Err}}_{\mathrm{CV}}$ must be close? We suspect that the answer involves how sensible the prediction formula is for the model. More importantly for statistical practice, we wonder how generally it can happen that a model-based estimate of prediction error is as good as (or even better than) a cross-validation estimate, even when the model is wrong. This is the question Efron addresses

Lorraine Denby is Research Scientist, Data Analysis Research Department, Avaya Labs, Basking Ridge, NJ 07920 (E-mail: *ld@research.avayalabs.com*). James M. Landwehr is Director, Data Analysis Research Department, Avaya Labs, Basking Ridge, NJ 07920 (E-mail: *jml@research.avayalabs.com*). Collin L. Mallows is Consultant, Data Analysis Research Department, Avaya Labs, Basking Ridge, NJ 07920 (E-mail: *colinm@research.avayalabs.com*).

© 2004 American Statistical Association
Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Theory and Methods
DOI 10.1198/016214504000000962

in the last sentence of his article. Having fitted a model, which may be inaccurate, but which suggests a prediction formula, we can estimate the predictive mean squared error by naively assuming the model (perhaps using simulation, which would qualify as a "parametric bootstrap" method, except that here we may be simulating from a model we know to be incorrect). When is this better than cross-validation?

In DLM (2002) we reported our findings for the second question in several simple situations, including linear regression using an inadequate model and linear regression with variance effects that are ignored. We also studied the effect of having high-leverage observations. Our studies suggest that (at least in the cases we studied)

1. A model-based calculation is often better than CV, even when the model is wrong.
2. Variance effects are unimportant.
3. Naive CV behaves badly when there are high-leverage observations.

Even though a model-based calculation of prediction error might be more accurate than CV in many situations, it is also the case that producing the MB estimate is more cumbersome and difficult than producing the CV estimate. Thus another problem area that could impact statistical practice involves defining modifications to simple CV that could improve accuracy in estimating predictive error. In DLM (2002) we proposed some modifications based on our analysis of the regression problem and did some numerical investigations under several scenarios. We did not find substantial, consistent improvements relative to standard CV. We view this as an open problem.

A third area that needs study is the variability of estimates of predictive error. A correction for small bias will not be helpful if it seriously degrades the precision of the estimate.

We welcome Efron's comments.

ADDITIONAL REFERENCES

Denby, L., Landwehr, J. M., and Mallows, C. L. (2001), "An Exercise in the Real World of Design and Analysis," *The American Statistician*, 55, 263–271.

Denby, L., Landwehr, J. M., and Mallows, C. L. (2002), "Estimating Predictive MSE by Cross-Validation," technical report, available at *http://www.research.avayalabs.com/techabstractY.html#ALR-2002-035*.

Comment

Xiaotong SHEN, Hsin-Cheng HUANG, and Jianming YE

In many scientific and engineering problems, a central issue is deciding among competing explanations of data, possibly of different types or from different sources, in the presence of diverse error that is difficult, if not impossible, to control. At the core of progress in science and engineering is model selection and combination. The key to model selection and combination is model assessment particularly in comparing models at different levels of complexity and stability through estimation. The author is to be congratulated for making important and fundamental contributions to model assessment from a prediction standpoint.

Over the past decades, scientists and engineers have used various statistical tools for model assessment but have lacked a clear understanding of the key issues involved. Within statistics, there are a number of theories that govern estimation/prediction, and yet there are subtle differences among them in application. The main results in this article alert statisticians to the importance of reducing estimation variability while controlling bias in model assessment. As a consequence, covariance penalties provide more accurate model assessment in general, yielding more precise guidance of model selection and combination.

In this discussion, we shall stress the fundamental importance of covariance penalties in model assessment. Not only does the covariance penalty cov_i account for the complexity of a model, it also provides an assessment of any modeling process, possibly involving many models at different levels of complexity. This is in contrast to any other complexity penalty that focuses solely on a single model. Further, we shall comment on and compare two methodologies for estimating cov_i, namely, parametric bootstrap (PB) and data perturbation (DP).

1. COVARIANCE PENALTIES AND OTHER INFORMATION CRITERIA

1.1 Information Criteria

In the literature, a large number of information criteria have been proposed and investigated as a means of model assessment. In a statistical framework, data are sampled from a true yet unknown distribution, and can be modeled in terms of likelihood $f(\text{Data}, \theta)$ and a parameter vector θ, which may be regarded as an approximation to the truth. Essentially all information criteria can be summarized in the form of $-\log f(\text{Data}, \hat{\theta}) + \lambda(\theta)$, where $\lambda(\theta)$ is a nonnegative model complexity penalty and is a function of θ. Placing it in the slightly more general framework of this article, we obtain

$$Q(\text{Data}, \hat{\theta}) + \lambda(\theta). \qquad (1)$$

This reflects a compromise between two important yet conflicting aspects of modeling: goodness of fit and model complexity. Goodness of fit, described by $Q(\text{Data}, \hat{\theta})$, refers to how well

Xiaotong Shen is Professor, School of Statistics, University of Minnesota, 224 Church Street S.E., Minneapolis, MN 55455 (E-mail: *xshen@stat.umn.edu*). His research was supported in part by National Science Foundation grant IIS-0328802 and Agreement No. 0112050. Hsin-Cheng Huang is Associate Research Fellow, Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan (E-mail: *hchuang@stat.sinica.edu.tw*). Jiamingy Ye is Associate Professor, Stan Ross Department of Accountancy, Baruch College, City University of New York, New York, NY 10010 (E-mail: *jimmy_ye@baruch.cuny.edu*). He gratefully acknowledges the financial support of the Zicklin School of Business, Baruch College.

© 2004 American Statistical Association
Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Theory and Methods
DOI 10.1198/016214504000000881

a model fits into observed data, while model complexity measures the level of complexity of the model generating the fit. In the context of model selection, the model minimizing (1) is often used as the best model that is expected to generalize well to predict unseen outcome.

Complexity of a Single Model. Traditionally, model complexity is described as the characteristic of a model that enables it to fit into a variety of patterns of data, primarily measured via its size and function form. Many information criteria such as Akaike's information criterion (AIC) and the Bayesian choose $\lambda(\theta)$ to be the model size k multiplied by a constant, ignoring the model's function form. Rissanen's modified stochastic complexity (Rissanen 1996) adds an adjustment factor such that $\lambda(\theta) = (k/2)\log(n/(2\pi)) + \log\int\sqrt{\mathrm{Det}(I(\theta))}\,d\theta$, with $\mathrm{Det}(I(\theta))$ being the determinant of the Fisher information matrix $I(\theta)$.

Complexity of a Modeling Process. In any modeling process, a modeling procedure employed to yield $\hat{\theta}$ can influence goodness of fit as well as generalizability, in addition to a model's size and function form. A modeling procedure is a mapping from the sample space to R^n, defined by $\hat{\theta}$ evaluated at n observations. Its complexity is apparently a more general concept than model complexity, as it can describe any situation particularly that with multiple and data-dependent models. However, it is more difficult to determine a good complexity measure for a modeling process. For instance, in a curve estimation, a free knot spline estimator with an estimated set of knots placed anywhere in a region should have a higher level of complexity than a dyadic spline estimator with a set of prespecified knot locations. The difficulty is that both modeling processes use data-dependent and data-independent models, defined by the knots, respectively.

Covariance Penalty as a Measure of Complexity of a Modeling Procedure. Earlier $\lambda(\theta)$ represents complexity of a model in (1). In the present context, we use $\lambda(\tilde{\theta}, \theta)$ to describe that of a modeling procedure $\tilde{\theta}$. Within this framework, we are able to derive cov_i, capturing complexity as well as stability of a modeling procedure.

Regardless of the interpretation, we now determine the optimal choice of $\lambda(\tilde{\theta}, \theta)$ by considering a more general version of (1):

$$Q(\mathrm{Data}, \hat{\theta}) + \lambda(\tilde{\theta}, \theta). \tag{2}$$

Ye (1998) and Shen and Ye (2002) argued that $\lambda(\tilde{\theta}, \theta)$ in (2) not only penalizes an increase in model size, but also can capture modeling uncertainty via $\hat{\theta}$. Efron (1986) and the present article suggested that $\lambda(\tilde{\theta}, \theta)$ is necessary to estimate $\sum_{i=1}^n \mathrm{Err}_i$ unbiasedly. Efron (1986) derived $\sum_{i=1}^n \mathrm{cov}_i$ in the form of "expected optimism" via unbiasedness, while Ye (1998) obtained it in terms of generalized degrees of freedom for the Gaussian distribution. Shen and Ye (2002) and Shen, Huang, and Ye (2003) derived $\sum_{i=1}^n \mathrm{cov}_i$ as the optimal penalty that minimizes an equivalent form of $E(\widehat{\mathrm{Err}} - \mathrm{Err})^2$ over all $\lambda(\tilde{\theta}, \theta)$ in a context of loss estimation for exponential-family models.

The covariance penalty cov_i is general for any modeling process, regardless of whether it is linear or nonlinear or candidate models are nested or not. Usually, it differs from the other model complexity penalty, although in some special cases such as linear regression, it may coincide with the penalty of AIC, Mallow's C_p, or Stein's unbiased risk estimator (SURE).

With an estimated covariance penalty $\widehat{\mathrm{cov}}_i$ in place, we obtain

$$Q(\mathrm{Data}, \hat{\theta}) + \sum_{i=1}^n \widehat{\mathrm{cov}}_i, \tag{3}$$

permitting model assessment for any arbitrary modeling process. This is in contrast to (1); for instance, (3) enables us to evaluate model averaging estimators based on models of different sizes, whereas (1) cannot. Most important, the aforementioned optimality of cov_i implies that it is expected to outperform any other penalty in (1) in terms of the accuracy of prediction.

Model Stability. Model stability or sensitivity measures the stability of model-based estimation relative to a change in μ via the fitted values $\hat{\mu}$. We now argue by example that cov_i captures model stability, which is an important aspect of a modeling process.

Now consider a model selection process, with $\hat{M}$ being a data-dependent model selected from a class of candidate models via a model selection criterion. A conventional treatment to this selected model is to estimate the loss using the complexity of $\hat{M}$. By putting this in the framework of covariance penalties, the complexity of $\hat{M}$ can be measured by generalized degrees of freedom as $\sum_{i=1}^n h_i^c \equiv \sum_{i=1}^n E(\partial\hat{\mu}_i(\hat{M})/\partial y_i)$, conditioning on $\hat{M}$, where $\partial\hat{\mu}_i(\hat{M})/\partial y_i$ is the sensitivity of the fitted value $\hat{\mu}_i$ to y_i holding $\hat{M}$ fixed. Similarly, the complexity of the selection procedure as a whole is $\sum_{i=1}^n h_i^u \equiv \sum_{i=1}^n E(\partial\hat{\mu}_i/\partial\mu_i)$, where $\partial\hat{\mu}_i/\partial y_i$ is the corresponding unconditional version without holding $\hat{M}$ fixed. Their difference, due to the selection process, describes stability of the selected model $\hat{M}$ when the data are perturbed locally. We refer to this difference as model stability, which is not captured by any complexity measure of $\hat{M}$. Usually, the difference $h_i^u - h_i^c = E(\partial(\hat{\mu}_i - \hat{\mu}_i(\hat{M}))/\partial y_i)$ is nonnegative. This is because $\hat{\mu}_i$ without holding $\hat{M}$ fixed typically has higher sensitivity than that holding $\hat{M}$ fixed, to y_i. Consequently, the complexity measure of $\hat{M}$ alone yields an underestimated prediction error $\widehat{\mathrm{Err}}$. In summary, cov_i, that is, h_i^u multiplied by error variance, takes into account the complexity of not only the selected model but also the whole selection process.

1.2 Cross-Validation

As pointed out by Efron (1983), cross-validation (CV) often yields unacceptably high variance as an estimate of prediction error. The Rao–Blackwell decomposition in Theorem 1 implies that the variance of CV is no less than that of $\widehat{\mathrm{Err}}$ defined by cov_i, provided of course that the higher-order term in (4.6) is ignorable. Usually, CV and $\widehat{\mathrm{Err}}$ are approximately unbiased for Err. This in turn translates into a more accurate estimate of Err via cov_i in view of this bias/variance property. On a related matter, the results in the article can be easily generalized to cover "delete-m" cross-validation.

2. ESTIMATION OF COVARIANCE PENALTY

PB and DP. Estimation of cov_i is highly nontrivial and may require MC approximation of some type, because $\lambda(\tilde{\theta}, \theta)$ captures modeling uncertainty that is usually difficult or impossible to describe analytically. To our knowledge, there are two general techniques available. The first is the DP method, developed in Ye (1998), Shen and Ye (2002), and Shen et al. (2003). The idea is to use the fitted values based on perturbed data to assess model accuracy. The pioneer work on data perturbation may be tracked to Breiman (1992) in linear regression. The second one is the proposed PB method in the present article. Although they substantially differ in their method of generating Y^*, there is an interesting connection between them. We shall explore this aspect next.

The DP method usually reduces to some type of PB when $c = 1$, where $0 < c \leq 1$ is the coefficient controlling the degree of shrinkage, as defined in (2.20). However, it is usually independent of candidate models. Equivalently, it samples Y^* from $N(Y, c\hat{\sigma}^2 I)$ in the simple Gaussian case while sampling Y^* from $\text{Bern}(\tilde{p})$ in the Bernoulli case. Here $\tilde{p}$, $0 < \tilde{p} < 1$, is a prespecified probability. On the other hand, the PB method based on a "moderately big" model samples from $N(\hat{\mu}, \hat{\sigma}^2 I)$ and $\text{Bern}(\hat{p})$, respectively, in the Gaussian and Bernoulli cases, with $\hat{\mu}$ and $\hat{p}$ being an estimated mean and probability via the "moderately big" model.

Both methods differ substantially when $c < 1$. Take the Bernoulli case for instance. The DP method generates a convex combination of $Y^* = (1 - c)Y + c\tilde{Y}$ with $\tilde{Y}$ sampled from $\text{Bern}(\tilde{p})$, resulting in a multinomial distribution for Y^*. For estimation, it is necessary to embed the Bernoulli distribution into a more general class of multinomial distributions to perform estimation. In 0–1 tree classification, when $c = .5$, Y^* assumes three values $\{0, .5, 1\}$ rather than $\{0, 1\}$, and classification is performed via three-category trees.

Two Approaches: A "Moderately Big" Model versus "Model-Free." In the present article, a "moderately big" model is advocated for parametric bootstrapping. In principle, we agree that a good estimate $\hat{\mu}$ of μ is expected to yield an accurate estimated prediction error. However, one major concern is that the bootstrap estimates generally depend on the model employed for bootstrapping. As a consequence, it is likely to produce a bootstrap estimate that favors large models in terms of the accuracy of predication, when a large model is used for bootstrap and vice versa for a small model. Further, in many problems, there are many "moderately big" models available, resulting in bias in any direction depending on the choice of the model used for bootstrapping. In our view, a "model-free" approach that does not involve candidate models is more appropriate for making fair model comparisons via $\widehat{\text{Err}}$ in model assessment. The increased variability due to a "model-free" approach may be reduced by suitably shrinking the coefficient c toward the origin, as in (2.20). We shall elaborate in what is to follow.

To illustrate the main points, we examine the mean μ_s and variance σ_s^2 of the generating distribution of Y^* in the Gaussian case. For the "moderately big" model approach, $\mu_s = \hat{\mu}$, with $\hat{\mu}$ generated using a "moderately big" model. This is in contrast to the "model-free" approach in which $\mu_s = Y$. Now consider the situation of variable selection in linear regression, in which $\widehat{\text{Err}}$ defined by the PB method is employed to compare different candidate models. In this situation, $\hat{\mu}$ is obtained via a "moderately big" model, which can be any model involving a reasonably large subset of candidate variables. Most important, $\widehat{\text{Err}}$ defined by $\widehat{\text{cov}}_i$ is likely to be biased for (or against) certain types of models that are most (or least) associated with the subset in use, regardless of which "moderately big" model is used. As a result, the accuracy of predication may vary dramatically over the choice of "moderately big" models, making accurate model comparisons difficult or impossible. In this context, the model involving all candidate variables is usually used as the "moderately big" model (cf. Freedman, Navidi, and Peters 1988).

Variance Reduction and Adjustment. As far as σ_s^2 is concerned, $\text{Var}(Y^*) = (1 + c^2)\,\text{Var}(Y)$ becomes larger when $\mu_s = Y$ is used, where Y^* is sampled from $N(Y, c\hat{\sigma}^2 I)$. This is the price to be paid for not requiring model assumptions. Fortunately, by shrinking c toward the origin, $\text{Var}(Y^*)$ decreases. One direct benefit of using $\mu_s = Y$ is that an adjustment can be made to further improve the accuracy of prediction. For instance, in a context of DP, Ye (1998) and Shen et al. (2003) suggested using $\widehat{\text{cov}}_i / c^2$ as opposed to $\widehat{\text{cov}}_i$. In contrast, it is generally difficult to make such an adjustment for the "moderately big" model approach because $\text{Var}(Y^*)$ usually depends on the unknown truth through $\text{Var}(\hat{\mu})$.

Example. The following simulation is designed to illustrate our main points regarding the choice of μ_s and σ_s^2 in PB and DP. Consider linear regression with Gaussian error, in which response Y_i depends on covariates $x_i \equiv (x_{i,0}, x_{i,1}, \ldots, x_{i,50})'$ as follows:

$$Y_i = \mu_i + \varepsilon_i = x_i'\beta + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_0^2), \tag{1}$$

where $\beta \equiv (\beta_0, \beta_1, \ldots, \beta_{50})'$ is the vector of regression parameters and $\sigma_0^2 = 1$ is assumed to be known for simplicity.

In this example, we compare the BP method with $Y^* \sim N(\hat{\mu}_{\text{full}}, I)$ to the DP method with $Y^* \sim N(Y, c^2\sigma_0^2 I)$ in terms of the prediction accuracy, as measured by $\widehat{\text{Err}} - \text{Err}$ evaluated at $\hat{\mu}(\hat{M})$ with $\hat{M}$ selected via a modeling procedure AIC. Here $\hat{\mu}_{\text{full}}$ is the least squares estimate based on all variables $\{x_0, \ldots, x_p\}$ and the DP method uses the aforementioned adjustment with $c = .5$.

A random sample $n = 200$ of $\{(Y_i, x_i)\}_{i=1}^n$ is generated according to (1), where x_i follows $N(0, I)$. In this simulation, five situations, corresponding to five different choices of β, indexed by $k = 1, 3, 5, 7, 9$, are examined, where $\beta_0 = 1$. The coefficient vector β takes the form: $\beta_i = \beta_{i-10}$ for $i = 11, \ldots, 20$, $\beta_i = \beta_{i-20}$ for $i = 21, \ldots, 30$, $\beta_i = \beta_{i-30}$ for $i = 31, \ldots, 40$, and $\beta_i = \beta_{i-40}$ for $i = 41, \ldots, 50$. In other words, β consists of β_0 and five replications of $(\beta_1, \ldots, \beta_{10})'$. For each k, the choice of $(\beta_1, \ldots, \beta_{10})'$ comprises the first k values that are equal to a constant B_k and 0s otherwise, and the values of B_k are chosen to give $\beta'X'X\beta/(\beta'X'X\beta + 200) = .75$, where $X \equiv (x_1, \ldots, x_n)'$ is an $n \times 51$ matrix. A similar example has been previously used in George and Foster (2000) and Shen and Ye (2002) for studying variable selection in (1).

The simulation is performed in R. For each case, the bias and the mean squared error (MSE) of $\widehat{\text{Err}} - \text{Err}$ for the two approaches

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,

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

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,

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

   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.

© 2004 American Statistical Association
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, $\mathrm{cov}_i = \mathrm{cov}(\widehat{\mu}_i, y_i)$, 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, $\mathrm{cov}(\widehat{\lambda}_i, y_i) \ge 0$, holds for the generalized q class of error measures Q and when it does not. Let $\mathbf{y}_{(i)} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$. 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 $\mathrm{cov}(\widehat{\lambda}_i, y_i)$ can be rewritten as

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

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 $\mathbf{y}_{(i)}$,

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

On the right side of (4), the choice of a concave function q, as introduced in Efron's article to define Q (and ensure $Q \ge 0$), entails $-q''(\widehat{\mu}_i(\mathbf{y}_{(i)}, y_i)) \ge 0$. 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 $\partial \widehat{\mu}_i(\mathbf{y}_{(i)}, y_i)/\partial y_i \ge 0$, (4) indicates that $\partial \widehat{\lambda}_i(\mathbf{y}_{(i)}, y_i)/\partial y_i \ge 0$. 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 $E\{\widehat{\lambda}_i(\mathbf{y}_{(i)}, y_i) \cdot (y_i - \mu_i)|\mathbf{y}_{(i)}\} \ge 0$, which, applied to (3), in turn induces $\mathrm{cov}(\widehat{\lambda}_i, y_i) \ge 0$.
2. On the contrary, if $\partial \widehat{\mu}_i(\mathbf{y}_{(i)}, y_i)/\partial y_i \le 0$, then $\mathrm{cov}(\widehat{\lambda}_i, y_i) \le 0$, 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, $2\,\widehat{\mathrm{cov}}_{(i)} = 2\,\mathrm{cov}_{\widetilde{f}_i}\{(\widehat{\lambda}_i(\mathbf{y}_{(i)}, y_i^*), y_i^*)|\mathbf{y}_{(i)}\}$. Efron shows that $E_{\widetilde{f}_i}\{\widetilde{O}_i^*|\mathbf{y}_{(i)}\} \doteq 2\,\widehat{\mathrm{cov}}_{(i)}$.

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 $2\,\widehat{\mathrm{cov}}_{(i)}$, similar to the term $2\,\widehat{\mathrm{cov}}_i$ in (2), aims to estimate an unknown deterministic quantity, $2\,\mathrm{cov}_i$. Henceforth, it may not strike the reader as particularly surprising that the variance of $\widetilde{O}_i^*$ exceeds that of $2\,\widehat{\mathrm{cov}}_{(i)}$.
2. To better appreciate the value of the covariance penalty method, it would be natural to quantify how much variance reduction is achieved by $2\,\widehat{\mathrm{cov}}_{(i)}$ 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 $\mathrm{var}(\widehat{\mathrm{Err}}_i^{\mathrm{CV}})$ versus $\mathrm{var}(\widehat{\mathrm{Err}}_i^{\mathrm{CP}})$. Generally, the original CV estimate, $\widehat{\mathrm{Err}}_i^{\mathrm{CV}}$, 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 $\mathbf{y}^{*b}$, $b = 1, \dots, B$, 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 $\sum_{i=1}^n \mathrm{cov}_i = \sigma^2 \mathrm{tr}(M)$. 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, $\mathrm{tr}(M_h)$ and $\mathrm{tr}(M_h^T M_h)$, 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, $\widehat{\sigma}^2 = \sum_{i=1}^n (y_i - \widehat{\mu}_i)^2/\{n - \mathrm{tr}(2M_h - M_h^T M_h)\}$ (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 $\mathrm{tr}(M_h) \doteq d\{(p+1-a) + \mathcal{C}n/(n-1)\mathcal{K}(0)|\Omega|/h\}$ and $\mathrm{tr}(M_h^T M_h) \doteq d\{(p+1-a) + \mathcal{C}n/(n-1)\mathcal{K} * \mathcal{K}(0)|\Omega|/h\}$ 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

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

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

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

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

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

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 coordinatewise 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 $\widehat{\mathbf{f}} \sim N(\widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2\mathbf{I})$ was used as in (2.14)–(2.15), with $\widehat{\sigma}^2$ obtained from the

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

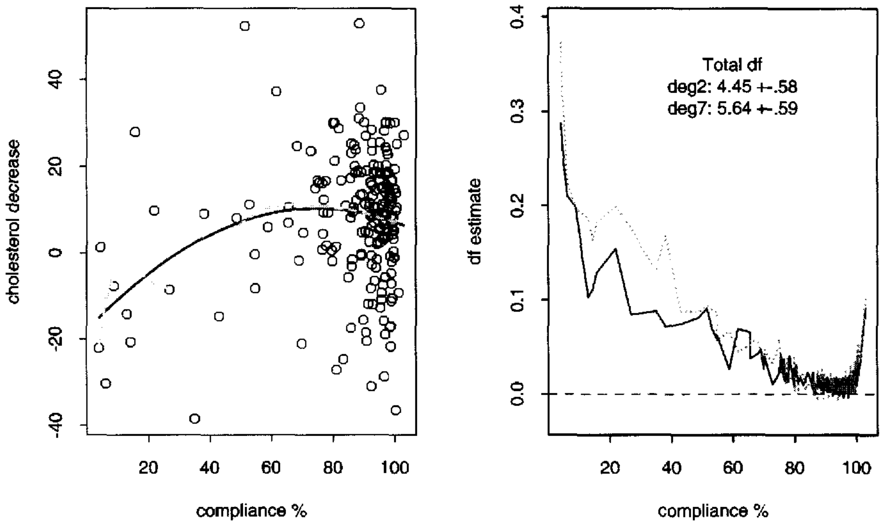

Figure 12. The Left Panel Shows Cholesterol Decrease versus Compliance Percentage for 201 Men in the Control Arm of a Clinical Trial; Quadratic Regression, Solid Curve, Minimized C_p Among Polynomial Predictors; the Dashed Curve Is Seventh-Degree Polynomial Regression; the Right Panel Shows Coordinatewise Parametric Bootstrap df_i Estimates $\widehat{cov}_i/\bar{\sigma}^2$, Bootstrapping From Quadratic Regression (solid line) or From Seventh-Degree Regression (dashed line); Here the Prediction Rule Uses Polynomial Regression With Degree Selected by C_p Minimization.

seventh-degree fit. Two different choices of $\widehat{\mu}$ were tried, from the quadratic fit and also from the seventh-degree polynomial, yielding reassuringly similar results.

Table 3 shows the C_p-selected "best" polynomial degrees in 250 bootstrap replications. Taking $\widehat{\mu}$ as the seventh-degree fit increases the selected degrees, but not drastically so. The point here is that model choice need not be crucial to parametric bootstrap calculations, even under adaptive model selection.

Of course this is just one example, and a simple one at that. Examples and simulations can easily become self-serving in the model selection arena, perhaps because it covers such an enormous range of situations.

The article was careless in its use of the terms "model-based" and "parametric." Almost any regression fit $\widehat{\mu}$ depends heavily on the assumed model, but the same is not necessarily true for estimating its predictive error. The parametric part of the parametric bootstrap is often less than crucial. This point is illustrated by the conditional covariance calculation in figure 5. Notice that the solid curve $m(\mathbf{y}_{(93)}, y^*_{93})$, with $\mathbf{y}_{(93)}$ fixed, *does not depend on the model at all.*

The Taylor series argument (3.23) gives a conditional degrees-of-freedom estimate

$$\widehat{\mathrm{df}}_{(i)} \equiv \widehat{\mathrm{cov}}_{(i)}/\bar{\sigma}^2 \doteq \dot{t}_i, \qquad \dot{t}_i = \left.\frac{\partial m(\mathbf{y}_{(i)}, y^*_i)}{\partial y^*_i}\right|_{\widehat{\mu}_i},$$

$m(\mathbf{y})$ being the lowess function in figure 5. To first order, $\widehat{\mathrm{df}}_{(i)}$ depends on the model only through the abscissa $\widehat{\mu}_i$ of the vertical dashed line (as in Ye 1998). Inspection of figure 1 suggests that any reasonable model will have $0 < \widehat{\mu}_i < 2$ for point $i = 93$, corresponding to $\widehat{\mathrm{df}}_{(i)} \doteq \dot{t}_i$ between roughly .08 and .10. Modeling assumptions are not very important in this case. Linear functions $m(\mathbf{y})$ furnish the extreme example of this phenomenon, where the solid curve in figure 5 becomes a straight line, and $\widehat{\mathrm{df}}_{(i)}$ does not depend on the model at all.

The "DP method" of Shen, Huang, and Ye evaluates $\dot{t}_i$ at $\widehat{\mu}_i = y_i$, yielding $\widehat{\mathrm{df}}_{(i)}$ about .06 in this case. DP tries to compensate for the increased variability from using $\widehat{\mu}_i = y_i$, by reducing the variance of y^*_i as in (2.20). This would not help in figure 5 but worked fine in their table 1. The best that can be said now is that there is a trade-off between "model-free" and estimating efficiency for predictive error, my own preferences being stated in remark A.

Zhang's nice monotonicity result is relevant to figure 5. The solid curve actually has negative slope for y^*_{93} less than -2 or greater than 5 (because lowess suppresses outliers). We might have seen such a value for y_{93}, which would have been troublesome for the DP calculations.

We might expect the equivalent of the solid curve of figure 5 to be discontinuous for a rule $m(\mathbf{y})$ that includes model selection, as in the cholesterol example. However, *usually not* is the actual fact, because changing a single y_i value seldom affects the selected model. This is a weakness of conditional calculations, not a strength. The unconditional parametric bootstrap method "shakes the data" more violently, and more realistically. It gives more honest assessments of prediction error when model selection is a major factor, as in Table 3.

Cross-validation is, of course, a useful tool, as Burman emphasizes. Its appealing rationale, nonparametric character, and easy implementation makes it a popular favorite. That does not obviate concern for its limitations, especially its reduced estimating efficiency. Cross-validation assessments of Err are often acceptable if for no other reason than err, the apparent error, is the main component of any estimate of total predictive error. The simulation results in tables 1 and 2 are perhaps typical. It pays to remember that cross-validation is not even the nonparametric maximum likelihood estimator (MLE) of prediction

Table 3. Polynomial Degrees Selected in 250 Parametric Bootstrap Replications of the C_p-Minimizing Rule for the Two Choices of $\widehat{\mu}$; ± Values for the Total Degrees of Freedom From (2.18). Total Covariance Penalty (2.16) Was Either 4.6% of 5.8% of err

$\widehat{\mu}$	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Total df*
Quadratic fit	3	28	166	21	11	15	2	4	4.45 ± .58
Seventh-degree fit	3	15	138	28	7	21	21	17	5.64 ± .59

error, that being a form of the nonparametric bootstrap (Efron 1983).

Cross-validation is less successful as an estimate of total optimism or degrees of freedom, this being the message of theorem 1 and figure 8. Degrees of freedom is important in its own right, a common currency that allows comparison of disparate estimators. The fact that the lowess rule of figure 2 has about 7 degrees of freedom rather than say 3 or 20, provides helpful intuition about its data-fitting properties. Optimism plays an essential role in model selection (because apparent error always decreases with increased model complexity) so it is reasonable to hope for better selection properties from better optimism estimates. Cross-validation is at its worst in diagnostic plots such as that of figure 2, where its estimates of individual df_i values have coefficients of variation near 2, as in (4.17).

My earlier criticism of conditional methods, that they may not shake the data hard enough to reveal a rule's model selection behavior, applies to cross-validation. Grouped cross-validation with group size $k = 20$ in Burman's notation, was applied to the C_p-minimizing rule of Figure 12. All 10 of the reduced datasets (each of size 181) still had a quadratic polynomial as the C_p minimizer. In this case cross-validation was really estimating the predictive error of an ordinary nonadaptive quadratic fit.

I appreciated the constructive nature of the commentaries, two of which are substantial essays in their own right. Here are a few more responses and remarks:

- The covariance penalty formulas (2.8) and (3.13) do not depend on the components of $\mathbf{y}$ being independent. We would, however, have to take correlation structure into account when implementing $\widehat{\mathbf{f}} \rightarrow \mathbf{y}^*$ in the parametric bootstrap algorithm (2.14), as Zhang points out in his fourth section. Correlation makes my conditional calculations more difficult. In figure 5, for example, the important vertical line would have to be located at $\widehat{E}\{y_i|y_{(i)}\}$ rather than at $\widehat{\mu}_i = \widehat{E}\{y_i\}$.
- "Parametric bootstrap" (PB) sounds exotic, but familiar Fisher information calculations for the variance of an MLE are themselves parametric bootstraps; see section 5 of Efron (1998). The PB algorithm (2.14)–(2.15) provides the MLE of cov_i. Outperforming PB, as does DP in Shen et al.'s table 1, usually involves biased estimates, shrunken toward a favorably chosen origin.
- Denby et al.'s warning about the difficulty of bias estimation is borne out by the top line of table 1; the bottom line is more encouraging. I agree with their points (1), (2), (3), assuming (2) refers to results such as $\widehat{\mathrm{df}}_{(i)} \doteq \dot{t}_i$ mentioned previously. My 1983 article directly concerned the question they raise of improvements on cross-validation, there in the context of fully nonparametric methods. Their "model-based" estimate $\widehat{\mathrm{Err}}_{\mathrm{MB}}$ must not be a covariance penalty rule because the latter would differ more from cross-validation for the linear predictor (1). Notice that it is okay for two error estimates to "agree even if the model is wrong," as long as they are both providing accurate estimates of Err.
- "Computational difficulty" tends to mean programming difficulty in our era of cheap and fast computers. All four corners of figure 11 are equally friendly in this regard because each method merely recomputes the original prediction rule for perturbed datasets "$\mathbf{y}^*$"; the $\mathbf{y}^*$'s are easier to generate for cross-validation while the parametric bootstrap has the advantage of keeping the sample size the same as in the original situation. Cross-validation uses less computer time if blocking is employed, but the parametric bootstrap's greater number of recomputations generates increased information, as in figure 2 or remark B. The relationship here is very much like that between jackknifing and bootstrapping for the accuracy of a point estimate $\widehat{\theta}$.
- In some situations covariance penalties can be calculated theoretically, without recourse to Monte Carlo methods; see Zhang's formulas (5) and (6), and also the "LARS" estimator of Efron, Hastie, Johnstone, and Tibshirani (2004).
- The relationship between optimism and expected optimism, O_i and Ω_i in (3.9)–(3.10), still has a mysterious aspect. The negative correlation phenomenon of figure 9 is endemic and disturbing. A fundamental question, can O_i itself be estimated, remains arguable; see remark K.

My thanks go to the discussants and the editor for focusing attention on the prediction error problem. Model selection has enjoyed a healthy burst of algorithmic growth without a corresponding boom in basic theory, but maybe that is due for a change.

ADDITIONAL REFERENCES

Efron, B. (1998), "R. A. Fisher in the 21st Century" (with discussion), *Statistical Science*, 13, 95–122.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression" (with discussion), *The Annals of Statistics*, 32, 407–499.

Efron, B., and Tibshirani, R. (1998), "The Problem of Regions," *The Annals of Statistics*, 26, 1687–1718.