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

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

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 $\widehat{\boldsymbol{\mu}}$: $\widehat{\text{Err}} = 538.8$ 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

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

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

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

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

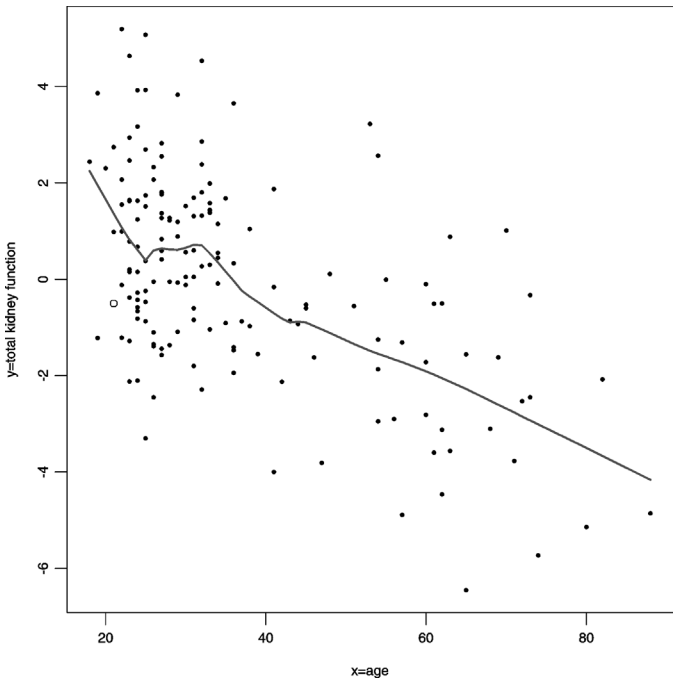

*Figure 1. Kidney Data: An Omnibus Measure of Kidney Function Plotted versus Age for n = 157 Healthy Volunteers. Fitted curve is lowess(**x**, **y**, f = 1/3); sum of squared residuals 495.1. How well can we expect this curve to predict future (x, y) pairs?*

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 $\mathrm{df} \ll n$. 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 $\widehat{\mathbf{f}}$,

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

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

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

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 $\widehat{\epsilon}_j = y_j - \widehat{\mu}_j$

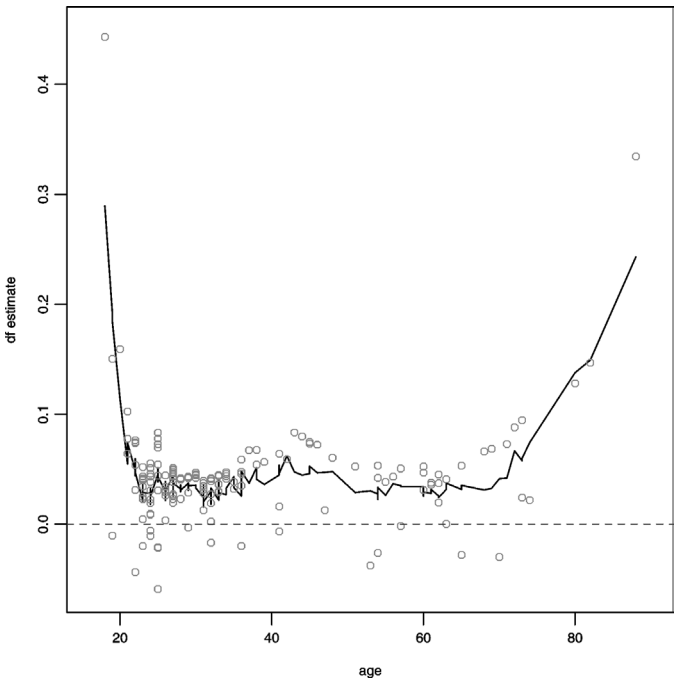

Figure 2. Coordinatewise Degrees of Freedom for Lowess Fit of Figure 1, Plotted versus Age. Open circles, SURE estimate $\widehat{\mathrm{df}}_i = \partial\widehat{\mu}_i/\partial y_i$; solid line, parametric bootstrap estimates $\widehat{\mathrm{cov}}_i/\widehat{\sigma}^2$, (2.14)–(2.15), $B = 1{,}000$. 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 $6.67 \pm .30$ with the bootstrap, the $\pm$ value indicating simulation error, estimated as

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

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

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

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) $\widehat{\mathbf{f}} = \mathrm{N}(\bar{\boldsymbol{\mu}}, \widehat{\sigma}^2\mathbf{I})$, with $\bar{\boldsymbol{\mu}}$ 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

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

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),

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

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

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

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

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

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 Δerr, 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.

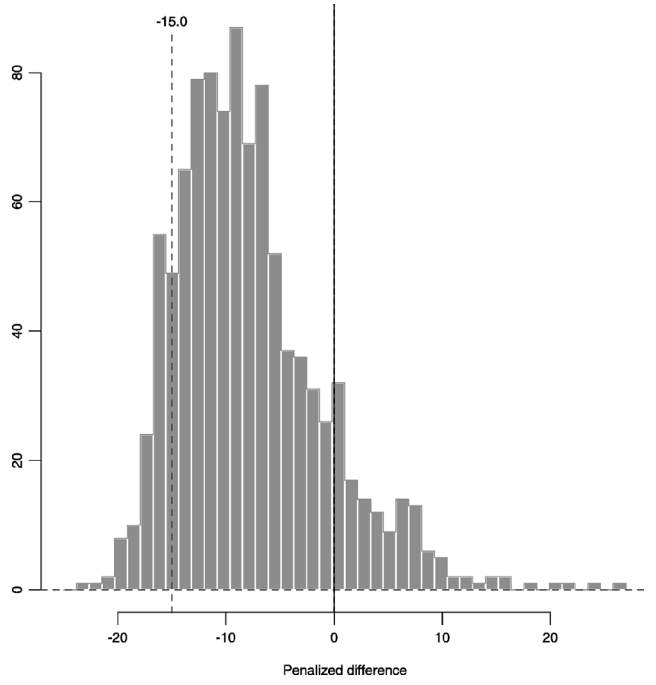

Figure 3. 1,000 Bootstrap Replications of $\Delta\widehat{Err}^*$ for the Difference Between lowess($\mathbf{x}, \mathbf{y}, 2/3$) and lowess($\mathbf{x}, \mathbf{y}, 1/3$), Kidney Data. The point estimate $\Delta\widehat{Err} = -15.0$ is in the lower part of the histogram.

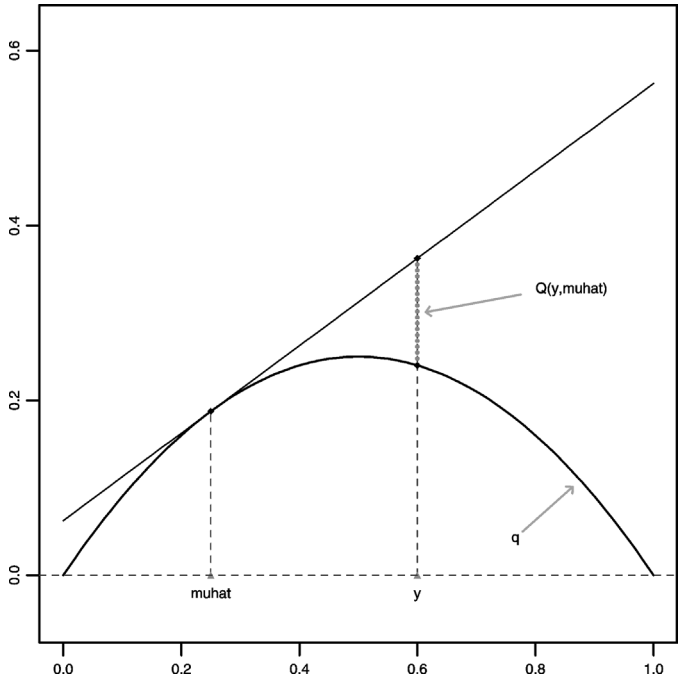

Figure 4. Tangency Construction (3.1) for General Error Measure $Q(y, \widehat{\mu})$; $q(\cdot)$ Is an Arbitrary Concave Function. The illustrated case has $q(\mu) = \mu(1-\mu)$ and $Q(y, \widehat{\mu}) = (y - \widehat{\mu})^2$.

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 $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, 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 $\mathbf{y} \sim \mathrm{N}(\widehat{\boldsymbol{\mu}}, \sigma^2\mathbf{I})$, 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 $\mathbf{y}^* \sim \mathrm{N}(\mathbf{y}, \sigma^2\mathbf{I})$ 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 $\mathbf{f} \to \mathbf{y}$. 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*

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

$E_{(i)}$ 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),

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

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 $\mathbf{y}^* \sim \text{N}(\widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2\mathbf{I})$, $\widehat{\sigma}^2 = 3.17$, 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.

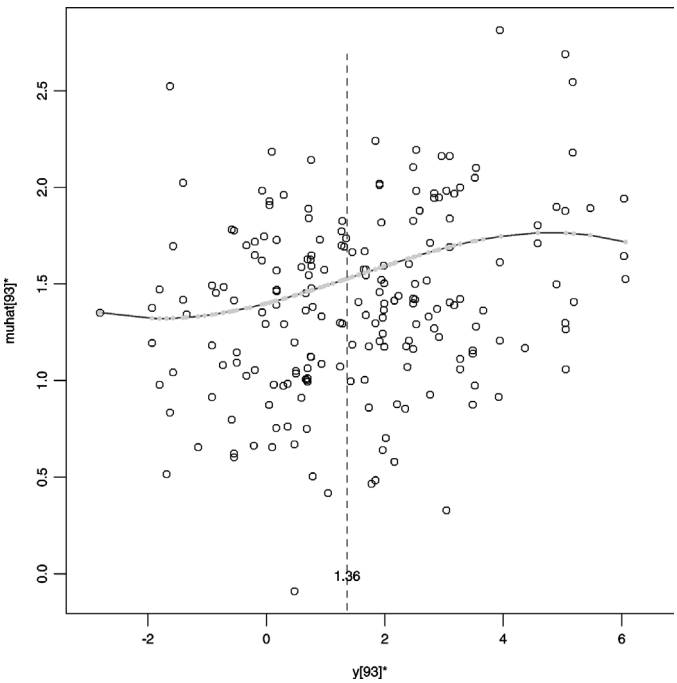

Figure 5. Conditional and Unconditional Covariance Calculations for Subject $i = 93$, Kidney Study. Open circles: 200 pairs $(y_i^*, \widehat{\mu}_i^*)$, unconditional resamples $\mathbf{y}^* \sim N(\widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2\mathbf{I})$; $\widehat{\text{cov}}_i = .218$; Dots: 100 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 $\mathbf{y} \sim \text{Be}(\boldsymbol{\mu})$. In this situation it is easy to see that

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

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

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

Calculating $\widehat{\text{cov}}_{(i)}$ 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 $\text{cov}_{(i)}$. 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 $\widehat{\text{cov}}_{(i)}$,

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

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

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

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):

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

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 $\text{err}_i = -2\log f_{\widehat{\mu}_i}(y_i) +$ constant. Combining (3.14), (3.24), and (3.25) gives

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

Choosing among competing models by minimizing $\widehat{\text{Err}}$ 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

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

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

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

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 $\rho = (\pi_0/\pi_1)^{1/2}$,

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

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

$$\widetilde{\text{Err}}_i = Q(y_i, \widetilde{\mu}_i), \tag{4.1}$$

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

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

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

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

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 $\mathrm{cov}_{(i)}$ 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

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

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

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,

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

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

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

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,

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

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

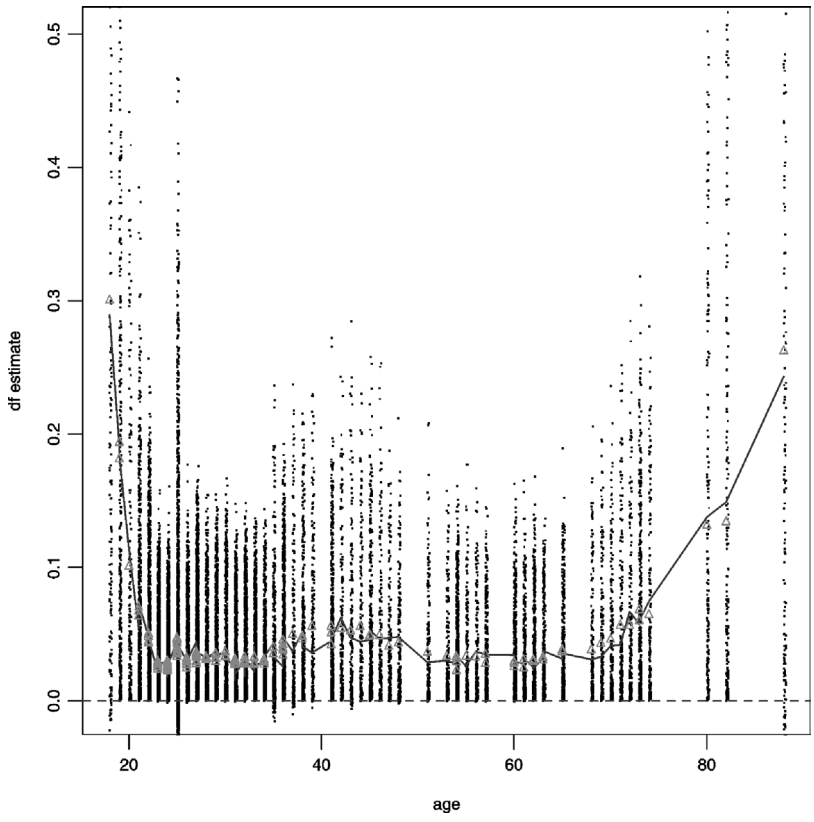

Figure 6. Small Dots Indicate 200 Bootstrap Replications of Cross-Validation Optimism Estimates (4.9); Triangles, Their Averages, Closely Match the Covariance Penalty Curve From Figure 2. (Vertical distance plotted in df units.)

penalty curve $\widehat{\mathrm{cov}}_i/\widehat{\sigma}^2$ 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 $\widetilde{O}_i^* = \sum_j \widetilde{O}_{ij}^*$, $\widehat{\mathrm{cov}}_{(i)} = \sum_j \widehat{\mathrm{cov}}_{(ij)}$, 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,

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

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

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

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

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

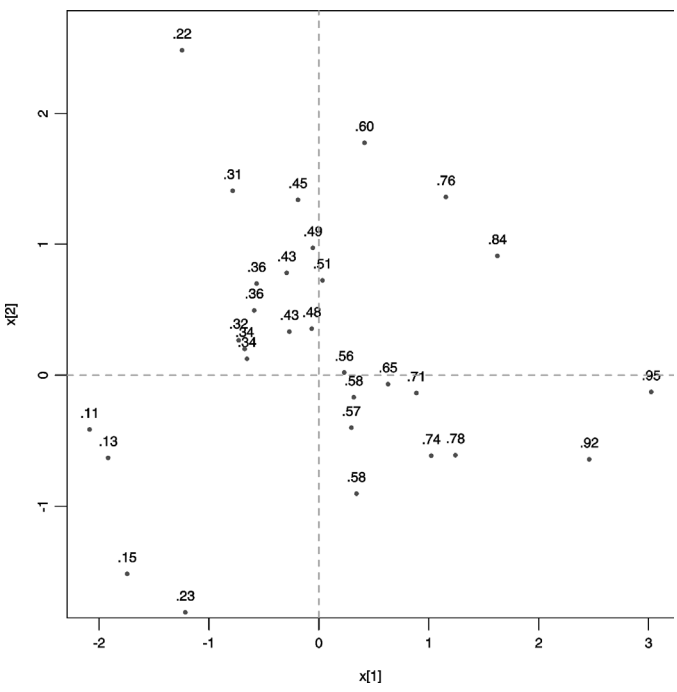

Figure 7. Underlying Model Used for Simulation Study: $n = 30$ Bivariate Vectors x_i and Associated Probabilities μ_i, (5.1).

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 $\widehat{\alpha} + \widehat{\beta}' x = 0$:

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

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,

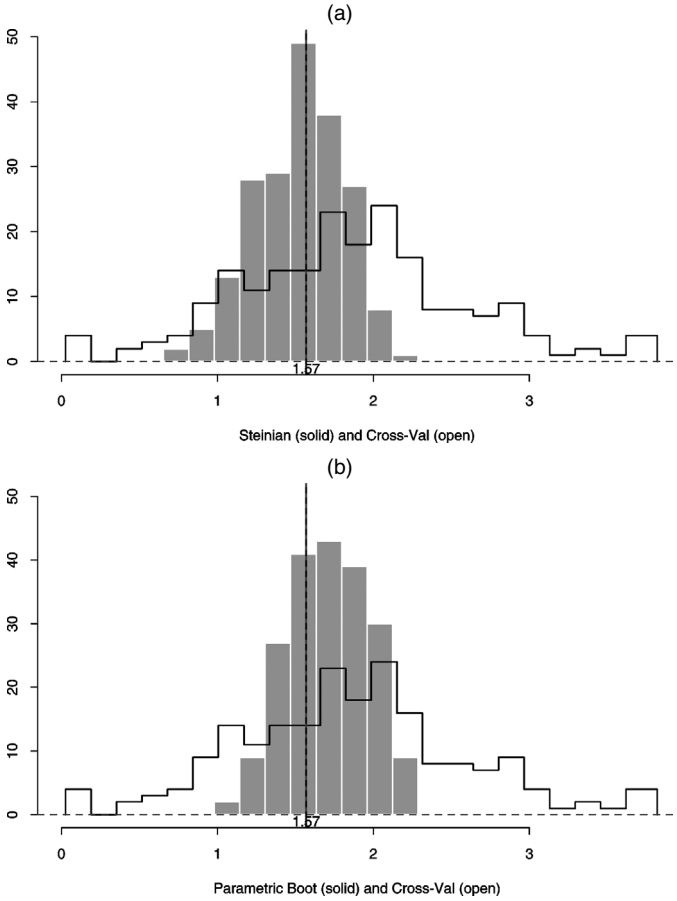

Figure 8. Degree-of-Freedom Estimates (optimism/2); 200 Simulations (5.2) and (5.4). The two covariance penalty estimates, Steinian (a) and parametric bootstrap (b), have about one-third the standard deviation of cross-validation. Error measured by binomial deviance (3.4); true $\Omega/2 = 1.57$.

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 $\boldsymbol{\mu}$ 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 $\mathbf{y}$,

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

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 $\text{Err} = \overline{\text{err}} + O$ with $\widehat{\text{Err}} = \overline{\text{err}} + \widehat{\Omega}$, then

$$\widehat{\text{Err}} - \text{Err} = \widehat{\Omega} - O, \tag{5.6}$$

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

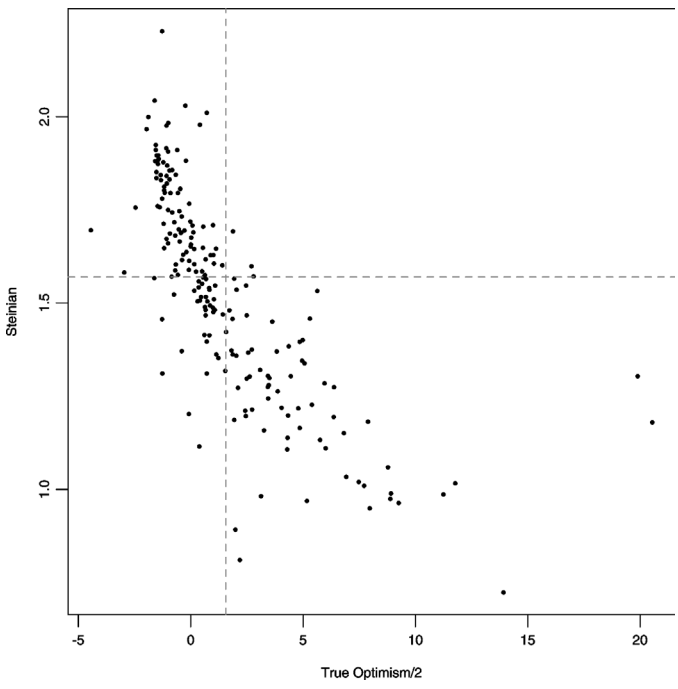

Figure 9. Steinian Estimate versus True Optimism/2, (5.5), for the 200 Simulations. Similar inverse relationships hold for parametric bootstrap or cross-validation.

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

	Steinian	*ParBoot*	*CrossVal*	*Apparent*
$E(\widehat{\text{Err}} - \text{Err})^2$	**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 $E(\widehat{\text{Err}} - \text{Err})^2$ for the prediction rule "ordinary logistic regression," rather than (5.4). Now all three methods handily beat the apparent error. The average prediction $\widehat{\text{Err}}$ 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. $\text{Err} = \sum \text{Err}_i$, (3.8), is the total prediction error at the n observed covariate points x_i. "Outsample error,"

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

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 Err_{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 $|\widehat{\text{Err}} - \text{Err}|$ 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 $|\widehat{\text{Err}} - \text{Err}_{\text{out}}|$, 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 $\mathbf{v}^* = (v_1^*, v_2^*, \ldots, v_n^*)$ 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 $\mathbf{v}^{*b}$, $b = 1, 2, \ldots, B$; define the indicator

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

$h = 0, 1, \ldots, n$; 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 $B(h)$:

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

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

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

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 $B(h) = (1, 0, 0, \ldots, 0)$ 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 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ 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 $\widehat{\mathrm{cov}}_{(i)} = \widehat{\sigma}^2 M_{ii}(h)$. 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 $[1 + \mathrm{O}(1/n)]$.

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 $\sim$ 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 $\bar{Q}_i(h)^{*a}$,

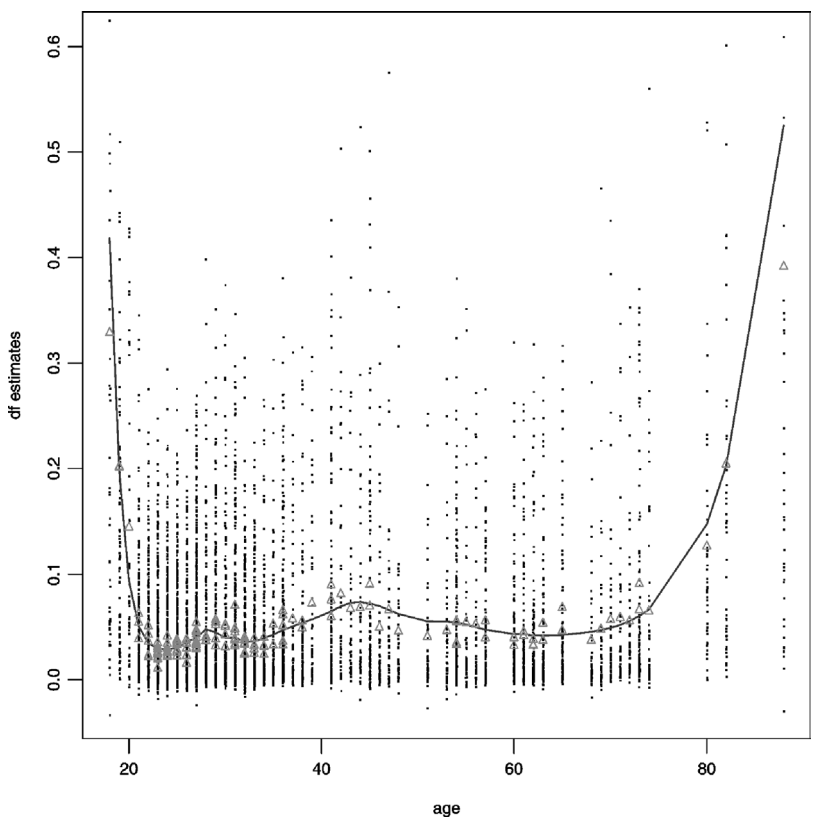

Figure 10. Small Dots Indicate 50 Parametric Bootstrap Replications of Unconditional Nonparametric Optimism Estimates (6.3); Triangles, Their Averages, Closely Follow the Covariance Penalty Estimates (solid curve). Vertical distance plotted in df units. Here the estimation rule is loess(span = .5). See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

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

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.

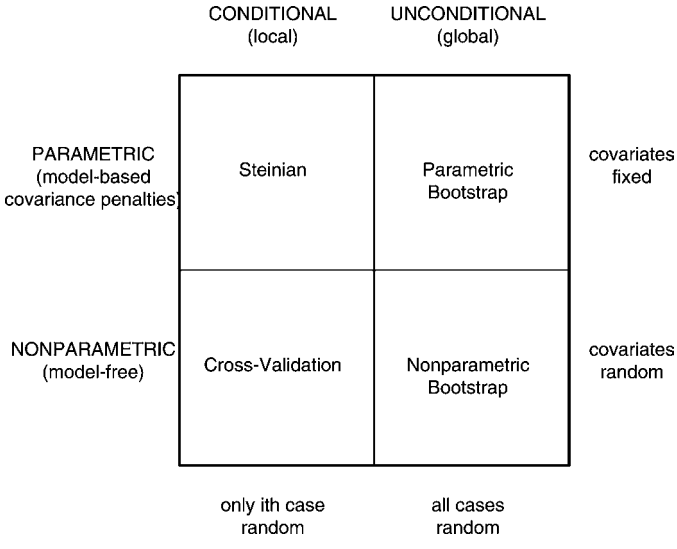

Figure 11. Two-Way Classification of Prediction Error Estimates Discussed in This Article. The conditional methods are local in the sense that only the ith case data are varied in estimating the ith error rate.

  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*).

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