On the Optimality of Prediction Based Selection Criteria and the Convergence Rates of Estimators

by

Naomi Altman
and
Christian Leger

BU-1230-MB August 1995

On the Optimality of Prediction Based Selection Criteria and the Convergence Rates of Estimators

Naomi Altman *
Biometrics Unit
Cornell University

Christian Léger †
Département de mathématiques et de statistique
Université de Montréal

August 1995

Abstract

A number of estimators of squared prediction error have been suggested for use in model and bandwidth selection problems. Among these are cross-validation, generalized cross-validation and a number of related techniques based on the residual sum of squares. For a number of situations with squared error loss, for example nonparametric smoothing, these estimators have been shown to be asymptotically optimal in the sense that in large samples the estimator minimizing the selection criterion also minimizes squared error loss. However, cross-validation is known not to be asymptotically optimal for some "easy" location problems. In this article, we consider selection criteria based on estimators of squared prediction risk for choosing among location estimators. We show that criteria based on adjusted residual sum of squares are not asymptotically optimal for choosing between asymptotically Normal location estimators that converge at rate $n^{1/2}$, but are when the rate of convergence is slower. We also show that leave-one-out cross-validation is not asymptotically optimal for choosing between $\sqrt{n}$-differentiable statistics but leave-d-out cross-validation is optimal when $d \to \infty$ at the appropriate rate.

Keywords: cross-validation, trimmed means, differentiable statistics, variable selection, generalized cross-validation, consistency, location estimation

*Supported by Hatch Grant 151410 NYF

†Supported by NSERC (Canada) and FCAR (Québec)

1 Introduction

Many modern data modelling techniques are adaptive in the sense that, rather than depending on a parametric model, they involve only mild regularity conditions and a family of estimators indexed by a tuning parameter which determines its properties. Under squared error loss, prediction risk and estimation risk differ by a constant independent of the estimator. Accordingly, many popular selection criteria, such as cross-validation (CV) and adjusted residual sum of squares, are based on minimizing estimators of prediction loss or risk.

Härdle and Marron (1985) showed that CV is asymptotically optimal for choice of bandwidth in kernel regression estimation, while Härdle, Hall and Marron (1988) showed the asymptotic optimality of a large class of squared prediction loss estimators based on adjusted residual sum of squares in the context of regression smoothing. However, Stone (1977) showed that CV is not asymptotically optimal for the "easier" problem of choosing between the mean and median in the Normal location problem with squared error loss, and Pruitt (1988) extended this result to show that CV is not asymptotically optimal for selecting the best trimming proportion in an adaptive trimmed mean.

In this article we show the relationship between the rate of convergence of the family of estimators and the optimality of prediction based selection criteria in the location problem. We consider a family including a finite number of asymptotically Normal location estimators (with rate n^p, $p \leq 1/2$, where n is the sample size). The lessons that can be learned in this simple context shed light on the behavior of similar methods in more complex situations. Prediction loss differs from estimation loss by a term depending only on the data and a cross-product term depending on the tuning parameter which converges to zero. Optimality is possible only if the cross-product term is of smaller order than the estimation loss term and that in turn depends on the rate of convergence of the estimator.

The paper is organized as follows. In Section 2, we show that selection criteria based on adjusted residual sum of squares or an independent validation sample of the same size are asymptotically optimal if and only if $p < 1/2$. We also show that if the estimators satisfy a weak differentiability condition, leave-one-out CV is not asymptotically optimal, thus generalizing results of Stone (1977) and Pruitt (1988). On the other hand, we show that leave-d-out CV

is asymptotically optimal if d/n converges to 1 while $n - d \to \infty$. This result is similar to Shao (1993) in the context of model selection in regression. Finally, it is argued that bootstrap estimates of risk, because they are expectations, do not include a cross-product term and so do not suffer from this problem. Section 3 reports on the results of a small simulation study examining the use of leave-d-out CV to choose between $\sqrt{n}-$convergent location estimators and between $\sqrt{n}$-convergent regression estimators. Section 4 summarizes our conclusions.

2 Non-optimality of Sum of Squared Prediction Error Estimators

Consider the location problem where $y_1, \ldots, y_n$ are identically and independently distributed (i.i.d.) from a distribution F with mean θ and variance σ^2 and let $\hat{\theta}(\mathbf{y})$ be any estimator of θ depending on the vector of data $\mathbf{y}$. We wish to select an estimator that minimizes the value of the squared error estimation loss:

$$(\hat{\theta}(\mathbf{y}) - \theta)^2. \tag{1}$$

Suppose that $\tilde{\theta}(\mathbf{y})$ is the minimizer of a selection criterion in a class of estimators Ω. We say that the selection criterion is asymptotically loss optimal if

$$\frac{(\tilde{\theta}(\mathbf{y}) - \theta)^2}{\min_{\Omega}(\hat{\theta}(\mathbf{y}) - \theta)^2} \xrightarrow{P} 1. \tag{2}$$

We also consider the weaker condition of risk optimality: a selection criterion is asymptotically risk optimal if the probability of selecting the estimator with the smallest risk goes to 1.

Li (1987) shows that asymptotic risk optimality implies asymptotic loss optimality in the context of bandwidth selection for nonparametric regression, but this need not always be the case. Consider estimating the mean of a Normal distribution using the sample mean or median. Clearly, the mean is asymptotically risk optimal. To see that there is no asymptotically loss optimal estimator for this case, note that the sample mean and median are asymptotically jointly normal. Assuming that the population mean is 0, loss optimality implies that for any $\epsilon > 0$, the probability of the two upper and upside down "V" regions delimited by the lines with slopes $1 + \epsilon$ and $-1 - \epsilon$ tends to 1 as $n \to \infty$. But that probability tends to the corresponding normal probability given by the joint asymptotic distribution which is different from 1.

Many selection criteria are based on minimizing an estimator of the squared error prediction loss:

$$[y^* - \hat{\theta}(\mathbf{y})]^2 = (y^* - \theta)^2 - 2(y^* - \theta)(\hat{\theta}(\mathbf{y}) - \theta) + (\hat{\theta}(\mathbf{y}) - \theta)^2 \tag{3}$$

where y^* is a new datum from the distribution. Note that because $E(y^* - \theta)(\hat{\theta}(\mathbf{y}) - \theta) = 0$, prediction risk and estimation risk differ by a constant not depending on $\hat{\theta}(y)$, but that the corresponding losses differ by a cross-product term that does depend on the location estimator. The main goal of this paper is to show that asymptotic optimality is only possible if the cross-product term converges faster than the estimation loss term. This, in turn, depends on the rate of convergence of the location estimator.

We consider three classes of estimators of squared error prediction loss for the purpose of choosing among a family of location estimators. The (leave-one-out) cross-validation (CV) estimator of average squared error prediction loss (Stone, 1974) is:

$$CV[\hat{\theta}_n(\mathbf{y})] = 1/n \sum_{i=1}^{n} [y_i - \hat{\theta}_{n-1}^{-i}(\mathbf{y})]^2, \tag{4}$$

where the subscript denotes the sample size on which the estimator is based, and the superscript $-i$ indicates that the i^{th} datum was not used in computing the estimator. (Generally we will suppress the subscript for $\hat{\theta}$, which always uses the full sample.) Intuitively, CV estimates prediction risk as a mean of prediction losses, each based on predicting y_i from the remainder of the data. More formally, $E(y_i - \hat{\theta}_{n-1}^{-i})^2 = \sigma^2 + E(\hat{\theta}_{n-1} - \theta)^2$, and for consistent estimators, $E(\hat{\theta}_{n-1} - \theta)^2$ and $E(\hat{\theta}_n - \theta)^2$ are close.

A second class of prediction risk estimators is based on adjusted residual sum of squares. The residual sum of squares has expectation

$$E\left(\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\theta}(\mathbf{y})]^2\right) = \sigma^2 + E\left([\hat{\theta}(\mathbf{y}) - \theta]^2\right) - 2\,Covariance[\bar{y}, \hat{\theta}(\mathbf{y})] \tag{5}$$

Estimates of average squared error prediction loss based on adjusted residual sum of squares have the form:

$$ARS(\hat{\theta}(\mathbf{y})) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\theta}(\mathbf{y})]^2 \Xi[n, \hat{\theta}(\mathbf{y})] \tag{6}$$

where $\Xi[n, \hat{\theta}(\mathbf{y})] = 1 + 2\,Covariance[\bar{y}, \hat{\theta}(\mathbf{y})]/\sigma^2 + o_p(1/n^{p+1/2})$ (in the case where $\hat{\theta}(\mathbf{y})$ converges at the rate n^p). This class has been used mainly in a regression context where it includes

generalized cross-validation (Craven and Wahba, 1979), Akaike's information criterion (Akaike 1974) and a large number of other selectors (for example, Härdle, Hall and Marron, 1988). Under appropriate assumptions, the term $2\frac{1}{n}\sum_{i=1}^{n} y_i^2\, Covariance[\bar{y}, \hat{\theta}(\mathbf{y})]/\sigma^2$ cancels the Covariance term in (5), so that the bias of the estimator is of smaller order than the prediction error.

Adjusted residual sum of squares and CV are averages of n prediction risk estimators, one for each data value. As a third alternative we consider average prediction loss of the form,

$$\begin{aligned} APL[\hat{\theta}(\mathbf{y})] &= \frac{1}{n}\sum_{i=1}^{n}[y_i^* - \hat{\theta}(\mathbf{y})]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}[y_i^* - \theta]^2 - 2\frac{1}{n}\sum_{i=1}^{n}[y_i^* - \theta][\hat{\theta}(\mathbf{y}) - \theta] + [\hat{\theta}(\mathbf{y}) - \theta]^2, \end{aligned} \tag{7}$$

where $y_1^*, \ldots, y_n^*$ is an i.i.d. validation sample from the same distribution as $y_1, \ldots, y_n$ and independent of them. Note that the cross-product term in this expansion has expectation zero, and that for reasonable estimators, $E(\theta - \hat{\theta}_n)^2$ converges to zero with n. If this term converges more rapidly than estimation loss then average prediction loss differs from estimation loss by a constant (which depends on the realization but not the estimator) and a negligible term so average prediction loss may be asymptotically loss optimal.

Theorem 1 shows that asymptotic loss optimality of average prediction loss (7) for choosing between two consistent asymptotically Normal estimators depends on their rate of convergence. In particular, if they converge slowly (rate less than $\sqrt{n}$) average prediction loss is asymptotically loss optimal, but if they converge rapidly it is not even risk optimal.

Theorem 1 *Suppose $y_1, \ldots, y_n, y_1^*, \ldots, y_n^*$ is an i.i.d. sample from distribution F with mean θ and finite variance σ_y^2. Suppose $\hat{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$ are both estimators of θ such that*

$$n^p \begin{pmatrix} \hat{\theta}_1(\mathbf{y}) - \theta \\ \hat{\theta}_2(\mathbf{y}) - \theta \end{pmatrix} \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix} \right)$$

where σ_i^2 is the asymptotic variance of $\hat{\theta}_i$ and ρ is the asymptotic correlation between the estimators. Assume also that $|\rho| \neq 1$. Then average prediction loss is asymptotically loss optimal if $p < 1/2$. If $p = 1/2$ then average prediction loss is not asymptotically risk optimal.

Proof: Without loss of generality, we may assume $\theta = 0$. Consider the difference in estimation loss

$$\begin{aligned} DEL[\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})] &= \hat{\theta}_1(\mathbf{y})^2 - \hat{\theta}_2(\mathbf{y})^2 \\ &= [\hat{\theta}_1(\mathbf{y}) - \hat{\theta}_2(\mathbf{y})][\hat{\theta}_1(\mathbf{y}) + \hat{\theta}_2(\mathbf{y})] \end{aligned}$$

and the difference in average prediction loss

$$\begin{aligned} DAPL[(\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})] &= -2\bar{y}^*[\hat{\theta}_1(\mathbf{y}) - \hat{\theta}_2(\mathbf{y})] + \hat{\theta}_1(\mathbf{y})^2 - \hat{\theta}_2(\mathbf{y})^2 \\ &= [\hat{\theta}_1(\mathbf{y}) - \hat{\theta}_2(\mathbf{y})][-2\bar{y}_n^* + \hat{\theta}_1(\mathbf{y}) + \hat{\theta}_2(\mathbf{y})] \end{aligned}$$

where $\bar{y}_n^* = \frac{1}{n}\sum_{i=1}^n y_i^*$.

By the asymptotic normality of the estimators,

$$n^p \begin{pmatrix} \hat{\theta}_1(\mathbf{y}) - \hat{\theta}_2(\mathbf{y}) \\ \hat{\theta}_1(\mathbf{y}) + \hat{\theta}_2(\mathbf{y}) \end{pmatrix} \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho & \sigma_1^2 - \sigma_2^2 \\ \sigma_1^2 - \sigma_2^2 & \sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2\rho \end{bmatrix} \right)$$

so $n^{2p} DEL[\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})]$ converges in distribution to a product of correlated Normals, which has support on the entire line.

Also $\sqrt{n}\bar{y}^* \xrightarrow{D} N(0, \sigma_y^2)$. If $p < 1/2$ then $n^p[-2\bar{y}_n^* + \hat{\theta}_1(\mathbf{y}) + \hat{\theta}_2(\mathbf{y})] = n^p[\hat{\theta}_1(\mathbf{y}) + \hat{\theta}_2(\mathbf{y})] + o_p(n^p)$ giving $DAPL[(\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})]/DEL[\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})] \xrightarrow{P} 1$ which is equivalent to asymptotic loss optimality.

However, if $p = 1/2$ then $nDAPL[(\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})] = nDEL[\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})] - 2n\bar{y}^*[\hat{\theta}_1(\mathbf{y}) - \hat{\theta}_2(\mathbf{y})]$ and the second term in this sum is not negligible. By definition $\bar{y}^*$ and $\hat{\theta}_1(\mathbf{y}) - \hat{\theta}_2(\mathbf{y})$ are uncorrelated, so this term goes to a product of independent Normals, which is symmetric about zero. Thus $DAPL[(\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})]/E(DEL[\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})])$ converges to a random variable.

Remark 1 Theorem 1 covers most cases of interest including many L- and M-estimators (Serfling 1980), and in particular, the mean, the median and trimmed means. Even in the optimistic case where we can use a validation sample of the same size, average prediction loss cannot be used to choose among two $\sqrt{n}$−convergent location estimators. The reason for this is the rate of convergence of the mean of the validation sample which is the same as that of $\hat{\theta}(\mathbf{y})$. To match adjusted residual sum of squares methods and leave-one-out CV, we have chosen the

sizes of the validation and estimation samples to be the same. Suppose instead we choose the validation sample to be of size n^ν where ν is a constant larger than $2p$. Then for sample size n we have $n^{\nu/2}\bar{y}^*_{n^\nu} \xrightarrow{D} N(0,\sigma_y^2)$ and $DAPL[(\hat{\theta}_1(\mathbf{y}),\hat{\theta}_2(\mathbf{y}))] = [\hat{\theta}_1(\mathbf{y})-\hat{\theta}_2(\mathbf{y})][-2\bar{y}^*_{n^\nu}+\hat{\theta}_1(\mathbf{y})+\hat{\theta}_2(\mathbf{y})]$. For this larger validation sample (in the case $p = 1/2$), $n^p[-2\bar{y}^*_{n^\nu} + \hat{\theta}_1(\mathbf{y}) + \hat{\theta}_2(\mathbf{y})] = n^p[\hat{\theta}_1(\mathbf{y}) + \hat{\theta}_2(\mathbf{y})] + o_p(n^p)$ and we have asymptotic loss optimality for choosing between $\hat{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$.

Remark 2 For an example of a class of location estimators with an asymptotically Normal distribution and $p < 1/2$, see the kernel estimates of the mode studied by Romano (1988).

Theorem 1 suggests that selection criteria based on average prediction loss may not be asymptotically loss optimal. Theorem 2 shows that prediction risk estimators based on adjusted residual sum of squares (Equation 6) are not asymptotically risk optimal for $\sqrt{n}$-convergent estimators under the conditions of Theorem 1 plus additional regularity conditions.

Theorem 2 *Suppose* $y_1,\ldots,y_n$, F, θ, $\hat{\theta}_1(\mathbf{y})$, $\hat{\theta}_2(\mathbf{y})$ *and* ρ *satisfy the conditions of Theorem 1 and that* $Covariance[\bar{y},\hat{\theta}_i(\mathbf{y})] = \sigma^2 K_i/n^{p+1/2}+o(1/n^{p+1/2})$. *Assume also that* $Correlation[\bar{y},\hat{\theta}_1(\mathbf{y})-\hat{\theta}_2(\mathbf{y})] \not\to \pm 1$. *Then prediction risk estimators of the form (6) are asymptotically loss optimal in choosing between* $\hat{\theta}_1(\mathbf{y})$ *and* $\hat{\theta}_2(\mathbf{y})$ *if* $p < 1/2$ *and are not asymptotically risk optimal if* $p = 1/2$.

Proof: Without loss of generality, we may assume $\theta = 0$. Consider the difference in the prediction risk estimators:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}[y_i-\hat{\theta}_1(\mathbf{y})]^2\Xi[n,\hat{\theta}_1(\mathbf{y})] \;-\; & \frac{1}{n}\sum_{i=1}^{n}[y_i-\hat{\theta}_2(\mathbf{y})]^2\Xi[n,\hat{\theta}_2(\mathbf{y})] \\
= \;& \hat{\theta}_1(\mathbf{y})^2-\hat{\theta}_2(\mathbf{y})^2-2\bar{y}[\hat{\theta}_1(\mathbf{y})-\hat{\theta}_2(\mathbf{y})]+2\frac{\sum_{i=1}^n y_i^2}{n^{3/2+p}}(K_1-K_2) \\
& +o_p\left[\hat{\theta}_1(\mathbf{y})^2-\hat{\theta}_2(\mathbf{y})^2+\bar{y}[\hat{\theta}_1(\mathbf{y})-\hat{\theta}_2(\mathbf{y})]+\frac{\sum_{i=1}^n y_i^2}{n^{3/2+p}}\right] \\
= \;& [\hat{\theta}_1(\mathbf{y})-\hat{\theta}_2(\mathbf{y})][\hat{\theta}_1(\mathbf{y})+\hat{\theta}_2(\mathbf{y})-2\bar{y}]+2\frac{\sum_{i=1}^n y_i^2}{n^{3/2+p}}(K_1-K_2) \\
& +o_p\left[\hat{\theta}_1(\mathbf{y})^2-\hat{\theta}_2(\mathbf{y})^2+\bar{y}[\hat{\theta}_1(\mathbf{y})-\hat{\theta}_2(\mathbf{y})]+\frac{\sum_{i=1}^n y_i^2}{n^{3/2+p}}\right].
\end{aligned}$$

Then for $p < 1/2$ we find that $n^p[-2\bar{y} + \hat{\theta}_1(\mathbf{y}) + \hat{\theta}_2(\mathbf{y})] = n^p[\hat{\theta}_1(\mathbf{y}) + \hat{\theta}_2(\mathbf{y})] + o_p(n^p)$ and $n^{2p}\sum_{i=1}^n y_i^2/n^{3/2+p} \to 0$ so the prediction risk estimator is asymptotically loss optimal.

If $p = 1/2$, the difference in risk estimators is

$$DEL \quad [\ \hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})] - 2\bar{y}[\hat{\theta}_1(\mathbf{y}) - \hat{\theta}_2(\mathbf{y})] + \frac{2}{n^2}\sum_{i=1}^{n} y_i^2(K_1 - K_2)$$
$$+o_p\left(DEL[\hat{\theta}_1(\mathbf{y}), \hat{\theta}_2(\mathbf{y})] + \bar{y}[\hat{\theta}_1(\mathbf{y}) - \hat{\theta}_2(\mathbf{y})] + \sum_{i=1}^{n} y_i^2/n^2\right).$$

and $n\bar{y}[\hat{\theta}_1(\mathbf{y}) - \hat{\theta}_2(\mathbf{y})]$ goes to a product of correlated Normals which has support on $(-\infty, \infty)$ as long as the correlation does not have absolute value 1. This shows the risk estimator is not asymptotically risk optimal.

We now turn to the question of asymptotic loss optimality of CV. Expanding CV in terms corresponding to those of average prediction loss we obtain:

$$CV[\hat{\theta}(\mathbf{y})] = 1/n\sum_{i=1}^{n}(y_i - \theta)^2 - 2/n\sum_{i=1}^{n}(y_i - \theta)(\hat{\theta}_{n-1}^{-i} - \theta) + 1/n\sum_{i=1}^{n}(\theta - \hat{\theta}_{n-1}^{-i})^2. \tag{8}$$

The three terms in (8) estimate the corresponding terms in (7). In light of Theorems 1 and 2 it is expected that leave-one-out CV will not be asymptotically loss or risk optimal for $p = 1/2$. On the other hand, Remark 1 suggests that for d increasing with n, leave-d-out CV might be asymptotically risk optimal. We first need some notation.

For a fixed n, let $d = d_n$ be an integer less than n and $r = n - d$. Following Shao and Wu (1989), define $S_{n,r}$ to be the collection of subsets of $\{1, \ldots, n\}$ which have size r. For any $S = \{i_1, \ldots, i_r\} \in S_{n,r}$, let $\hat{\theta}^S = \hat{\theta}(y_{i_1}, \ldots, y_{i_r})$. The leave-$d$-out CV estimator of risk is

$$CV_d(\hat{\theta}_j) = \frac{1}{dN}\sum_{S}\left(\sum_{i \in S^C}(y_i - \hat{\theta}_j^S)^2\right),$$

where S^C is the complement of the set S and $N = \binom{n}{r}$ is the number of subsets of size r.

Shao (1993) has studied risk optimality for the problem of estimating the "true" model in a linear regression model. He discusses the well-known fact that with probability tending to 1, leave-1-out cross-validation (also known as PRESS), adjusted residual sum of squares and other asymptotically equivalent criteria such as C_p, will include all variables with non-zero coefficients, but may include more variables. He provides a convergent estimator of the "true" model based on leave-d-out CV with $r \to \infty$ and $r/n \to 0$. As with the location problem, the regression estimators are $\sqrt{n}$-convergent.

We now show that, for choosing among $\sqrt{n}$-differentiable location estimators satisfying a weak differentiability condition, leave-d-out CV is asymptotically risk optimal under the conditions on r introduced by Shao.

Theorem 3 *Suppose $y_1, \ldots, y_n$ is an i.i.d. sample from distribution F with mean θ and finite variance σ_y^2. Suppose $\hat{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$ are both estimators of θ such that*

$$\hat{\theta}_j = \hat{\tau}_j + R_{n,j}$$

where $\hat{\tau}_j = \frac{1}{n}\sum_{i=1}^n h_j(y_i)$, $E[h_j(y_i)] = \theta$, $Var[h_j(y_i)] = \sigma_j^2 < \infty$. Assume that $E(R_{n,j}^2) = o(1/n)$, $r \to \infty$, and $r/n \to 0$. Then with probability converging to 1, the leave-d-out cross-validation criterion will choose the estimator with smallest asymptotic variance.

Proof: Without loss of generality, we may assume that $\theta = 0$. Let $\hat{\tau}_j^S = 1/r\sum_{i\in S} h_j(y_i)$, $R_{n,j}^S = \hat{\theta}_j^S - \hat{\tau}_j^S$ and $U_j^S = R_{n,j}^S - R_{n,j}$. It can be shown that

$$\begin{aligned} CV_d(\hat{\theta}_1(\mathbf{y})) \quad - \quad & CV_d(\hat{\theta}_2(\mathbf{y})) = \frac{1}{r}s_{h,1}^2 - \frac{1}{r}s_{h,2}^2 && (9)\\ & + \frac{1}{n(n-1)}\sum\sum_{i\neq k} h_1(y_i)h_1(y_k) - \frac{1}{n(n-1)}\sum\sum_{i\neq k} h_2(y_i)h_2(y_k) && (10)\\ & - \frac{2}{n(n-1)}\sum\sum_{i\neq k} y_i h_1(y_k) + \frac{2}{n(n-1)}\sum\sum_{i\neq k} y_i h_2(y_k) && (11)\\ & + o_p\left(\frac{1}{r}s_{h,1}^2\right) \end{aligned}$$

where $s_{h,j}^2 = 1/n\sum_{i=1}^n (h_j(y_i) - \bar{h}_j)^2$, $\bar{h}_j = 1/n\sum_{i=1}^n h_j(y_i)$. Since $s_{h,j}^2$ converges a.e. to the asymptotic variance of $\hat{\theta}_j$, (9) is $O_P(1/r)$. Terms (10) and (11) are U-estimators and are $O_P(1/n)$. Since $r/n \to 0$, these terms are of smaller order. The largest remaining terms are $O_P(E(U_j^S)^2) = o_P(1/r)$ and $o_P(1/nd)$. In either case, these terms are negligible compared to $s_{h,j}^2/r$.

Corollary 1 *Suppose $y_1, \ldots, y_n$, F, θ, σ_y^2, $\hat{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$ satisfy the conditions of Theorem 3 and suppose that the asymptotic correlation between them is not ± 1. Then for any fixed d, leave-d-out CV is not asymptotically risk optimal for choosing between $\hat{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$.*

Proof: For fixed d, $r/n \to 1$, so (9) is $O_P(1/n)$, which is the same size as (10) and (11). All other terms are $o_P(1/n)$.

Remark 3 The differentiability condition of Theorem 3 is weak and covers a wide variety of cases of interest. Similar conditions were discussed in Shao and Wu (1989) in the context of delete-d jackknife estimates of variance. Section 6 of that paper verifies our condition, under regularity conditions on the distribution F, for a number of statistics including quantiles, some L- and M-estimators. In particular, under the conditions $r/n \to 0$ and $r \to \infty$, leave-d-out CV can choose between the mean and the median, unlike leave-one-out CV.

Remark 4 The number of computations in computing a leave-d-out CV estimate increases rapidly with d. It is possible to reduce the amount of computation by means of balanced subsampling or by Monte Carlo simulation as in Shao (1993). In the first case, subsets are chosen so that each observation y_i appears in the same number of subsets and each pair (y_i, y_j) appears in the same number of subsets. It is easy to see that in this case Theorem 3 remains valid. By choosing subsets at random, the same properties are satisfied on average.

Remark 5 The need to use most observations to validate is not inherent to the resampling procedure. As in Theorem 1, the average prediction risk estimator of the form (7) using an independent validation sample of the same size as the original sample would not be risk optimal due to the size of the cross-product term. But as in Remark 1, using a much larger sample size would render the cross-product negligible and lead to an asymptotically risk optimal selection criterion.

Remark 6 Another class of methods for choosing among competing estimators is based on bootstrap estimates of risk. By bootstrapping, one can estimate *estimation* risk directly rather than going through *prediction* risk estimates. For instance, to choose among a fixed (and finite) number of unbiased location estimators, one can compute a bootstrap estimate of variance and select the estimator with the smallest bootstrap estimate of variance. The resulting adaptive estimator would be asymptotically risk optimal provided that the bootstrap estimate of variance for each estimator is consistent, a condition met by most $\sqrt{n}-$convergent location estimators.

For more details, see Léger and Romano (1990a,b). Note that minimizing bootstrap estimates of prediction risk would also be asymptotically risk optimal because bootstrap estimates of prediction risk, being prediction risks for the same estimator but under a different distribution, are the sum of a variance estimate (the same for all location estimators) and the bootstrap estimate of estimation risk discussed above. Hence the cross-product term is 0.

Remark 7 The results in this paper are for families with a finite number of estimators. Usually, the family contains an infinite number of estimators, such as the class of trimmed means. Clearly, our non-optimality results are valid in this case. Asymptotic optimality requires smoothness of the risk (or loss) estimator in the tuning parameter λ. This often depends on the smoothness of the family of estimators $\hat{\theta}_\lambda$. This issue is discussed in more detail in, for example, Léger and Romano (1990a,b).

3 Simulations

To verify the small sample loss and risk behavior of leave-d-out CV selection criteria we performed 2 small simulation studies, one for ordinary location estimation and the other for regression estimation. They confirmed that deleting a large proportion of observations improves the risk behavior of unbiased $\sqrt{n}$-convergent estimators.

For each simulation, we simulated 1,000 samples of size 100, computed the delete-1 CV criterion exactly, and simulated 100 CV samples to approximate the delete-10 and delete-90 CV estimators. For each adaptive estimator, we computed the probability that the selected estimator corresponds to the one with smaller risk, and the relative efficiencies of the adaptive estimator compared to the risk optimal estimator.

The 2 studies were:

1. choosing between the mean and the median for Normal and double exponential data

2. choosing between weighted and unweighted least squares regression when the data are generated by the model $y = 1 + x + s(x)\epsilon$ where ϵ is generated i.i.d. $N(0, .2)$ and $s(x)$ is either the constant 1 or a normalized version of the inverse square root of the weight function. The weights are $w(x) = 1 + 4(.25 - (x - .5)^2)$

The sample mean is the minimum variance unbiased estimator of the mean for the Normal, and the median is optimal for the double exponential. The probabilities of selecting the optimal estimator by CV for the Normal distribution were 0.50, 0.62 and 0.92 for the delete-1, delete-10 and delete-90 criteria, respectively. The corresponding results for the double exponential were 0.71, 0.77, and 0.88 respectively. In this example, deleting a larger proportion of observations improves the probability of choosing the estimator with the smaller risk. For the Normal distribution, the efficiencies of the adaptive estimators (computed as the variance of the better estimator over the variance of the adaptive estimator) were 88%, 92% and 97% for the delete-1, delete-10, and delete-90 CV criteria, respectively. The corresponding results for the double exponential were 74%, 73% and 90%, respectively. Note that the efficiencies of the delete-90 CV are always much better than the other two.

For the linear regression problem, we considered ordinary least squares versus weighted least squares regression. The 100 x's were equally spaced on $[0, 1]$. To avoid variance inflation in regression estimates caused by too small a range of x in the delete-90 samples, 3 points were selected at random on the interval [0,.3], 4 at random on (.3,.7] and 3 at random on (.7,1]. The delete-10 samples used the complement of these selected points.

When the variance of the errors is constant, the relative efficiency of the weighted to unweighted regression estimator is 80%. The ASE of the weighted estimator was smaller than that of the unweighted 38% of the time and the relative risk of using the (ideal but unattainable) estimator with smaller loss was 115%. The probabilities of selecting the estimator with smaller risk were 0.85, 0.81 and 0.95 for delete-1, delete-10, and delete-90 CV criteria respectively, with respective relative efficiencies of 99%, 98% and 99%. Although delete-1 CV has high relative efficiency, it is picking the "wrong" estimator a large proportion of times, which would lead, for example, to incorrect confidence intervals if the intervals are based on the selected model.

When the variance of the errors varies with x, the weighted estimator is optimal. The relative efficiency of the unweighted to the weighted estimator is 71%. The ASE of the unweighted estimator was smaller than that of the weighted 34% of the time and the relative risk of using the estimator with smaller loss was 121%. The probabilities of selecting the estimator with smaller risk were 0.83, 0.80 and 0.93 for delete-1, delete-10, and delete-90 CV criteria

respectively, with respective relative efficiencies of 86%, 86% and 98%.

Remark 8 The claim is often made (e.g. Hart, 1995) that CV is estimating prediction *loss* and that it therefore may do better at estimating the loss optimal estimator for the data set at hand than methods clearly aimed at estimating risk. However, the results of this study strongly suggest that CV is estimating risk. In the regression problem, the risk optimal estimator has the smaller ASE only about 2/3 of the time, but it has smaller CV 80-99% of the time, depending on the number of data points deleted. There was no apparent tendency of CV to select the "wrong" estimator when that estimator had smaller ASE.

4 Conclusion

The increasing availability of fast, convenient desk-top computing, has spurred statisticians to develop data-analytic methods which do not require stringent distributional assumptions. To work well, procedures such as nonparametric smoothing and other "self-modeling" techniques require data-adaptive selection of tuning parameters from a class of available estimators. Cross-validation and adjusted residual sum of squares have been suggested in a number of contexts – examples include nonparametric regression, (Härdle and Marron, 1985; Wahba and Wold, 1975) and variable selection in multiple linear regression, (Allen, 1974) – due to their ease of use and intuitive appeal. Although recent research has shown that the convergence of tuning parameters selected by these methods can be slow compared to competing methods (Härdle, Hall and Marron, 1988; Jones, Marron and Sheather, 1992), they are heavily used in practice.

In this article we have shown that for squared error loss, the cross-product term of prediction risk estimators plays a crucial role in the asymptotic optimality of selection criteria. While its expectation is 0, so that prediction risk and estimation risk are equivalent for the purpose of choosing a tuning parameter, its size may be just as large as the estimate of estimation risk or loss. This has been shown for $\sqrt{n}$-convergent estimators of location and so leave-one-out CV, adjusted residual sum of squares, and even average prediction loss estimates based on an independent validation sample do not lead to asymptotically loss optimal criteria. On the other hand, we have shown that the latter two are asymptotically loss optimal for choosing between two asymptotically Normal estimators which converge at the rate n^p with $p < 1/2$.

For $\sqrt{n}$-convergent estimators which satisfy a weak differentiability condition, risk optimality can be obtained by using a leave-d-out CV where $(n-d)/n \to 0$ with $(n-d) \to \infty$ is required. This latter result is in agreement with Shao (1993) in the context of model selection in multiple linear regression with a fixed number of variables.

It is worth noting that bootstrap estimators of prediction risk do not suffer from this problem, because the bootstrap method computes actual expectations rather average prediction losses and so no cross-product is involved. Moreover, the bootstrap can also be used to estimate estimation risk directly rather than going through prediction risk. This is particularly important if the loss is different from squared error.

References

Akaike, H. (1974). A new look at the statistical model identification. *I.E.E.E. Trans. Auto. Control* **19** 716–723.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16** 1307-1325.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31** 377–403.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83** 86–95.

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.

Hart, J. D. (1995). Smoothing Time-Dependent Data: A Survey of Data-Driven Methods. *J. Nonpar. Stat.* (to appear).

Jones, M. C., Marron, J. S. and Sheather, S. J. (1992). Progress in Data-Based Bandwidth Selection for Kernel Density Estimation.Ann. Statist. Mimeo Series 2088, Dept. of Statistics, University of North Carolina, Chapel Hill.

Léger, C. and Romano, J. P. (1990a). Bootstrap choice of tuning parameters. *Ann. Inst. Statist. Math.* **42** 709–735.

Léger, C. and Romano, J. P. (1990b). Bootstrap adaptive estimation: The trimmed-mean example. *Canad. J. Statist.* **4** 297–314.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15** 958–975.

Pruitt, R. C. (1988). Cross-validation in the one sample location problem. Tech. report No. 510, School of Statistics, Univ. of Minnesota.

Romano, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* **16** 629–647.

Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions. *J. Statist. Plann. Inference* **35** 65–75.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494.

Shao, J. and Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *Ann. Statist.* **17** 1176–1197.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika* **64** 29–35.

Wahba, G. and Wold, S. (1975). A completely automatic French curve: Fitting spline functions by cross validation. *Comm. Statist. A—Theory Methods* **4** 1–17.