Asymptotics of Cross-Validated Risk Estimation in Model Selection and Performance Assessment

Sandrine Dudoit and Mark J. van der Laan
Division of Biostatistics, University of California, Berkeley
sandrine@stat.berkeley.edu

February 4, 2003

Abstract

Risk estimation is an important statistical question for the purposes of selecting a good predictor (i.e., model selection) and assessing its performance (i.e., estimating its generalization error). This article derives distributional properties of cross-validated risk estimators in the context of model selection and performance assessment. General loss functions are considered, including the absolute and squared error loss functions for continuous outcomes and the indicator loss function for polychotomous outcomes. A broad definition of cross-validation is used in order to cover leave-one-out cross-validation, V-fold cross-validation, and Monte Carlo cross-validation. For model selection, the asymptotic optimality of cross-validation procedures is established, in the sense that a selector based on a cross-validated risk estimator performs asymptotically as well as an optimal benchmark selector based on the risk for the true unknown distribution. An important condition for the theorems is that the size of the validation sets converges to infinity; this rules out leave-one-out cross-validation. For predictor performance assessment, cross-validated risk estimators are shown to be consistent and asymptotically linear for the risk for the true underlying distribution.

Keywords: Classification, cross-validation, loss function, model selection, prediction, risk estimation.

1 Introduction

Risk estimation, i.e., the estimation of prediction error as measured by the expected value of a loss function comparing predicted and true responses, is as important statistical problem for at least two purposes. Risk estimation is used for: (i) *model or predictor selection*, where the best predictor is chosen to minimize risk over a given class of predictors; (ii) *predictor performance assessment*, i.e., estimating the *generalization error* of the selected predictor when it is used to predict the response Y corresponding to a future observation with explanatory variables X. These two fundamental problems have been referred to in the statistical literature as "submodel selection and evaluation" (Breiman, 1992) and "choice and assessment of statistical predictions" (Stone, 1974). An immediate difficulty is that the joint distribution P of the responses Y and explanatory variables X is typically unknown. This means that the available data have to be used for both tasks (i) and (ii), that is, to select a good model (specifically, estimate the risk criteria used to select the model) and to assess the performance of this selected model. As discussed by Breiman (1992) in the context of dimensionality selection in regression, criteria such as Mallow's C_p, Akaike information's criterion (AIC), and the Bayesian information criterion (BIC), do not account for the data-driven selection of the sequence of submodels and thus provide biased assessment of prediction error in finite sample situations. Instead, risk estimation methods based on sample reuse have been favored. The main procedures include: leave-one-out cross-validation, V-fold cross-validation (i.e., random partition of the learning set into V mutually exclusive and exhaustive sets), Monte Carlo cross-validation (i.e., repeated random splits of the learning set into a training and a validation set), and the bootstrap (Chapter 3 in Breiman et al. (1984), Breiman and Spector (1992), Breiman (1996a), Breiman (1996b), Chapter 17 in Efron and Tibshirani (1993), Chapters 7 and 8 in Györfi et al. (2002), Chapter 7 in Hastie et al. (2001), Chapter 3 in Ripley (1996), Stone (1974), Stone (1977)).

Thus, a variety of cross-validation procedures are available for estimating the risk of a predictor. A natural question then concerns the distributional properties of the resulting risk estimators, i.e., their performance as estimators of generalization error, their performance in terms of identifying a good predictor (model selection), and also the impact of the particular cross-validation procedure (e.g., the choice of V in V-fold cross-validation, the use of V-fold vs. Monte Carlo cross-validation). Aside from empiri-

cal assessment of different estimation procedures, previous theoretical work has focused primarily on the distributional properties of leave-one-out cross-validation (Stone, 1974, 1977).

The present article examines distributional properties of cross-validated risk estimators in the context of both model selection and predictor performance assessment. A broad definition of cross-validation (CV) is considered in order to cover leave-one-out cross-validation (LOOCV), V-fold cross-validation, and Monte Carlo cross-validation. In this general framework, a learning set is divided into a training set and a validation set based on the value of a random n-vector S_n. For a given S_n, the risk of a predictor built using the training set is assessed by the empirical mean of the loss function on the validation set. These individual risk estimators are then averaged over S_n to yield the cross-validated risk estimator. The particular distribution of S_n determines the flavor of the cross-validation procedure. Note that some bootstrap-based risk estimation methods, such as the .632 bootstrap, can also be handled within this framework (Section 2.3.2).

For model selection, the asymptotic optimality of cross-validation procedures is established (Theorems 1 and 2, and Corollary 1), in the sense that a selector based on a cross-validated risk estimator performs asymptotically as well as an optimal benchmark selector based on the risk for the true unknown distribution P. An important condition for the theorems is that the size of the validation sets converges to infinity; this rules out leave-one-out cross-validation. For predictor performance assessment, cross-validated risk estimators are shown to be consistent and asymptotically linear for the risk for the true underlying distribution (Theorems 3 and 4). The results derived in this article apply to general loss functions (e.g., squared and absolute error loss, indicators) and distributions of S_n.

Note that we are concerned with prediction error in the *X-random* case, i.e., the prediction error is evaluated for a new pair of (X, Y) random variables. This is in contrast to the *X-fixed* case, where prediction error is evaluated for the same set of X variables as in the available dataset. Breiman (1992) examines the X-fixed case in detail and proposes a new procedure, the *little bootstrap*, for estimating X-fixed prediction error in the context of dimensionality selection is regression.

The article is organized as follows. Section 2 reviews basic notions in prediction and risk estimation, and introduces a general framework for cross-validated risk estimation. Section 3 derives distributional properties of cross-validated risk estimators in model selection, while Section 4 concerns distri-

butional properties of cross-validated risk estimators of generalization error. Finally, Section 5 summarizes our findings and discusses related work and open questions.

2 Cross-validated risk estimation

2.1 Prediction

Consider random variables X and Y with joint distribution P and ranges $\mathcal{X}$ and $\mathcal{Y}$, respectively. We are concerned with using the explanatory variables X to predict the responses Y. A *predictor* is a mapping C from $\mathcal{X}$ into $\mathcal{Y}$, $C: \mathcal{X} \to \mathcal{Y}$, where $\hat{y} = C(x)$ denotes the prediction corresponding to explanatory variable x.

Classification is a prediction, or learning problem, in which the variable Y to be predicted assumes one of K predefined and unordered values, $\{c_1, c_2, \ldots, c_K\}$, arbitrarily relabeled by the integers $\{1, 2, \ldots, K\}$ or $\{0, 1, \ldots, K-1\}$, and sometimes $\{-1, 1\}$ in binary classification. A *classifier* C corresponds to a *partition* of the space $\mathcal{X}$ into K disjoint and exhaustive subsets, $A_1, \ldots, A_K$, such that an observation with explanatory variable $x \in A_k$ has predicted class $\hat{y} = k$.

Predictors are built, or *trained*, from past experience, i.e., from a *learning set* (LS) $\mathcal{L} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of observations with both explanatory variables X and responses Y. The (X_i, Y_i) are typically assumed to be independent and identically distributed (i.i.d.) random variables with distribution P. Denote the empirical distribution by P_n and a classifier built from this empirical distribution by $C(\cdot \mid P_n)$.

Standard classification methods include linear discriminant analysis, k-nearest neighbor classifiers, classification trees, and support vector machines. Predictors for continuous outcomes include linear and non-linear regression, and regression trees. An introduction to prediction methods can be found in the texts by Breiman et al. (1984), Hastie et al. (2001), McLachlan (1992), and Ripley (1996).

2.2 Loss and risk functions

To quantify errors in prediction, we introduce a *loss function* $L: \mathcal{Y}^2 \to \Re^+$, where $L(y, \hat{y})$ elaborates the loss incurred when predicting y by $\hat{y}$. Common

choices of loss function for continuous responses are the absolute and squared error losses, also known as L_1 and L_2 losses, respectively

$$\begin{aligned} L(y, \hat{y}) &= |\hat{y} - y| \\ L(y, \hat{y}) &= (\hat{y} - y)^2. \end{aligned}$$

In classification, one often uses the indicator loss function

$$L(y, \hat{y}) = I(y \neq \hat{y}),$$

i.e., making an error of one type is equivalent to making an error of a different type. Another type of loss function is $L(h, l) = L_h I(h \neq l)$, that is, the loss incurred from misclassifying a class h observation is the same irrespective of the predicted class l.

The *risk function* for a predictor C is the expected loss, that is,

$$R(C, P) = E_P[L(Y, C(X))] = \int L(y, C(x)) dP(x, y).$$

In classification, for a loss function $L(y, \hat{y}) = I(y \neq \hat{y})$, the risk is simply the *misclassification rate*, $Pr(C(X) \neq Y)$.

When (unrealistically) P is known, it is possible to define an optimal predictor, C_{opt}, which minimizes the risk function: $C_{opt} = \text{argmin}_C R(C, P)$. This situation gives an upper bound on the performance of predictors in the more realistic setting where the distribution P is unknown. For the squared and absolute error losses, the optimal predictors are, respectively, the conditional mean and median of Y given X, i.e., $C_{opt}(X) = E_P(Y|X)$ and $C_{opt}(X) = \text{median}_P(Y|X)$. In classification, for the simple indicator loss function, the optimal predictor is given by the *Bayes rule*, i.e., the predicted class is that with maximum posterior probability given X, $C_{opt}(X) = \text{argmax}_y Pr(y|X)$. For the general loss function in classification

$$C_{opt}(X) = \text{argmin}_y \sum_{y'=1}^{K} L(y', y) Pr(y'|X).$$

In practice, however, predictors are trained using a learning set $\mathcal{L}$ and it is useful to distinguish between the risk conditional on the learning set $\mathcal{L}$, i.e., conditional on the empirical distribution P_n, and the marginal risk.

The *conditional risk* (conditional on P_n) is defined as

$$\tilde{\theta}_n = \int L(y, C(x \mid P_n)) dP(x, y) \tag{1}$$

and the *marginal risk* as

$$\theta_n = E_P[\tilde{\theta}_n] = E_P\left[\int L(y, C(x \mid P_n))dP(x,y)\right]. \tag{2}$$

We also define the *asymptotic risk*, or risk of the predictor based on the distribution P, as

$$\theta = \int L(y, C(x \mid P))dP(x,y). \tag{3}$$

The (marginal) *optimal risk* is

$$\theta_{opt} = R(C_{opt}, P) = \min_C \int L(y, C(x))dP(x,y). \tag{4}$$

Note that, unlike θ_{opt}, the quantities θ, θ_n, and $\tilde{\theta}_n$ are specific to the predictor C under consideration. The asymptotic risk θ and the optimal risk θ_{opt} will coincide if the predictor $C(X \mid P_n)$ is consistent for $C_{opt}(X)$, i.e., if $C(X \mid P) = C_{opt}(X)$ a.e. The optimal risk θ_{opt} will be the quantity of interest in model selection (Section 3), while the asymptotic risk θ will be of interest for assessing predictor performance (Section 4).

In practice, one does not have access to P and therefore the above four quantities, which are functions of P, are unknown. Note that the optimal risk, θ_{opt}, the marginal risk, θ_n, and the asymptotic risk, θ, are unknown *parameters*, while the conditional risk, $\tilde{\theta}_n$, is an unknown *random variable*, as it depends on the empirical distribution P_n in addition to P. Table 1 p. 12 summarizes the various definitions of risk.

2.3 Risk estimation

One may be interested in evaluating the risk of a predictor for at least two purposes: (i) *model or predictor selection*, where the *best* predictor is chosen to minimize risk over a given class of predictors; (ii) *predictor performance assessment*, i.e., estimating the *generalization error* of the selected predictor when it is used to predict the response Y corresponding to a future observation with explanatory variables X (sampled from P independently of the empirical distribution P_n used to build the predictor). It is thus important in practice to derive accurate estimators of the risks defined in equations (1)–(4).

Before presenting various risk estimation procedures, it will be useful to introduce the following terminology. In many situations, the sample of

(X_i, Y_i) available to the investigator has to be used for both tasks above: (i) building the predictor, including the predictor selection, or training, aspect, and (ii) assessing its performance. It is then common practice to divide the entire sample into two sets, a learning set and a test set. The *learning set* is used to select and build the predictor (task (i)) and the *test set* is used to estimate its generalization error (task (ii)), i.e., to assess the overall performance of the predictor selected in (i). For the purpose of selecting a good predictor in (i), the learning set can be further divided into two sets, a *training set*, on which the predictors are built, and a *validation set*, to which the predictors are applied and based on which their risks are estimated. The predictor with the smallest risk estimated from the validation sets is retained. The different sets are represented in Figure 1.

The risk estimation procedures described in this article are equally applicable for both tasks (i) and (ii), that is, the same procedure can be applied but to different empirical distributions. For (i), the empirical distribution will be that of the learning set, while for (ii), the empirical distribution will be that of the entire dataset. When one is concerned with both tasks (i) and (ii), a *double* or *nested cross-validation* study can be performed: for each random division of the dataset into a learning set and a test set, the learning set is in turn randomly divided into a training set and a validation set.

Below, we use P_n to refer to a learning set of n observations, to be divided into a training set and a validation set (as in (i), for model selection). The same results hold when P_n refers to an entire dataset to be divided into a learning set and test set (as in (ii), for performance assessment)

2.3.1 Resubstitution risk estimation

A naive risk estimator is the *resubstitution error*, where the same dataset is used to build the predictor and to assess its performance

$$\hat{\theta}_n^{LS} = \int L(y, C(x \mid P_n)) dP_n(x, y). \tag{5}$$

That is, the predictor is trained using the entire learning set $\mathcal{L}$, and an estimate of the prediction error is obtained by running the *same* learning set $\mathcal{L}$ through the predictor and comparing predicted and actual responses. Although this is a simple approach, the resubstitution risk estimator $\hat{\theta}_n^{LS}$ can be severely biased downward as an estimator of the conditional risk $\tilde{\theta}_n$. Consider the trivial and extreme classification example where the classifier partitions

the space $\mathcal{X}$ into n sets, each containing a learning set observation. In this extreme over-fitting situation, the resubstitution error rate is zero. However, such a classifier is unlikely to generalize well, that is, the classification error rate as estimated from an independent test set is likely to be high.

2.3.2 Cross-validated risk estimation

In this article, we are concerned with distributional properties of cross-validated (CV) risk estimators. Such estimators involve random divisions of the learning set into two sets: a *training set*, used to build the predictor, and a *validation set*, used to estimate the risk. To derive a general representation for CV risk estimators, we introduce a binary random n-vector $S_n \in \{0,1\}^n$, independent of the empirical distribution P_n. A realization of $S_n = (S_{n,1}, \ldots, S_{n,n})$ defines a particular split of the sample of n observations into a training set $\{i \in \{1,\ldots,n\} : S_{n,i} = 0\}$ and validation set $\{i \in \{1,\ldots,n\} : S_{n,i} = 1\}$. Let P^0_{n,S_n} and P^1_{n,S_n} denote the empirical distributions of the training and validation sets, respectively. The particular distribution of S_n defines the type of cross-validation procedure (see examples below). The general form of the *cross-validated risk estimator* is

$$\hat{\theta}_{n(1-p)} = E_{S_n} \int L(y, C(x \mid P^0_{n,S_n}))dP^1_{n,S_n}(x,y). \tag{6}$$

The proportion of observations in the validation sets, $p = \sum_i S_{n,i}/n$, is typically a pre-specified parameter of the CV procedure, with $p \in (0,1)$. When needed, we will use the notation p_n to emphasize the dependence of p on the sample size n.

Monte Carlo cross-validation. In *Monte Carlo cross-validation*, the learning set is repeatedly and randomly divided into two sets, a training set of $n_0 = n(1-p)$ observations and a validation set of $n_1 = np$ observations. A common choice for p in the machine learning literature is 10% (Breiman, 1998). The corresponding distribution of S_n places mass $1/\binom{n}{np}$ on each binary vector $s_n = (s_{n,1}, \ldots, s_{n,n})$ such that $\sum_i s_{n,i} = np$. In practice, the support of the distribution of S_n can be very large and one approximates the expected value over S_n by an empirical average based on a random sample of S_n's.

V-fold cross-validation. In *V-fold cross-validation*, the learning set $\mathcal{L}$ is randomly divided into V mutually exclusive and exhaustive sets, $\mathcal{L}_v$, $v = 1, \ldots, V$, of as nearly equal size as possible. Predictors are built on training sets $\mathcal{L} - \mathcal{L}_v$, error rates are computed for the validation sets $\mathcal{L}_v$, and averaged over v.

V-fold CV amounts to using a random vector S_n with a distribution that places mass $1/V$ on each of the V binary vectors s_n^v, $v = 1, \ldots, V$, defined as follows. Let $n_V = \lfloor n/V \rfloor$ denote the integer part, or floor, of n/V. Then, for $v = 1, \ldots, V-1$, let $s_{n,i}^v = 1$ for $i = 1 + (v-1)n_V, \ldots, vn_V$ and 0 otherwise. For $v = V$, let $s_{n,i}^V = 1$ for $i = 1 + (V-1)n_V, \ldots, n$ and 0 otherwise. The proportion p of observations in the validation sets is approximately $1/V$.

Leave-one-out cross-validation. A commonly used form of cross-validation is *leave-one-out cross-validation* (LOOCV), where $V = n$ and $p_n = 1/n$. In LOOCV, each observation in the learning set is used in turn as the validation set and the remaining $n-1$ observations are used as the training set. The corresponding distribution of S_n places mass $1/n$ on each binary vector $s_n = (s_{n,1}, \ldots, s_{n,n})$ such that $\sum_i s_{n,i} = 1$. Note that, for model selection, the asymptotic optimality results in Section 3 below for CV risk estimators require that $np_n \to \infty$; this is not the case for LOOCV.

Intuitively, there is a *bias-variance trade-off* in the selection of p. Large p's typically produce estimators of the conditional risk $\tilde{\theta}_n$ with a large bias, but a small variance. In particular, LOOCV, with $p = 1/n$, often results in low bias but high variance estimators. The simulation studies in Breiman and Spector (1992) and Breiman (1996a) for model selection show that leave-one-out cross-validation is inferior to leave-many-out cross-validation (e.g., $V = 10$-fold CV). In particular, LOOCV is found to behave poorly in selection from an unstable sequence of predictors.

Bootstrap cross-validation. A number of cross-validation procedures based on bootstrap samples have been proposed for estimating prediction error (Ambroise and McLachlan, 2002; Efron and Tibshirani, 1993). The standard *leave-one-out bootstrap* procedure, $B1$, can be viewed as producing training sets that are random samples of size n drawn *with replacement* from the learning set. For each bootstrap sample, about one-third ($(1 - 1/n)^n \approx e^{-1} \approx .368$) of the cases are left out; these observations form the validation set. The definition of the random vector S_n needs to be modi-

fied for bootstrap-based CV to account for multiple occurrences of the same observation in the training sets. This can be done by allowing weights in the empirical distribution P^0_{n,S_n}. In this setting, one could define $s_{n,i}$ as the number of occurrences of observation (X_i, Y_i) in the training set, so that $S_n \in \{0, \ldots, n\}^n$ and there are n^n possible random vectors S_n. In practice, one approximates the expected value over S_n by an empirical average based on a random sample of S_n's. For bootstrap-based CV, the proportion of observations in the validation sets, $p_n = \sum_i I(S_{n,i} = 0)/n$, is a random variable, with $E[p_n] = (1 - 1/n)^n \approx e^{-1} \approx .368$ (note that $S_{n,i} = 0$ now correspond to validation set observations). Shortcomings of this approach include the occurrence of ties in the training sets and the lack of control over p_n. In addition, the resulting risk estimator, $\hat{\theta}^{B_1}_n$, tends to be upwardly biased (it is based roughly on only two-thirds of the data) and estimators which combine it with the downwardly biased resubstitution estimator have been suggested. The *.632 bootstrap estimator* is a linear combination of these two estimators: $\hat{\theta}^{.632}_n = .368 \times \hat{\theta}^{LS}_n + .632 \times \hat{\theta}^{B_1}_n$, where the factor .632 corresponds to the expected proportion of learning set observations included in the bootstrap training samples (Chapter 17 in Efron and Tibshirani (1993)).

2.4 Risk estimation in model selection

An important application of risk estimation is to predictor selection or training, a form of model selection. Consider a family $\{C_k\}$ of predictors, indexed by a parameter k, $k = 1, \ldots, K(n)$. For example, k might represent the number of neighbors in k-nearest neighbor classification, the kernel used in support vector machines, or the number of explanatory variables used in linear regression. Model selection generally involves a trade-off between bias and variance: variance in the predictions typically increases with model complexity (e.g., as measured by the number of variables in a regression setting), while bias tends to decrease. An optimal predictor should achieve a balance between bias and variance. The issue of bias-variance trade-off is discussed in Breiman (1992) in the context of dimensionality selection in regression.

A fundamental and practical problem is the selection of a $\hat{k}$ in such a manner that the risk of the predictor $C_{\hat{k}}(\cdot \mid P_n)$ converges optimally to that of the optimal predictor C_{opt}. Ideally, given the empirical distribution P_n, one seeks k that minimizes the conditional risk

$$\tilde{\theta}_n(k) = \int L(y, C_k(x \mid P_n)) dP(x, y). \tag{7}$$

However, P is unknown. One could envisage using the empirical distribution, P_n, in place of the true P as in the resubstitution estimator $\hat{\theta}_n^{LS}$, but this generally leads to over-fitting. Instead, we turn to cross-validated risk estimation as described above. In this setting, the learning set $\mathcal{L}$ is split at random into two sets, a training set and a validation set. A predictor C_k is built for each $k \in \{1, \ldots, K(n)\}$ using the training set only and the empirical distribution for the validation set is used in place of the true P in equation (7) to estimate the conditional risk $\tilde{\theta}_n(k)$. Specifically, we define the *cross-validated risk criterion* as

$$\hat{\theta}_{n(1-p)}(k) = E_{S_n} \int L(y, C_k(x \mid P^0_{n,S_n}))dP^1_{n,S_n}(x, y) \tag{8}$$

and the corresponding optimal choice $\hat{k}$, or *cross-validated selector*,

$$\hat{k} = \text{argmin}_{k \in \{1,\ldots,K(n)\}} \hat{\theta}_{n(1-p)}(k).$$

To obtain a benchmark for the selected $\hat{k}$, we also define the conditional risk based on $n(1-p)$ training observations

$$\tilde{\theta}_{n(1-p)}(k) = E_{S_n} \int L(y, C_k(x \mid P^0_{n,S_n}))dP(x, y) \tag{9}$$

and its minimizer

$$\tilde{k} = \text{argmin}_{k \in \{1,\ldots,K(n)\}} \tilde{\theta}_{n(1-p)}(k).$$

When needed (e.g., Corollary 1), we will also refer to $\tilde{k}$ as $\tilde{k}_{n(1-p)}$, to distinguish it from $\tilde{k}_n$, the minimizer of the conditional risk $\tilde{\theta}_n(k)$ for the entire P_n.

In his article on model selection, Breiman (1996a) defines the *predictive loss* (PL) as the difference between the risks for a *fallible* and a *crystal ball* estimator of k. In our case, the predictive loss would compare the conditional risks for the fallible CV selector $\hat{k}$ and the crystal ball benchmark selector $\tilde{k}$, that is, $\tilde{PL}_{n(1-p)} = \tilde{\theta}_{n(1-p)}(\hat{k}) - \tilde{\theta}_{n(1-p)}(\tilde{k})$. The results in Section 3 below show that the cross-validated selector $\hat{k}$ is asymptotically optimal, in the sense that the ratio of expected risk differences $(E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt})/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to one, and thereby $E\tilde{PL}_{n(1-p)}/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to zero (Theorems 1 and 2). Under additional conditions, one can show that the same results hold for the conditional risk $\tilde{\theta}_n(k)$ using the entire empirical distribution P_n (Corollary 1).

Having selected a $\hat{k}$ using cross-validation applied to the learning set and built a predictor $C_{\hat{k}}(\cdot|P_n)$, the next task is to assess the generalization error of this final predictor, i.e., estimate $\tilde{\theta}_n(\hat{k})$. This can be done again by (double) cross-validation, where the learning set is obtained by random divisions of the entire dataset into a learning set and a test set. As a practical issue, when an empirical mean is used in place of the expected value over S_n in equation (8), the same splits (i.e., realizations of S_n) should be used for each value of k.

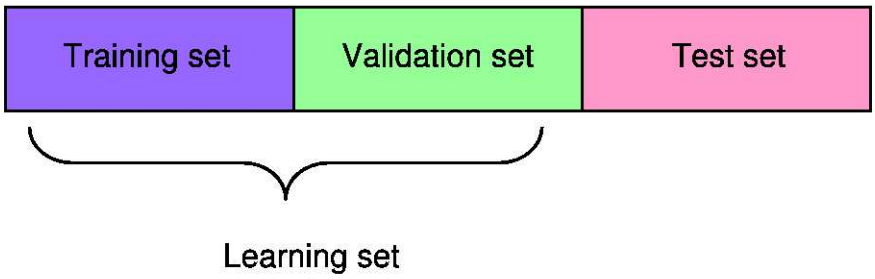

Figure 1: Terminology for the datasets used in cross-validation.

Table 1: *Definitions of risk and risk estimators.*

Name	Symbol	Definition
Optimal risk	θ_{opt}	$R(C_{opt}, P) = \min_C \int L(y, C(x))dP(x, y)$
Asymptotic risk	θ	$\int L(y, C(x \mid P))dP(x, y)$
Marginal risk	θ_n	$E_P \int L(y, C(x \mid P_n))dP(x, y)$
Conditional risk	$\tilde{\theta}_n$	$\int L(y, C(x \mid P_n))dP(x, y)$
Conditional risk	$\tilde{\theta}_{n(1-p)}$	$E_{S_n} \int L(y, C(x \mid P^0_{n,S_n}))dP(x, y)$
CV estimator	$\hat{\theta}_{n(1-p)}$	$E_{S_n} \int L(y, C(x \mid P^0_{n,S_n}))dP^1_{n,S_n}(x, y)$
Resubstitution estimator	$\hat{\theta}^{LS}_n$	$\int L(y, C(x \mid P_n))dP_n(x, y)$

3 Results for model selection

We derive two main results concerning the asymptotic optimality of the cross-validated model selectors described in Section 2.4. In this context, the centered conditional risk for the cross-validated selector $(\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt})$ is compared to the centered conditional risk for the benchmark selector $(\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$. Finite sample bounds are obtained for the expected value of the predictive loss, $E\tilde{PL}_{n(1-p)} = E(\tilde{\theta}_{n(1-p)}(\hat{k}) - \tilde{\theta}_{n(1-p)}(\tilde{k}))$, that imply convergence to zero in expectation and in probability of this risk difference, at rate $O(\log(K(n))/\sqrt{np})$ for general loss functions (Theorem 1) and $O(\log(K(n))/np)$ for the squared error loss (Theorem 2). Consequently, if the risk difference $E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}$ converges to zero slower than these rates, then the ratio of expected risk differences $(E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt})/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to one. This implies, in particular, that $E\tilde{PL}_{n(1-p)}/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to zero. Convergence in probability of the ratios of risk differences follows from Lemma 2 below. Theorem 1 applies to general loss functions, including the squared and absolute error loss functions commonly used for continuous outcomes, and general loss functions for polychotomous outcomes. In the special case of the squared error, or L_2, loss function, Theorem 2 provides a stronger convergence result than Theorem 1: for the L_2 loss function, the rate of convergence is shown to be $O(\log(K(n))/np)$ rather than the slower $O(\log(K(n))/\sqrt{np})$ applicable to general loss functions. Both theorems consider general distributions of S_n, i.e., general cross-validation procedures with an arbitrary proportion p_n of observations included the validation sets. Note that the finite sample results hold for any p_n, while the asymptotic results require that $np_n \to \infty$; the later condition rules out LOOCV.

The proofs of Theorems 1, 2, and 3 rely on Bernstein's inequality, which we state here as a lemma for ease of reference. A proof is given in Györfi et al. (2002) for Lemma A.2, p. 564.

Lemma 1 Bernstein's inequality. *Let Z_i, $i = 1, \ldots, n$, be independent real valued random variables such that $Z_i \in [a, b]$ with probability one. Let $0 < \sum_{i=1}^{n} VAR(Z_i)/n \leq \sigma^2$. Then, for all $\epsilon > 0$,*

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - EZ_i) > \epsilon\right) \leq \exp\left(-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + \epsilon(b-a)/3}\right).$$

This implies

$$Pr\left(\frac{1}{n} \mid \sum_{i=1}^{n}(Z_i - EZ_i) \mid > \epsilon\right) \leq 2\exp\left(-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + \epsilon(b-a)/3}\right).$$

Theorems 1 and 2 first establish convergence in expectation of the conditional risk for the cross-validated selector to that of the benchmark selector. Convergence in probability follows from the next lemma based on Markov's inequality.

Lemma 2 *Consider a sequence of random variables $Z_1, Z_2, \ldots$, with finite expectation $E|Z_n| = O(g(n))$, for a positive function $g(n)$. Then $Z_n = O_P(g(n))$.*

Proof. We wish to show that $\forall\, \epsilon > 0$, $\exists N$ and $B > 0$ such that $Pr(|Z_n|/g(n) > B) < \epsilon\ \forall\ n \geq N$. The proof is a direct application of Markov's inequality. Since $E|Z_n| = O(g(n))$, then $\exists N$ and $C > 0$ such that $E|Z_n|/g(n) < C$ $\forall\ n \geq N$. Thus, letting $B = C/\epsilon$,

$$Pr\left(\frac{|Z_n|}{g(n)} > B\right) \leq \frac{E|Z_n|}{Bg(n)} \leq \frac{C}{B} = \epsilon \qquad \forall\ n \geq N.$$

□

3.1 General loss function

Theorem 1 *Suppose that $\sup_{X,Y} L(Y, C_k(X \mid P_n)) \leq M < \infty$ a.s. for all k, where the supremum is over a support of the distribution of X and Y. Let $m = 2M$ and $v = M^2$, and define*

$$f(M, K(n), np) \equiv 2\left[u_n + \int_{u_n}^{\infty} K(n)\exp\left(-\frac{1}{2}\frac{x^2(np)}{v + mx/3}\right)dx\right], \tag{10}$$

where

$$u_n \equiv \frac{\log(K(n))m/3 + \sqrt{\log^2(K(n))m^2/3^2 + 2np\log(K(n))v}}{np}.$$

We have the following finite sample result

$$0 \leq E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \leq E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt} + f(M, K(n), np). \tag{11}$$

Suppose that $\log(K(n))/\sqrt{np} \to 0$, *as* $n \to \infty$, *then* $f(M, K(n), np) = O(\log(K(n))/\sqrt{np})$. *This implies*

$$E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} = E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt} + O\left(\frac{\log(K(n))}{\sqrt{np}}\right),$$

and, in particular,

$$\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} = \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt} + O_P\left(\frac{\log(K(n))}{\sqrt{np}}\right).$$

If $\frac{\log(K(n))}{\sqrt{np}(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})} \to 0$ *for* $n \to \infty$, *then*

$$\frac{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \to 1.$$

Similarly, if $\frac{\log(K(n))}{\sqrt{np}(\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})} \xrightarrow{P} 0$ *for* $n \to \infty$, *then*

$$\frac{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \xrightarrow{P} 1.$$

Proof. We have

$$\begin{aligned}
0 \;\leq\; & \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \\
= \;& E_{S_n} \int L(y, C_{\hat{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP(x, y) \\
& - E_{S_n} \int L(y, C_{\hat{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP^1_{n,S_n}(x, y) \\
& + E_{S_n} \int L(y, C_{\hat{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP^1_{n,S_n}(x, y) \\
\leq \;& E_{S_n} \int L(y, C_{\hat{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) d(P - P^1_{n,S_n})(x, y) \\
& + E_{S_n} \int L(y, C_{\tilde{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) d(P^1_{n,S_n} - P)(x, y) \\
& + E_{S_n} \int L(y, C_{\tilde{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP(x, y),
\end{aligned}$$

where the first inequality follows by definition of θ_{opt} and the second by definition of $\hat{k}$, such that $\hat{\theta}_{n(1-p)}(\hat{k}) \leq \hat{\theta}_{n(1-p)}(k)\ \forall\ k$. Denote the first two terms by

$T_{n,\hat{k}}$ and $-T_{n,\tilde{k}}$, respectively; the last term is the benchmark $\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}$. Hence

$$0 \leq \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \leq \tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt} + T_{n,\hat{k}} - T_{n,\tilde{k}}. \tag{12}$$

In the sequel, we show that both $|ET_{n,\hat{k}}|$ and $|ET_{n,\tilde{k}}|$ are bounded by $f(M, K(n), np)/2$, so that $ET_{n,\hat{k}} - ET_{n,\tilde{k}} \leq f(M, K(n), np)$. Represent $T_{n,k}$ as $T_{n,k} = E_{S_n} T_{n,k}(S_n)$. For convenience, introduce the following notation for the relevant random variables

$$\begin{aligned} \tilde{H}_k &\equiv \int L(y, C_k(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP(x,y) \\ \bar{H}_k &\equiv \int L(y, C_k(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP^1_{n,S_n}(x,y) \end{aligned}$$

and rewrite $T_{n,k}(S_n)$ as

$$T_{n,k}(S_n) = \tilde{H}_k - \bar{H}_k.$$

Note that

$$\begin{aligned} Pr(T_{n,\hat{k}}(S_n) > s \mid P^0_{n,S_n}, S_n) &= Pr\left(\tilde{H}_{\hat{k}} - \bar{H}_{\hat{k}} > s \mid P^0_{n,S_n}, S_n\right) \\ &\leq K(n) \max_{k \in \{1,\ldots,K(n)\}} Pr\left(\tilde{H}_k - \bar{H}_k > s \mid P^0_{n,S_n}, S_n\right). \end{aligned}$$

The same bound applies to $T_{n,\tilde{k}}(S_n)$. Conditional on P^0_{n,S_n} and S_n, consider the random variable

$$Z_k = L(Y, C_k(X \mid P^0_{n,S_n})) - L(Y, C_{opt}(X)).$$

Let $Z_{k,i}$, $i = 1, \ldots, np$, be the np i.i.d. copies of Z_k corresponding with X_i, given $S_{n,i} = 1$. Note that $\bar{H}_k = 1/np \sum_{i=1}^{np} Z_{k,i}$ and $\tilde{H}_k = E(Z_k \mid P^0_{n,S_n}, S_n)$, so that $\tilde{H}_k - \bar{H}_k = E(Z_k \mid P^0_{n,S_n}, S_n) - 1/np \sum_{i=1}^{np} Z_{k,i}$ represents an empirical mean of np centered i.i.d. random variables. The random variables Z_k are bounded, with $\mid Z_k \mid \leq M$ a.s. and $\mathrm{VAR}(Z_k \mid P^0_{n,S_n}, S_n) \leq v = M^2$. Thus, from Bernstein's inequality (Lemma 1), for $s > 0$,

$$Pr\left(\tilde{H}_k - \bar{H}_k > s \mid P^0_{n,S_n}, S_n\right) \leq \exp\left(-\frac{1}{2}\frac{s^2(np)}{v + ms/3}\right),$$

where $m = 2M$. This proves that, for $s > 0$,

$$Pr(T_{n,\hat{k}}(S_n) > s \mid P^0_{n,S_n}, S_n) \leq K(n) \exp\left(-\frac{1}{2}\frac{s^2(np)}{v + ms/3}\right).$$

In particular, this provides us with the same bound for the marginal probabilities

$$Pr(T_{n,\hat{k}}(S_n) > s) \leq K(n) \exp\left(-\frac{1}{2}\frac{s^2(np)}{v + ms/3}\right).$$

Note that, for any random variable Z,

$$EZ \leq EI(Z > 0)Z = \int_0^\infty Pr(Z > z)dz.$$

Thus, for each $u > 0$, we have

$$\begin{aligned} ET_{n,\hat{k}} &= EE_{S_n}T_{n,\hat{k}}(S_n) \leq \int_0^\infty Pr(T_{n,\hat{k}}(S_n) > x)dx \\ &\leq u + \int_u^\infty K(n) \exp\left(-\frac{1}{2}\frac{x^2(np)}{v + mx/3}\right)dx. \end{aligned}$$

The solution u_n in the statement of the theorem corresponds with the minimizer of the integral on the right-hand side as a function of u, which is also a solution of the equation obtained by setting the derivative with respect to u equal to zero. The same bound can be derived for $-ET_{n,\hat{k}}$ by applying Bernstein's inequality to $\bar{H}_k - \tilde{H}_k$. Hence, $|ET_{n,\hat{k}}| \leq f(M, K(n), np)/2$. A similar proof as above shows that $|ET_{n,\tilde{k}}| \leq f(M, K(n), np)/2$. Thus, taking the expected values of the quantities in equation (12), we have the finite sample result

$$0 \leq E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \leq E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt} + f(M, K(n), np),$$

where

$$f(M, K(n), np) = 2\left[u_n + \int_{u_n}^\infty K(n) \exp\left(-\frac{1}{2}\frac{x^2(np)}{v + mx/3}\right)dx\right].$$

The remaining statements of the theorem involve proving that when $\log(K(n))/\sqrt{np} \to 0$ as $n \to \infty$, then $f(M, K(n), np) = O(\log(K(n))/\sqrt{np})$. First, note that $u_n = O(\sqrt{\log(K(n))/np})$. Next, using the substitution $x = [\log(K(n))/\sqrt{np}]y$, the integral in $f(M, K(n), np)$ can be rewritten as

$$\begin{aligned} &\int_{u_n}^\infty K(n) \exp\left(-\frac{1}{2}\frac{x^2(np)}{v + mx/3}\right)dx \\ &\quad = \frac{\log(K(n))}{\sqrt{np}} \int_{\frac{\sqrt{np}}{\log(K(n))}u_n}^\infty K(n) \exp\left(-\frac{1}{2}\frac{y^2 \log^2(K(n))}{v + \frac{m\log(K(n))}{3\sqrt{np}}y}\right)dy. \end{aligned}$$

Note that the integrand in the expression for $f(M, K(n), np)$ in equation (10) is a decreasing function of x for $x > 0$, which achieves a value of 1 at $x = u_n$ (by definition of u_n) and tends to 0 as x approaches ∞. Hence, for $y > [\sqrt{np}/\log(K(n))]u_n$, the integrand in the last expression is bounded by 1. Since $u_n = O(\sqrt{\log(K(n))/np})$, then $\exists N_1 > 0$ and some constant $1 < A < \infty$, such that $[\sqrt{np}/\log(K(n))]u_n \leq A \ \forall \ n \geq N_1$. Thus, for $n \geq N_1$,

$$\begin{aligned}
&\int_{u_n}^{\infty} K(n) \exp\left(-\frac{1}{2}\frac{x^2(np)}{v + mx/3}\right) dx \\
&\quad = \frac{\log(K(n))}{\sqrt{np}} \int_{\frac{\sqrt{np}}{\log(K(n))}u_n}^{A} K(n) \exp\left(-\frac{1}{2}\frac{y^2 \log^2(K(n))}{v + \frac{m\log(K(n))}{3\sqrt{np}}y}\right) dy \\
&\qquad + \frac{\log(K(n))}{\sqrt{np}} \int_{A}^{\infty} K(n) \exp\left(-\frac{1}{2}\frac{y^2 \log^2(K(n))}{v + \frac{m\log(K(n))}{3\sqrt{np}}y}\right) dy \\
&\quad \leq \frac{\log(K(n))}{\sqrt{np}} \left[A + \int_{A}^{\infty} K(n) \exp\left(-\frac{1}{2}\frac{y^2 \log^2(K(n))}{v + \frac{m\log(K(n))}{3\sqrt{np}}y}\right) dy\right].
\end{aligned}$$

Consider now the second term in the above expression. Since $\log(K(n))/\sqrt{np} \to 0$ as $n \to \infty$, then $\exists N_2 > 0$ such that $m\log(K(n))/3\sqrt{np} < \epsilon \ \forall \ n \geq N_2$. Hence, for $n \geq N_2$ and $1 < A < y$,

$$y^2/(v + m\log(K(n))/3\sqrt{np}y) > y^2/(v + \epsilon y) > y/(v + \epsilon).$$

Let $g(y) = y/(v + \epsilon)$. Then, for $n \geq \max(N_1, N_2)$,

$$\begin{aligned}
&\int_{u_n}^{\infty} K(n) \exp\left(-\frac{1}{2}\frac{x^2(np)}{v + mx/3}\right) dx \\
&\quad \leq \frac{\log(K(n))}{\sqrt{np}} \left[A + \int_{A}^{\infty} K(n) \exp\left(-\frac{1}{2}\log^2(K(n))g(y)\right) dy\right].
\end{aligned}$$

The above expression will be $O(\log(K(n))/\sqrt{np})$ as desired if the integral is uniformly bounded in n. The integrand may be rewritten as $K(n)^{1-\frac{1}{2}g(y)\log(K(n))}$ and is decreasing in $K(n)$ for each y. To see this, let $K(n) > K(n_0)$ and note that $0 < g(A) < g(y)$ for $y > A$. Then,

$$\frac{K(n_0)^{1-\frac{1}{2}g(y)\log(K(n_0))}}{K(n)^{1-\frac{1}{2}g(y)\log(K(n))}} \geq \left(\frac{K(n_0)}{K(n)}\right)^{1-\frac{1}{2}g(y)\log(K(n))}$$

$$\geq \left(\frac{K(n)}{K(n_0)}\right)^{\frac{1}{2}g(A)\log(K(n))-1}$$

which is greater than 1 for each y when A is chosen so that $\frac{1}{2}g(A)\log(K(n)) > 1$. It suffices to show that the integral is finite for some $K(n)$, which is immediate from

$$\begin{aligned}
&\int_A^\infty K(n)\exp\left(-\frac{1}{2}\log^2(K(n))g(y)\right)dy \\
&= \int_A^\infty K(n)\exp\left(-\frac{1}{2}\frac{\log^2(K(n))}{v+\epsilon}y\right)dy \\
&= \frac{2(v+\epsilon)K(n)}{\log^2(K(n))}\exp\left(-\frac{1}{2}\frac{\log^2(K(n))}{v+\epsilon}A\right) < \infty.
\end{aligned}$$

Thus, $f(M, K(n), np) = O(\log(K(n))/\sqrt{np})$. By definition of $\tilde{k}$, $\tilde{\theta}_{n(1-p)}(\tilde{k}) \leq \tilde{\theta}_{n(1-p)}(\hat{k})$ and from the finite sample result (11), it follows that

$$E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} = E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt} + O\left(\frac{\log(K(n))}{\sqrt{np}}\right).$$

Convergence in probability follows from Lemma 2. This completes the proof. □

We now present a simple bound for $f(M, K(n), np)$. Let $c = 2(v + m/3)$ and $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp(-u^2/2)du$ be the standard normal cumulative distribution function (c.d.f.). By using the facts that $x^2np/(v + mx/3) \geq xnp/(v + m/3)$ for $x \geq 1$ and $x^2np/(v + mx/3) \geq x^2np/(v + m/3)$ for $x \leq 1$, the integral f can be bounded by the following analytical expression:

$$\begin{aligned}
f(M, K(n), np) &\leq 2\Bigg[u_n + K(n)\frac{c}{np}\exp\left(-\max(1, u_n)\frac{np}{c}\right) \\
&+ I(u_n \leq 1)K(n)\sqrt{\frac{c\pi}{np}}\left(\Phi\left(\sqrt{\frac{2np}{c}}\right) - \Phi\left(u_n\sqrt{\frac{2np}{c}}\right)\right)\Bigg].
\end{aligned}$$

Theorem 1 provides a finite sample bound $f(M, K(n), np)$ for the expected value of the predictive loss $\tilde{PL}_{n(1-p)} = \tilde{\theta}_{n(1-p)}(\hat{k}) - \tilde{\theta}_{n(1-p)}(\tilde{k})$, which compares the performance of the cross-validated selector $\hat{k}$ to the benchmark $\tilde{k}$ in terms of the conditional risk $\tilde{\theta}_{n(1-p)}(k)$ based on $n(1-p)$ training

observations. This bound is used to prove that the ratio of expected risk differences $(E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt})/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to one and also that $E\tilde{PL}_{n(1-p)}/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to zero.

However, one would like the cross-validated selector $\hat{k}$ to perform as well as a benchmark selector based on the whole sample of size n, rather than only $n(1-p)$ as above. The following is an immediate corollary of Theorem 1, which relates the conditional risk of the cross-validated selector, $\tilde{\theta}_{n(1-p)}(\hat{k})$, to that of a benchmark selector based on n observations, $\tilde{\theta}_n(\tilde{k}_n)$. In this corollary, we use the notation $p = p_n$ to emphasize the dependence of the validation set proportion p on n.

Corollary 1 *Let*

$$\tilde{\theta}_n(k) = \int L(y, C_k(x \mid P_n))dP(x, y)$$

and let

$$\tilde{k}_n = \text{argmin}_{k\in\{1,\ldots,K(n)\}}\tilde{\theta}_n(k)$$

denote its minimizer. Let $\tilde{k}_{n(1-p_n)}$ now denote the previously defined $\tilde{k}$ which minimizes

$$\tilde{\theta}_{n(1-p_n)}(k) = E_{S_n}\int L(y, C_k(x \mid P^0_{n,S_n}))dP(x, y).$$

If $\sup_{X,Y} L(Y, C_k(X \mid P_n)) \leq M < \infty$ a.s. for all k, where the supremum is over a support of the distribution of X and Y, and, as $n \to \infty$, $p = p_n \to 0$, $\log(K(n))/\sqrt{np_n} \to 0$, $\frac{\log(K(n))}{\sqrt{np_n}(\tilde{\theta}_{n(1-p_n)}(\tilde{k})-\theta_{opt})} \xrightarrow{P} 0$, and

$$\frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}} \xrightarrow{P} 1, \tag{13}$$

then

$$\frac{\tilde{\theta}_{n(1-p_n)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}} \xrightarrow{P} 1.$$

A sufficient condition for (13) is that there exists $\gamma > 0$ such that

$$\left(n^\gamma\left(\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}\right),\ (n(1-p_n))^\gamma\left(\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}\right)\right) \xRightarrow{D} (Z, Z),$$

for a random variable Z with $Pr(Z > a) = 1$ for some $a > 0$. In particular, for the single split case with $Pr(S_n = s) = 1$ for some $s \in \{0,1\}^n$, it suffices to assume that there exists $\gamma > 0$ such that $n^\gamma\left(\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}\right) \xRightarrow{D} Z$, for a random variable Z with $Pr(Z > a) = 1$ for some $a > 0$.

We expect the last condition in the corollary to be sufficient for averages over multiple splits as well as single splits.

Proof. Firstly note that by Theorem 1

$$\frac{\tilde{\theta}_{n(1-p_n)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}} \frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}} \quad \xrightarrow{P} \quad 1.$$

This proves the first statement of the corollary. We now show that (13) holds under the first sufficient condition. Define

$$\begin{aligned} Z_{1,n} &= n^{\gamma} \left(\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt} \right) \\ Z_{2,n} &= (n(1-p_n))^{\gamma} \left(\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt} \right). \end{aligned}$$

If $(Z_{1,n}, Z_{2,n}) \overset{D}{\Rightarrow} (Z, Z)$, then by the continuous mapping theorem we have $\frac{Z_{1,n}}{Z_{2,n}} \overset{D}{\Rightarrow} 1$. However, note that

$$\frac{Z_{1,n}}{Z_{2,n}} = \frac{1}{(1-p_n)^{\gamma}} \frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}}.$$

Thus, if $p_n \to 0$,

$$\frac{\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt}}{\tilde{\theta}_{n(1-p_n)}(\tilde{k}_{n(1-p_n)}) - \theta_{opt}} \xrightarrow{P} 1,$$

and (13) holds. If there is only one split, i.e., $P(S_n = s) = 1$ for some $s \in \{0,1\}^n$, then $Z_{1,n} \overset{D}{=} Z_{2,\frac{n}{1-p_n}}$, and hence $Z_{1,n} \overset{D}{\Rightarrow} Z$ implies $(Z_{1,n}, Z_{2,n}) \overset{D}{\Rightarrow} (Z, Z)$. This completes the proof.

□

An important and practical issue is the impact of the cross-validation proportion p on the risk estimators. The following discussion provides some intuition regarding the behavior of the conditional risk $\tilde{\theta}_{n(1-p)}(k)$ compared to the conditional risk $\tilde{\theta}_n(k)$ for a predictor based on the entire empirical distribution P_n. One can argue that, due to the expectation with respect to (w.r.t.) S_n in the definition of $\tilde{\theta}_{n(1-p)}(k)$, the first order linear approximation of $\tilde{\theta}_{n(1-p)}(k) - \tilde{\theta}_n(k)$ equals zero for each fixed $p \in (0,1)$. This is formalized by the following argument. Let $\theta_k = \int L(y, C_k(x \mid P)) dP(x,y)$ denote the

parameter corresponding with $\tilde{\theta}_n(k)$, i.e., the asymptotic risk for the predictor C_k. Suppose

$$\tilde{\theta}_n(k) - \theta_k = \frac{1}{n}\sum_{i=1}^{n} IC_k(X_i \mid P) + R_k(P_n, P)$$

for some function $IC_k(\cdot \mid P)$ of X and remainder term $R_k(P_n, P)$. Then

$$\tilde{\theta}_{n(1-p)}(k) - \theta_k = E_{S_n}\frac{1}{n(1-p)}\sum_{i=1}^{n} IC_k(X_i \mid P)I(S_{n,i} = 0) + E_{S_n}R_k(P^0_{n,S_n}, P).$$

Now, note that, due to the expectation w.r.t. S_n, the first term actually equals $\sum_{i=1}^{n} IC_k(X_i \mid P)/n$. Consequently,

$$\tilde{\theta}_n(k) - \tilde{\theta}_{n(1-p)}(k) = R_k(P_n, P) - E_{S_n}R_k(P^0_{n,S_n}, P).$$

Thus, even for a fixed $p \in (0, 1)$, $\tilde{\theta}_{n(1-p)}(k)$ can be viewed as a decent approximation of $\tilde{\theta}_n(k)$. This suggests that averaging over S_n significantly reduces the sensitivity of the cross-validated selector $\hat{k}$ to the choice of p compared to single split cross-validation. Preliminary sensitivity analysis results concerning the validation proportion p are discussed in a related article on likelihood-based cross-validation (van der Laan et al., 2003).

3.2 Squared error loss function

In this section, we derive a stronger analog of Theorem 1 for the squared error loss function. In this special case, one can prove that the ratio of expected risk differences $(E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt})/(E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt})$ converges to one at rate $O(\log(K(n))/np)$ rather than the slower $O(\log(K(n))/\sqrt{np})$ applicable to general loss functions. The proof is related to that of Theorem 7.1, p. 101, in Györfi et al. (2002) for the single split case (i.e., without averaging over S_n). We correct the proof of Theorem 7.1 do deal with the fact that, in the notation of Györfi et al. (2002), H is a random variable in the expressions for $Pr(T_{1,n} \geq s|D_{n_l})$ on the last line of p. 102 and the first line of p. 103. Our result in Theorem 2 below is more general than the single split result of Györfi et al. (2002), as we consider risk estimators based on multiple random splits of the learning set based on the random vector S_n. In order to deal with the expected value over S_n, an extra term similar to T_{1n} is introduced. Finally, we point out that a finite sample result similar to that in Györfi et al. (2002) implies the asymptotic optimality of the cross-validated selector $\hat{k}$ under appropriate conditions.

Theorem 2 *Suppose that* $Pr(|Y| < M) = 1$ *and* $\sup_X |C_k(X \mid P_n)| \leq M < \infty$ *a.s. for all* k*, where the supremum is over a support of the marginal distribution of* X*. Let* $M_1 = 8M^2$ *and* $M_2 = 16M^2$*. For any* $\delta > 0$*, we have*

$$0 \leq E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \leq (1+2\delta)\left\{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\right\} + 2c(M,\delta)\frac{1+\log(K(n))}{np},$$

where

$$c(M,\delta) = 2(1+\delta)^2\left(\frac{M_1}{3} + \frac{M_2}{\delta}\right).$$

This finite sample result has the following asymptotic implications. If

$$\frac{\log(K(n))}{(np)\{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\}} \to 0 \text{ for } n \to \infty,$$

then

$$\frac{E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \to 1 \text{ for } n \to \infty.$$

Similarly, if

$$\frac{\log(K(n))}{(np)\{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\}} \xrightarrow{P} 0 \text{ for } n \to \infty,$$

then

$$\frac{\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt}}{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}} \xrightarrow{P} 1 \text{ for } n \to \infty.$$

Proof. We have

$$\begin{aligned}
0 \;\leq\; & \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \\
= \;& E_{S_n}\int L(y, C_{\hat{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x))dP(x,y) \\
& -(1+\delta)E_{S_n}\int L(y, C_{\hat{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x))dP^1_{n,S_n}(x,y) \\
& +(1+\delta)E_{S_n}\int L(y, C_{\hat{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x))dP^1_{n,S_n}(x,y) \\
\leq \;& E_{S_n}\int L(y, C_{\hat{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x))dP(x,y) \\
& -(1+\delta)E_{S_n}\int L(y, C_{\hat{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x))dP^1_{n,S_n}(x,y) \\
& +(1+\delta)E_{S_n}\int L(y, C_{\tilde{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x))dP^1_{n,S_n}(x,y)
\end{aligned}$$

$$
\begin{aligned}
= \; & E_{S_n} \int L(y, C_{\hat{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP(x,y) \\
& -(1+\delta) E_{S_n} \int L(y, C_{\hat{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP^1_{n,S_n}(x,y) \\
& +(1+\delta) E_{S_n} \int L(y, C_{\tilde{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP^1_{n,S_n}(x,y) \\
& -(1+2\delta) E_{S_n} \int L(y, C_{\tilde{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP(x,y) \\
& +(1+2\delta) E_{S_n} \int L(y, C_{\tilde{k}}(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP(x,y),
\end{aligned}
$$

where the first inequality follows by definition of θ_{opt} and the second by definition of $\hat{k}$, such that $\hat{\theta}_{n(1-p)}(\hat{k}) \leq \hat{\theta}_{n(1-p)}(k) \; \forall \; k$. Denote the first two terms in the last expression by $R_{n,\hat{k}}$ and the third and fourth terms by $T_{n,\tilde{k}}$; the last term is the benchmark $(1+2\delta)\{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\}$. Hence

$$
0 \leq \tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \leq (1+2\delta)\{\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\} + R_{n,\hat{k}} + T_{n,\tilde{k}}. \tag{14}
$$

In the sequel, we show that $ER_{n,\hat{k}} + ET_{n,\tilde{k}} \leq 2c(M,\delta)(1+\log(K(n)))/np$. Represent $R_{n,k}$ and $T_{n,k}$ as $R_{n,k} = E_{S_n} R_{n,k}(S_n)$ and $T_{n,k} = E_{S_n} T_{n,k}(S_n)$, respectively. For convenience, introduce the following notation for the relevant random variables

$$
\begin{aligned}
\tilde{H}_k &\equiv \int L(y, C_k(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP(x,y) \\
\bar{H}_k &\equiv \int L(y, C_k(x \mid P^0_{n,S_n})) - L(y, C_{opt}(x)) dP^1_{n,S_n}(x,y),
\end{aligned}
$$

where, by definition of C_{opt}, $\tilde{H}_k \geq 0 \; \forall \; k$. Rewrite $R_{n,k}(S_n)$ and $T_{n,k}(S_n)$ as

$$
R_{n,k}(S_n) = (1+\delta)\left[\tilde{H}_k - \bar{H}_k\right] - \delta \tilde{H}_k
$$

and

$$
T_{n,k}(S_n) = (1+\delta)\left[\bar{H}_k - \tilde{H}_k\right] - \delta \tilde{H}_k.
$$

Note that

$$
\begin{aligned}
& Pr(R_{n,\hat{k}}(S_n) > s \mid P^0_{n,S_n}, S_n) \\
& \quad = \; Pr\left(\tilde{H}_{\hat{k}} - \bar{H}_{\hat{k}} > \frac{1}{1+\delta}\left\{s + \delta \tilde{H}_{\hat{k}}\right\} \mid P^0_{n,S_n}, S_n\right) \\
& \quad \leq \; K(n) \max_{k \in \{1,\ldots,K(n)\}} Pr\left(\tilde{H}_k - \bar{H}_k > \frac{1}{1+\delta}\left\{s + \delta \tilde{H}_k\right\} \mid P^0_{n,S_n}, S_n\right).
\end{aligned}
$$

Similarly, for $T_{n,\tilde{k}}(S_n)$,

$$\begin{aligned} & Pr(T_{n,\tilde{k}}(S_n) > s \mid P^0_{n,S_n}, S_n) \\ & \quad = K(n) \max_{k \in \{1,\ldots,K(n)\}} Pr\left(\bar{H}_k - \tilde{H}_k > \frac{1}{1+\delta}\left\{s + \delta \tilde{H}_k\right\} \mid P^0_{n,S_n}, S_n\right). \end{aligned}$$

Conditional on P^0_{n,S_n} and S_n, consider the random variable

$$Z_k = L(Y, C_k(X \mid P^0_{n,S_n})) - L(Y, C_{opt}(X)).$$

Let $Z_{k,i}$, $i = 1, \ldots, np$, be the np i.i.d. copies of Z_k corresponding with X_i, given $S_{n,i} = 1$. Note that $\bar{H}_k = 1/np \sum_{i=1}^{np} Z_{k,i}$ and $\tilde{H}_k = E(Z_k \mid P^0_{n,S_n}, S_n)$, so that $\tilde{H}_k - \bar{H}_k = E(Z_k \mid P^0_{n,S_n}, S_n) - 1/np \sum_{i=1}^{np} Z_{k,i}$ represents an empirical mean of np centered i.i.d. random variables. For the squared error loss function, the random variables Z_k are bounded with $\mid Z_k \mid \leq 4M^2$ a.s. We will apply Bernstein's inequality to the centered empirical mean $\tilde{H}_k - \bar{H}_k$ and exploit the following special property of Z_k for the squared error loss function, to obtain an $\exp(-nps/c)$ tail probability instead of the usual $\exp(-nps^2/(a+bs))$, for constants a, b, $c < \infty$. This will show that the risk differences converge at a $\log(K(n))/np$ rate instead of the usual $\log(K(n))/\sqrt{np}$ as in Theorem 1.

Lemma 3 *Conditional on P^0_{n,S_n} and S_n, let $Z_k = L(Y, C_k(X \mid P^0_{n,S_n})) - L(Y, C_{opt}(X))$, where L denotes the squared error loss function $L(y, \hat{y}) = (\hat{y} - y)^2$. Then,*

$$\sigma_k^2 \equiv VAR(Z_k \mid P^0_{n,S_n}, S_n) \leq M_2 E(Z_k \mid P^0_{n,S_n}, S_n) = M_2 \tilde{H}_k,$$

where $M_2 = 16M^2$.

Proof. Note that by definition of C_{opt}, $\tilde{H}_k = E(Z_k \mid P^0_{n,S_n}, S_n) \geq 0 \ \forall k$. In addition, for the squared error loss, $C_{opt}(X) = E(Y|X)$ and

$$\begin{aligned} Z_k &= (Y - C_k(X \mid P^0_{n,S_n}))^2 - (Y - C_{opt}(X))^2 \\ &= (C_{opt}(X) - C_k(X \mid P^0_{n,S_n}))(2Y - C_k(X \mid P^0_{n,S_n}) - C_{opt}(X)). \end{aligned}$$

Thus,

$$\begin{aligned} E(Z_k \mid P^0_{n,S_n}, S_n) &= E\left(E(Z_k \mid P^0_{n,S_n}, S_n, X) \mid P^0_{n,S_n}, S_n\right) \\ &= E((C_{opt}(X) - C_k(X \mid P^0_{n,S_n})) \\ &\qquad (2E(Y|X) - C_k(X \mid P^0_{n,S_n}) - C_{opt}(X)) \mid P^0_{n,S_n}, S_n) \\ &= E\left((C_{opt}(X) - C_k(X \mid P^0_{n,S_n}))^2 \mid P^0_{n,S_n}, S_n\right). \end{aligned}$$

Hence, using the fact that $|2Y - C_k(X \mid P^0_{n,S_n}) - C_{opt}(X)| \leq 4M$ a.s.

$$\begin{aligned} VAR(Z_k \mid P^0_{n,S_n}, S_n) &\leq E(Z_k^2 \mid P^0_{n,S_n}, S_n) \\ &\leq (4M)^2 E\left((C_{opt}(X) - C_k(X \mid P^0_{n,S_n}))^2 \mid P^0_{n,S_n}, S_n\right) \\ &= M_2 E(Z_k \mid P^0_{n,S_n}, S_n) = M_2 \tilde{H}_k. \end{aligned}$$

□

From the above Lemma and Bernstein's inequality (Lemma 1), for $s > 0$ and $M_1 = 8M^2$,

$$\begin{aligned} &Pr(R_{n,k}(S_n) > s \mid P^0_{n,S_n}, S_n) \\ &\quad = Pr\left(\tilde{H}_k - \bar{H}_k > \frac{1}{1+\delta}\left[s + \delta \tilde{H}_k\right] \mid P^0_{n,S_n}, S_n\right) \\ &\quad \leq Pr\left(\tilde{H}_k - \bar{H}_k > \frac{1}{1+\delta}\left[s + \delta \sigma_k^2/M_2\right] \mid P^0_{n,S_n}, S_n\right) \\ &\quad \leq \exp\left(-\frac{np}{2(1+\delta)^2} \frac{\left(s + \delta\sigma_k^2/M_2\right)^2}{\sigma_k^2 + \frac{M_1}{3(1+\delta)}(s + \delta\sigma_k^2/M_2)}\right). \end{aligned}$$

Note that

$$\begin{aligned} \frac{\left(s + \delta\sigma_k^2/M_2\right)^2}{\sigma_k^2 + \frac{M_1}{3(1+\delta)}(s + \delta\sigma_k^2/M_2)} &= \frac{(s + \delta\sigma_k^2/M_2)}{\frac{\sigma_k^2}{s+\delta\sigma_k^2/M_2} + \frac{M_1}{3(1+\delta)}} \geq \frac{(s + \delta\sigma_k^2/M_2)}{\frac{M_2}{\delta} + \frac{M_1}{3}} \\ &\geq \frac{s}{\frac{M_2}{\delta} + \frac{M_1}{3}}. \end{aligned}$$

This shows that, for $s > 0$,

$$Pr(R_{n,\hat{k}}(S_n) > s \mid P^0_{n,S_n}, S_n) \leq K(n) \exp\left(-\frac{np}{c(M,\delta)} s\right),$$

where $c(M,\delta) = 2(1+\delta)^2\left(\frac{M_1}{3} + \frac{M_2}{\delta}\right)$, with $M_1 = 8M^2$ and $M_2 = 16M^2$. In particular, this provides us with a bound for the marginal probability

$$Pr(R_{n,\hat{k}}(S_n) > s) \leq K(n) \exp\left(-\frac{np}{c(M,\delta)} s\right).$$

As in the proof of Theorem 1, for each $u > 0$, we have

$$ER_{n,\hat{k}} \leq u + \int_u^\infty K(n) \exp\left(-\frac{np}{c(M,\delta)} s\right) ds.$$

The minimum is attained at $u_n = c(M,\delta)\log(K(n))/np$ and is given by $c(M,\delta)(\log(K(n))+1)/np$. Thus,

$$ER_{n,\hat{k}} \leq c(M,\delta)\frac{1+\log(K(n))}{np}.$$

Similarly for $ET_{n,\tilde{k}}$. Taking the expected values of the quantities in equation (14) yields the finite sample result

$$0 \leq E\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt} \leq (1+2\delta)\left\{E\tilde{\theta}_{n(1-p)}(\tilde{k}) - \theta_{opt}\right\} + 2c(M,\delta)\frac{1+\log(K(n))}{np},$$

where

$$c(M,\delta) = 2(1+\delta)^2\left(\frac{M_1}{3} + \frac{M_2}{\delta}\right).$$

Again, as in Theorem 1, convergence in probability follows from convergence in expectation by Lemma 2. This completes the proof of Theorem 2.

□

Note that Corollary 1 applies also in this setting, with suitable modifications to reflect the assumptions of Theorem 2 and the improved rate of convergence.

3.3 Translation of risk convergence into predictor convergence

An interesting issue is the translation of rate of convergence results for cross-validated risk estimators into rate of convergence results for the corresponding selected predictors. Specifically, we are examining conditions under which convergence to zero in probability at a given rate of the risk difference

$$\tilde{\theta}_n(\hat{k}) - \theta_{opt} = \int L(y, C_{\hat{k}}(x \mid P_n)) - L(y, C_{opt}(x))dP(x,y)$$

implies convergence of the selected predictor $C_{\hat{k}}(\cdot \mid P_n)$ to the optimal predictor C_{opt}, in the sense that

$$\int L(C_{\hat{k}}(x \mid P_n), C_{opt}(x))dP(x) \xrightarrow{P} 0$$

at that rate. This result is immediate for the squared error loss function, as

$$\tilde{\theta}_n(\hat{k}) - \theta_{opt} \;\; = \;\; \int L(C_{\hat{k}}(x \mid P_n), C_{opt}(x))dP(x),$$

where we use the fact that $C_{opt}(X) = E_P(Y|X)$.

We are investigating similar inversion results for other loss functions under appropriate assumptions.

4 Results for predictor performance assessment

In this section, we establish consistency and asymptotic linearity of the CV estimator $\hat{\theta}_{n(1-p)}$ and the resubstitution estimator $\hat{\theta}_n^{LS}$ for the conditional risk $\tilde{\theta}_n$. The results hold for general loss functions and cross-validation procedures.

4.1 Asymptotic linearity of the cross-validated risk estimator

We first derive a consistency and asymptotic linearity result for the cross-validated estimator $\hat{\theta}_{n(1-p)}$ as an estimator of the conditional risk $\tilde{\theta}_{n(1-p)}$ based on $n(1-p)$ training observations.

Theorem 3 *Assume that* $\sup_{X,Y} L(Y, C(X \mid P_n)) \leq M < \infty$ *a.s., where the supremum is over a support of the distribution of* X *and* Y*, and*

$$\frac{E_{S_n}\sqrt{\int \left\{L(y, C(x \mid P^0_{n,S_n})) - L(y, C(x \mid P))\right\}^2 dP(x,y)}}{\sqrt{p_n}} = o_P(1). \qquad (15)$$

Then

$$\hat{\theta}_{n(1-p)} - \tilde{\theta}_{n(1-p)} = \frac{1}{n}\sum_{i=1}^{n} \{L(Y_i, C(X_i \mid P)) - \theta\} + o_P(1/\sqrt{n}).$$

Proof. We have

$$\begin{aligned}
&\hat{\theta}_{n(1-p)} - \tilde{\theta}_{n(1-p)} \\
&\quad = E_{S_n} \int L(y, C(x \mid P^0_{n,S_n})) - L(y, C(x \mid P)) d(P^1_{n,S_n} - P)(x,y) \\
&\qquad + E_{S_n} \int L(y, C(x \mid P)) d(P^1_{n,S_n} - P)(x,y).
\end{aligned}$$

The last term equals

$$\int L(y, C(x \mid P))d(P_n - P)(x, y) = \frac{1}{n}\sum_{i=1}^{n}\left\{L(Y_i, C(X_i \mid P)) - \theta\right\}.$$

Denote the first term as $T_n = E_{S_n}T_n(S_n)$. We wish to show that it is $o_P(1/\sqrt{n})$. Conditional on P^0_{n,S_n} and S_n, consider the random variable

$$Z_n = L(Y, C(X \mid P^0_{n,S_n})) - L(Y, C(X \mid P)),$$

and let $Z_{n,i}$, $i = 1, \ldots, np$, be the np i.i.d. copies of Z_n corresponding with X_i, given $S_{n,i} = 1$. Then $T_n(S_n)$ can be written as an empirical mean of np centered random variables $1/np\sum_{i=1}^{np} Z_{n,i} - E(Z_n \mid P^0_{n,S_n}, S_n)$, with $|Z_n| < M$ a.s. Let

$$\sigma_n^2(S_n) = \max((np)^{-1}, E(Z_n^2 \mid P^0_{n,S_n}, S_n)).$$

Then, $VAR(Z_n \mid P^0_{n,S_n}, S_n) \leq \sigma_n^2(S_n)$. From Bernstein's inequality, with $W = 2M$,

$$Pr(\mid T_n(S_n) \mid > x \mid P^0_{n,S_n}, S_n) \leq 2\exp\left(-\frac{1}{2}\frac{npx^2}{\sigma_n^2(S_n) + Wx/3}\right).$$

Thus

$$\begin{aligned}
E_{S_n}|T_n(S_n)| &= E_{S_n}\int_0^\infty Pr(\mid T_n(S_n) \mid > x \mid P^0_{n,S_n}, S_n)dx \\
&\leq E_{S_n}\int_0^\infty 2\exp\left(-\frac{1}{2}\frac{npx^2}{\sigma_n^2(S_n) + Wx/3}\right)dx \\
&= E_{S_n}\frac{\sigma_n(S_n)}{\sqrt{np}}\int_0^\infty 2\exp\left(-\frac{1}{2}\frac{y^2}{1 + \frac{W}{3\sqrt{np}\sigma_n(S_n)}y}\right)dy,
\end{aligned}$$

where we carried out the substitution $x = [\sigma_n(S_n)/\sqrt{np}]y$. Since $\sigma_n(S_n) > (np)^{-0.5}$, we have that the integral is bounded uniformly in n by

$$C = \int_0^\infty 2\exp\left(-\frac{1}{2}\frac{y^2}{1 + Wy/3}\right)dy.$$

By assumption (15),

$$\frac{E_{S_n}\sqrt{E(Z_n^2 \mid P^0_{n,S_n}, S_n)}}{\sqrt{p_n}} = o_P(1),$$

which proves that

$$E_{S_n}|T_n(S_n)| \leq \frac{C}{\sqrt{np}} E_{S_n}\sigma_n(S_n) = o_P(1/\sqrt{n}).$$

□

Note that by Jensen's inequality, a sufficient condition for (15) is

$$\frac{1}{p_n} E_{S_n} \int \left\{ L(y, C(x \mid P^0_{n,S_n})) - L(y, C(x \mid P) \right\}^2 dP(x,y) = o_P(1).$$

The argument following Corollary 1 can also be applied in this context, to suggest that the risk difference $\tilde{\theta}_{n(1-p)} - \tilde{\theta}_n$ is zero in first order. Thus, by virtue of the expectation w.r.t. to S_n, one expects the cross-validated risk estimator $\hat{\theta}_{n(1-p)}$ to be a decent approximation of the conditional risk $\tilde{\theta}_n$ for a fixed validation proportion $p \in (0,1)$.

The asymptotic linearity result in Theorem 3 allows us to derive confidence intervals for the conditional risk $\tilde{\theta}_{n(1-p)}$. Specifically, let

$$IC(X, Y \mid P) = L(Y, C(X \mid P)) - \theta.$$

Then $E_P(IC(X, Y \mid P)) = 0$ and

$$\sigma^2 = VAR_P(IC(X, Y \mid P)) = \int (IC(x, y \mid P))^2 dP(x, y).$$

From the Central Limit Theorem, as $n \to \infty$, $\sqrt{n}(\hat{\theta}_{n(1-p)} - \tilde{\theta}_{n(1-p)})/\sigma$ converges in distribution to a standard normal random variable. Apply the following resubstitution estimators for $IC(X, Y \mid P)$ and its variance σ^2,

$$\hat{IC}(X, Y \mid P) = IC(X, Y \mid P_n) = L(Y, C(X \mid P_n)) - \hat{\theta}^{LS}_n$$

$$\hat{\sigma}^2_n = \int (IC(x, y \mid P_n))^2 dP_n(x, y).$$

An approximate asymptotic $(1-\alpha)100\%$ confidence interval for $\tilde{\theta}_{n(1-p)}$ is given by

$$\hat{\theta}_{n(1-p)} \pm z_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}},$$

where $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ for the standard normal cumulative distribution function $\Phi(\cdot)$.

4.2 Asymptotic linearity of the resubstitution risk estimator

The previous theorem for the cross-validated estimator $\hat{\theta}_{n(1-p)}$ applies to the resubstitution estimator $\hat{\theta}_n^{LS}$ as well, but under different conditions. In particular, the proof follows a different approach than the proofs of Theorems 1, 2, and 3, and relies on the weak convergence theory for empirical processes and the definition of a P-Donsker class (van der Vaart and Wellner, 1996).

Theorem 4 *Suppose $\mathcal{C}$ is a class of functions of X so that $Pr(C(X \mid P_n) \in \mathcal{C}) \to 1$, $\int \{L(y, C(x \mid P_n)) - L(y, C(x \mid P))\}^2 dP(x,y) = o_P(1)$, and let $\{(x,y) \to L(y, C(x)) - L(y, C(x \mid P)) : C \in \mathcal{C}\}$ be a P-Donsker class. Then*

$$\hat{\theta}_n^{LS} - \tilde{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \{L(Y_i, C(X_i \mid P)) - \theta\} + o_P(1/\sqrt{n}).$$

Proof. We have

$$\begin{aligned} \hat{\theta}_n^{LS} - \tilde{\theta}_n &= \int \{L(y, C(x \mid P_n)) - L(y, C(x \mid P))\} \, d(P_n - P)(x,y) \\ &\quad + \int L(y, C(x \mid P)) d(P_n - P)(x,y). \end{aligned}$$

The second term is the desired linear component

$$\frac{1}{n} \sum_{i=1}^{n} \{L(Y_i, C(X_i \mid P)) - \theta\}.$$

In order to show that the first term is $o_P(1/\sqrt{n})$, we appeal to Lemma 2.3.11, p. 115, in van der Vaart and Wellner (1996). Consider the empirical process

$$G_n(f) = \int f(x,y) d\sqrt{n}(P_n - P)(x,y),$$

indexed by $f \in \mathcal{F}$, where $\mathcal{F}$ is the assumed P-Donsker class $\{(x,y) \to L(y, C(x)) - L(y, C(x \mid P)) : C \in \mathcal{C}\}$. Then, by Lemma 2.3.11, $\{G_n(f) : f \in \mathcal{F}\}$ is tight, and thereby

$$\sup_{\{f \in \mathcal{F} : \int f^2 dP \leq \delta_n\}} |G_n(f)| \xrightarrow{P} 0$$

for any sequence $\delta_n \downarrow 0$. Let $f_n(X,Y) \equiv L(Y, C(X \mid P_n)) - L(Y, C(X \mid P))$. By the assumptions of the theorem, we have

$$Pr\left(\int f_n^2(x,y)dP(x,y) \leq \delta_n\right) \to 1$$

for some sequence $\delta_n \downarrow 0$. Consequently,

$$Pr\left(|G_n(f_n)| \leq \sup_{\{f \in \mathcal{F}: \int f^2 dP \leq \delta_n\}} |G_n(f)|\right) \to 1.$$

But, $\sup_{\{f \in \mathcal{F}: \int f^2 dP \leq \delta_n\}} |G_n(f)| \xrightarrow{P} 0$. This proves

$$G_n(f_n) = \int f_n(x,y) d\sqrt{n}(P_n - P)(x,y) = o_P(1).$$

□

In spite of the good asymptotic behavior of resubstitution risk estimators under the conditions of Theorem 4, a number of caveats are in order. Firstly, resubstitution estimators can be severely biased downward in finite sample situations. Secondly, our proof of consistency and asymptotic linearity for the resubstitution risk estimator requires stronger assumptions than the analogue for the CV risk estimators. Finally, resubstitution estimators perform poorly in model selection as they tend to over-fit.

5 Summary and discussion

This article derived distributional properties of cross-validated risk estimators in the context of model selection and performance assessment. General loss functions were considered, including the absolute and squared error loss functions for continuous outcomes and the indicator loss function for polychotomous outcomes. A broad definition of cross-validation was used in order to cover leave-one-out cross-validation, V-fold cross-validation, and Monte Carlo cross-validation.

For model selection, the asymptotic optimality of cross-validation procedures was established, in the sense that a selector based on a cross-validated risk estimator performs asymptotically as well as an optimal benchmark selector based on the risk for the true unknown distribution. That is, for a fixed validation set proportion $p \in (0,1)$, the ratio of conditional risk differences

comparing the cross-validated selector $\hat{k}$ to the benchmark selector $\tilde{k}_{n(1-p)}$, $(\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt})/(\tilde{\theta}_{n(1-p)}(\tilde{k}_{n(1-p)}) - \theta_{opt})$, converges to one in probability (Theorems 1 and 2). For a sequence $p = p_n$ converging to zero slowly enough with sample size n, we showed asymptotic equivalence of $\hat{k}$ and the absolutely optimal benchmark selector $\tilde{k}_n$ based on the entire empirical distribution P_n, that is, $(\tilde{\theta}_{n(1-p)}(\hat{k}) - \theta_{opt})/(\tilde{\theta}_n(\tilde{k}_n) - \theta_{opt})$ converges to one in probability (Corollary 1). In the special case of the squared error loss function, Theorem 2 provides a stronger convergence result than Theorem 1: for the L_2 loss function, the rate of convergence is shown to be $O(\log(K(n))/np)$ rather than the slower $O(\log(K(n))/\sqrt{np})$ applicable to general loss functions. An important condition for the theorems is that the size np of the validation sets converges to infinity; this rules out leave-one-out cross-validation.

For predictor performance assessment, cross-validated risk estimators were shown, under certain conditions, to be consistent and asymptotically linear for the conditional risk $\tilde{\theta}_{n(1-p)}$ based on the true underlying distribution P (Theorem 3). This asymptotic linearity result allowed us to derive confidence intervals for the conditional risk $\tilde{\theta}_{n(1-p)}$. We are performing simulation studies to examine the coverage properties of these confidence intervals.

The argument following Corollary 1 provided some intuition regarding the behavior of the conditional risk $\tilde{\theta}_{n(1-p)}$, averaged over empirical distributions P^0_{n,S_n} for $n(1-p)$ training observations, compared to the conditional risk $\tilde{\theta}_n$, based on the entire empirical distribution P_n. It suggested, in particular, that averaging over S_n significantly reduces the sensitivity of the cross-validated selector $\hat{k}$ to the choice of p compared to single split cross-validation. We plan to study the sensitivity to the validation proportion p in more detail and to develop and test a proposal for a data adaptive choice $\hat{p}$.

Section 3.3 raised the issue of translating rate of convergence results for cross-validated risk estimators into rate of convergence results for the corresponding selected predictors. The translation was shown to be immediate in the case of the squared error loss function. Similar inversion results can be derived for other loss functions under appropriate assumptions and will be investigated in ongoing research.

Finally, we have used a similar general framework as in the present article to derive optimality results for likelihood-based cross-validation (van der Laan et al., 2003) and cross-validated model selection for regression on censored outcomes (Keleş et al., 2003).

References

C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci.*, 99(10): 6562–6566, 2002.

L. Breiman. The little bootstrap and other methods for dimensionality selection in regression: x-fixed prediction error. *Journal of the American Statistical Association*, 87(419):738–754, 1992.

L. Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383, 1996a.

L. Breiman. Out-of-bag estimation. Technical report, Department of Statistics, U.C. Berkeley, 1996b.

L. Breiman. Arcing classifiers. *Annals of Statistics*, 26:801–824, 1998.

L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth International Group, 1984.

L. Breiman and P. Spector. Submodel selection and evaluation in regression. the x random case. *International Statistical Review*, 60:291–319, 1992.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

L. Györfi, M. Kohler, A. Krzyźak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer-Verlag, New York, 2002.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. Springer Verlag, 2001.

S. Keleş, M. J. van der Laan, and S. Dudoit. Asymptotically optimal model selection method for regression on censored outcomes. *Bernoulli*, 2003. (Submitted).

G. J. McLachlan. *Discriminant analysis and statistical pattern recognition*. Wiley, New York, 1992.

B. D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge, New York, 1996.

M. Stone. Cross-validatory choice and assessment of statistics predictions. *Journal of the Royal Statistical Society, Series B*, 36(2):111–147, 1974.

M. Stone. Asymptotics for and against cross-validation. *Biometrika*, 64(1): 29–35, 1977.

M. J. van der Laan, S. Dudoit, and S. Keleş. Asymptotic optimality of likelihood based cross-validation. *Biometrika*, 2003. (Submitted).

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.