University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2003 *Paper* 126

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

Sandrine Dudoit* Mark J. van der Laan†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, sandrine@stat.berkeley.edu

†Division of Biostatistics, School of Public Health, University of California, Berkeley

http://biostats.bepress.com/ucbbiostat/paper126

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

Sandrine Dudoit and Mark J. van der Laan

Abstract

Risk estimation is an important statistical question for the purposes of selecting a good estimator (i.e., model selection) and assessing its performance (i.e., estimating generalization error). This article introduces a general framework for cross-validation and derives distributional properties of cross-validated risk estimators in the context of estimator selection and performance assessment. Arbitrary classes of estimators are considered, including density estimators and predictors for both continuous and polychotomous outcomes. Results are provided for general full data loss functions (e.g., absolute and squared error, indicator, negative log density). A broad definition of cross-validation is used in order to cover leave-one-out cross-validation, V-fold cross-validation, Monte Carlo cross-validation, and bootstrap procedures. For estimator selection, finite sample risk bounds are derived and applied to establish the asymptotic optimality of cross-validation, in the sense that a selector based on a cross-validated risk estimator performs asymptotically as well as an optimal oracle selector based on the risk under the true, unknown data generating distribution. The asymptotic results are derived under the assumption that the size of the validation sets converges to infinity and hence do not cover leave-one-out cross-validation. For performance assessment, cross-validated risk estimators are shown to be consistent and asymptotically linear for the risk under the true data generating distribution and confidence intervals are derived for this unknown risk. Unlike previously published results, the theorems derived in this and our related articles apply to general data generating distributions, loss functions (i.e., parameters), estimators, and cross-validation procedures.

Contents

1 Introduction

1.1 Motivation

Risk estimation is an important statistical question for at least two purposes. Risk estimation is used for: (i) *estimator selection*, or *model selection*, where the "best" estimator is chosen to minimize risk over a given class of estimators; (ii) *estimator performance assessment*, i.e., the estimation of *generalization error*. These two fundamental problems have been referred to variously in the statistical literature as "submodel selection and evaluation" (Breiman, 1992) and "choice and assessment of statistical predictions" (Stone, 1974). For example, regression problems often involve the data-driven selection of a predictor for an outcome Y given covariates W (e.g., linear model with k explanatory variables, regression tree with k terminal nodes), with the intention of predicting the outcome of interest for future observations. A common measure of performance in this context is the mean squared error between predicted and true responses, i.e., the risk for the quadratic loss function. An immediate difficulty is that the risk of a given estimator is the expected value of a loss function under the typically *unknown* data generating distribution. This means that the available data (i.e., the learning set or empirical distribution) have to be used for both tasks (i) and (ii), that is, to select a good estimator (specifically, estimate the risk criterion used to select the estimator) and to assess the performance of this selected estimator.

A number of approaches have been proposed for selection and performance assessment. As discussed by Breiman (1992) in the context of dimensionality selection in regression, criteria such as Mallow's C_p, Akaike's information criterion (AIC), and the Bayesian information criterion (BIC), do not account for the data-driven selection of the sequence of estimators (i.e., submodels) and thus provide biased assessment of generalization error. Instead, risk estimation methods based on sample reuse have been favored in the recent literature. 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), the jackknife, and the bootstrap (Chapter 3, Breiman et al. (1984); Breiman and Spector (1992); Breiman (1996a,b); Burman (1989); Devroye et al. (1996); Efron (1983); Chapter 17, Efron and Tibshirani (1993); Geisser (1975); Gong (1986); Chapters 7 and 8, Györfi et al. (2002a); Chapter 7,

Hastie et al. (2001); Li (1987); McCarthy (1976); Picard and Cook (1984); Chapter 3, Ripley (1996); Shao (1993); Stone (1974, 1977); Zhang (1993)). Another important class of approaches for model selection, described in Barron et al. (1999), uses sieve theory to define penalized empirical loss criteria. Connections with cross-validation methods are discussed in Birgé and Massart (1997).

A variety of *cross-validation* (CV) procedures are available for estimating the risk of a given estimator and for performing estimator selection. A natural question then concerns the distributional properties of the resulting risk estimators, i.e., their performances in terms of identifying a good estimator (model selection) and as estimators of generalization error, 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 empirical assessment of different cross-validation procedures, previous theoretical work has focused primarily on distributional properties for *leave-one-out cross-validation* (Stone, 1974, 1977). However, this special form of CV has well-known limitations, both theoretical and practical, and a number of authors have considered more general *multifold cross-validation* procedures (Breiman et al., 1984; Breiman and Spector, 1992; Burman, 1989; Devroye et al., 1996; Geisser, 1975; Györfi et al., 2002a; McCarthy, 1976; Picard and Cook, 1984; Ripley, 1996; Shao, 1993; Zhang, 1993). With the exception of the general moment derivations of Burman (1989), theoretical investigations of multifold cross-validation procedures have concentrated on linear models (Li, 1987; Shao, 1993; Zhang, 1993). In particular, Li (1987) derives asymptotic optimality results for leave-one-out cross-validation applied to linear model selection using a squared error loss function. Optimality is defined as convergence to one in probability of the ratio of the squared error for the CV selector to the squared error for an optimal benchmark selector based on the unknown mean parameter (Equation (1.6), p. 961, Li (1987)). Again in the special case of linear models, Shao (1993) establishes consistency results for Monte Carlo and balanced incomplete cross-validation, in the sense that the probability of selecting the linear model with the best predictive ability converges to one. Using similar criteria as Shao (1993), Zhang (1993) compares the performances of V-fold, Monte Carlo, and leave-one-out cross-validation in variable selection for linear models. Results of Devroye et al. (1996) and Györfi et al. (2002a) are discussed in Section 3.

Section 2, below, introduces our general loss-based framework for estimator construction, selection, and performance assessment, and provides further

motivation for the present investigation of cross-validation procedures for risk estimation (van der Laan and Dudoit, 2003).

1.2 Outline

The present article proposes a general framework for cross-validation and derives distributional properties of cross-validated risk estimators. This investigation of cross-validation was motivated by the development of a general loss-based methodology for estimator construction, selection, and performance assessment (Section 2). Arbitrary classes of estimators are considered, including density estimators (e.g., non-parametric kernel density estimators, parametric maximum likelihood estimators) and predictors for both continuous and polychotomous outcomes (e.g., estimators based on linear and non-linear models, classification and regression trees, support vector machines). Results are derived for general full data loss functions, such as the absolute and squared error loss functions for the prediction of continuous outcomes, the indicator or general matrix loss functions for the prediction of polychotomous outcomes, and the negative log density loss function for density estimation. A broad definition of cross-validation is considered in order to cover leave-one-out cross-validation, V-fold cross-validation, Monte Carlo cross-validation, and bootstrap procedures. In this general cross-validation framework, the learning set is divided into a training set and a validation set based on the value of a random split vector B_n. For a given B_n, the risk of an estimator 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 B_n to yield the cross-validated risk estimator. The particular distribution of the split vector B_n determines the flavor of the cross-validation procedure.

For estimator 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 (in terms of risk) asymptotically as well as an optimal benchmark or oracle selector based on the risk under the true, unknown data generating distribution. Theorem 2 applies to general full data loss functions (e.g., truncated absolute and squared error, matrix, and negative log density loss functions). In the special case of the quadratic (i.e., squared error or L_2) loss function, Theorem 1 provides a stronger convergence result than Theorem 2: for the L_2 loss function, the rate of convergence is shown to be $O(\log(K_n)/np_n)$ rather than

the slower $O(\log(K_n)/\sqrt{np_n})$ applicable to general loss functions, where p_n denotes the proportion of observations in the validation sets. Note that the related article by van der Laan et al. (2004a) focuses on likelihood-based cross-validation and derives, as in Theorem 1, a stronger $O(\log(K_n)/np_n)$ convergence result for the negative log density loss function. The asymptotic results are derived under the assumption that the size of the validation sets converges to infinity and hence do not cover leave-one-out cross-validation. For estimator performance assessment, cross-validated risk estimators are shown to be consistent and asymptotically linear for the risk under the true data generating distribution (Theorems 3 and 4). The asymptotic linearity result allows the derivation of confidence intervals for the unknown risk under the true distribution (Section 4.3).

The article is organized as follows. Section 2 introduces our general loss-based estimation framework and cross-validation methodology for estimator selection and performance assessment. Section 3 establishes distributional properties and the asymptotic optimality of cross-validated risk estimators in model selection. Section 4 concerns distributional properties of cross-validated risk estimators of generalization error and derives confidence intervals for the risk under the true, unknown data generating distribution. Finally, Section 5 summarizes our findings and discusses related work. We stress that, unlike previously published results, the theorems derived in this and our related articles apply to general data generating distributions, loss functions (i.e., parameters), estimators, and cross-validation procedures.

2 Framework for loss-based estimation with cross-validation

2.1 Estimation road map

In a series of related articles, we have developed a unified *loss-based cross-validation methodology* for estimator construction, selection, and performance assessment in the presence of censoring. Our proposed estimation road map can be stated in terms of the following three main steps.

Step 1. First, define the parameter of interest as the minimizer of the expected loss, or risk, for a full (uncensored) data loss function chosen

to represent the desired measure of performance. Apply the estimating function methodology of van der Laan and Robins (2003) to map the full data loss function into an observed (censored) data loss function having the same expected value and leading to an efficient estimator of this risk.

Step 2. Next, construct a finite collection of candidate estimators for the parameter of interest, based on a sieve of increasing dimension approximating the complete parameter space. For each element of the sieve, the candidate estimator is chosen as the minimizer of the empirical risk.

Step 3. Apply cross-validation to select an optimal estimator among the candidates and to assess the overall performance of the resulting estimator.

The formulation, in **Step 1**, of the estimation problem in terms of a loss function allows the unification and generalization of a broad range of problems that are traditionally treated separately in the statistical literature, including density estimation and the prediction of both (possibly censored and/or multivariate) polychotomous and continuous outcomes (i.e., classification and regression, respectively). For example, for maximum likelihood estimation one would use the negative log density loss function, for least squares regression one would use the squared error (i.e., quadratic or L_2) loss function, and for classification one could use the indicator or a general matrix loss function. In contrast to existing approaches, this unified loss-based framework reconciles censored and full data estimation methods, in the sense that standard full data estimators are recovered as special cases of censored data estimators. The ability to select an appropriate loss function and parameterization of the parameter space confers great flexibility and generality to this approach.

For **Step 2** of the road map, Sinisi and van der Laan (2004) propose general *Deletion/Substitution/Addition algorithms*, or in short *D/S/A algorithms*, for risk minimization over a given parameter subspace. Such D/S/A algorithms are more flexible and aggressive than standard forward/backward or tree-structured approaches and are especially well-suited to handle high-dimensional estimation problems, with higher-order interactions among variables.

The present article is concerned with **Step 3** of the road map in the case of uncensored data, namely, cross-validation for estimator selection and

performance assessment with full data loss functions that do not depend on a nuisance parameter. It establishes finite sample and asymptotic optimality results for the cross-validated selector, for general data generating distributions, full data loss functions (i.e., parameters), estimators, and cross-validation procedures. The results apply, in particular, to regression, classification, and density estimation. The asymptotic optimality results state that the cross-validated selector performs (in terms of risk) asymptotically as well as an optimal oracle selector based on the true, unknown data generating distribution.

The general loss-based estimation framework and its theoretical foundations are established in van der Laan and Dudoit (2003). Special cases and applications are described in a collection of related articles: estimator selection with censored data (Keleş et al., 2004); likelihood-based cross-validation (van der Laan et al., 2004a); cross-validated adaptive ϵ–net estimation (van der Laan et al., 2004b); tree-based estimation with censored data (Molinaro et al., 2004); D/S/A algorithm for generating candidate estimators (Dudoit et al., 2003; Molinaro and van der Laan, 2004; Sinisi and van der Laan, 2004).

2.2 Loss-based parameter definition

Model. Consider a full data structure $X = (W, Y)$, consisting of two components, an (polychotomous or continuous) outcome $Y \in \mathbb{R}$ (i.e., dependent variable, response) and a J–dimensional covariate vector $W = (W(j) : j = 1, \ldots, J) \in \mathbb{R}^J$ (i.e., independent, explanatory, or predictor variables). Assume that the *data generating distribution* P belongs to a *statistical model* $\mathcal{M}$ (i.e., a set of possibly non-parametric distributions): $X \sim P \in \mathcal{M}$.

Parameters. For data generating distributions $P \in \mathcal{M}$, define a *parameter mapping*, $\Psi : \mathcal{M} \to \mathcal{D}(\mathcal{X})$, from the model $\mathcal{M}$ into a space $\mathcal{D}(\mathcal{X})$ of real-valued functions defined on a $(J+1)$–dimensional Euclidean set $\mathcal{X} \subseteq \mathbb{R}^{J+1}$. A *parameter value* (or in short, parameter) is a realization, $\psi \equiv \Psi(P)$, of Ψ for a given $P \in \mathcal{M}$. Thus, the parameter $\psi \in \mathcal{D}(\mathcal{X})$ is a function, $\psi : \mathcal{X} \to \mathbb{R}$, from $\mathcal{X} \subseteq \mathbb{R}^{J+1}$ into the real line $\mathbb{R}$. The *parameter space*, corresponding to the parameter mapping Ψ, is $\mathbf{\Psi} \equiv \{\Psi(P) : P \in \mathcal{M}\} \subseteq \mathcal{D}(\mathcal{X})$.

Loss functions and risk. A *loss function*, $L : (X, \psi) \to L(X, \psi) \in \mathbb{R}$, is a real-valued function of a parameter value $\psi \in \mathbf{\Psi}$ and an observation $X \sim P$.

For a given loss function L, with $\psi \in \mathbf{\Psi}$ and $X \sim P$, the *risk* is the expected value of $L(X, \psi)$ with respect to (w.r.t.) P,

$$\Theta(\psi, P) \equiv \int L(x, \psi) dP(x) = E[L(X, \psi)]. \tag{1}$$

It is assumed that given $X \sim P$ and a parameter value ψ, one can identify a loss function $L(X, \psi)$, so that ψ can be defined as the *risk minimizer* for this loss function, that is,

$$\psi = \Psi(P) \equiv \text{argmin}_{\psi' \in \mathbf{\Psi}} \Theta(\psi', P) = \text{argmin}_{\psi' \in \mathbf{\Psi}} \int L(x, \psi') dP(x). \tag{2}$$

Denote the *optimal risk*, corresponding to the parameter $\psi = \Psi(P)$, by θ, that is,

$$\theta \equiv \Theta(\psi, P) = \min_{\psi' \in \mathbf{\Psi}} \Theta(\psi', P) = \min_{\psi' \in \mathbf{\Psi}} \int L(x, \psi') dP(x). \tag{3}$$

For example, in *regression*, the conditional expected value, $\psi(W) = E[Y|W]$, of an outcome Y given covariates W, is the risk minimizer for the *quadratic loss function* (i.e., L_2 or squared error loss function), $L(X, \psi) = (Y - \psi(W))^2$. In *classification*, the *Bayes classifier*, $\psi(W) = \text{argmax}_y Pr(y|W)$, is the risk minimizer for the *indicator loss function*, $L(X, \psi) = \text{I}(Y \neq \psi(W))$. The optimal risk θ is then referred to as the *Bayes risk*. In *maximum likelihood* and *density estimation*, one uses the *negative log density loss function*, $L(X, \psi) = -\log \psi(X)$ (cf. entropy, Kullback-Leibler divergence). For censored data, van der Laan and Dudoit (2003) apply the estimating function methodology of van der Laan and Robins (2003) to map the full data loss function into a censored data loss function having the same expected value and leading to an efficient estimator of this risk. Unlike the uncensored data loss functions considered in the present article, censored data loss functions depend on a nuisance parameter. Table 1 provides examples of loss functions for a variety of common full data estimation problems.

2.3 Loss-based estimation

Suppose one has available as *learning set* a random sample, $\mathcal{L}_n = \{X_1, \ldots, X_n\}$, of n independent and identically distributed (i.i.d.) random variables, $X_i \sim P \in \mathcal{M}$, $i = 1, \ldots, n$. Our goal is to use the learning set $\mathcal{L}_n$ to estimate a

parameter $\psi = \Psi(P)$ of the unknown data generating distribution P. Let P_n denote the *empirical distribution* of the learning set, which places probability $1/n$ on each X_i, $i = 1, \ldots, n$. An *estimator mapping* $\hat{\Psi}$ is a function from empirical distributions to the parameter space $\mathbf{\Psi}$. An *estimator value* (or in short, estimator) is a realization of this mapping corresponding to a particular empirical distribution P_n and is denoted by $\psi_n \equiv \hat{\Psi}(P_n)$. Note that estimator mappings $\hat{\Psi}$ can be viewed simply as black box algorithms one applies to data, i.e., to empirical distributions P_n.

Given an estimator $\psi_n \in \mathbf{\Psi}$ of a parameter $\psi = \Psi(P)$, the *conditional risk* is defined as the risk of ψ_n with respect to the true, unknown data generating distribution P, that is,

$$\tilde{\theta}_n \equiv \Theta(\psi_n, P) = \int L(x, \psi_n) dP(x). \tag{4}$$

Note that the conditional risk is a *random variable*, as it depends on the data, $X_1, \ldots, X_n$, via the empirical distribution P_n on which ψ_n is based.

A naive risk estimator is the *resubstitution* or *empirical risk estimator*, which replaces the unknown data generating distribution P, in Equation (4), by the known empirical distribution P_n,

$$\bar{\theta}_n \equiv \Theta(\psi_n, P_n) = \int L(x, \psi_n) dP_n(x) = \frac{1}{n} \sum_{i=1}^{n} L(X_i, \psi_n). \tag{5}$$

A corresponding naive estimator for the parameter ψ is the *resubstitution* or *plug-in estimator*, which seeks to minimize the empirical risk,

$$\bar{\psi}_n = \bar{\Psi}(P_n) \equiv \text{argmin}_{\psi' \in \mathbf{\Psi}} \Theta(\psi', P_n) = \text{argmin}_{\psi' \in \mathbf{\Psi}} \sum_{i=1}^{n} L(X_i, \psi'). \tag{6}$$

The special cases of the quadratic and negative log density loss functions correspond, respectively, to *least squares estimation* (LSE) and *maximum likelihood estimation* (MLE). However, unless one considers small enough parameter spaces $\mathbf{\Psi}$ (e.g., corresponding to very specific models $\mathcal{M}$, such as linear models in regression or exponential families in density estimation), minimizing risk over the entire parameter space can result in highly-variable and possibly ill-defined estimators (cf. over-fitting).

Instead, according to **Step 2** of the above road map and other common estimation approaches (e.g., tree-based estimation), one approximates the

parameter space $\mathbf{\Psi}$ by a sequence of K_n subspaces of increasing dimension and generates candidate estimators, $\{\psi_{k,n} = \hat{\Psi}_k(P_n) : k = 1, \ldots, K_n\}$, as the empirical risk minimizers for each subspace (van der Laan et al., 2004b; Molinaro et al., 2004; Molinaro and van der Laan, 2004; Sinisi and van der Laan, 2004). The optimal estimator is then defined as the conditional risk minimizer among the K_n candidates.

Accurate risk estimation is therefore a key aspect of the above estimation approaches for two purposes: for selecting an optimal minimum risk estimator and for assessing the overall performance of this "final" estimator, i.e., its generalization error. Specifically, given a candidate estimator $\psi_n = \hat{\Psi}(P_n)$ of a parameter $\psi = \Psi(P)$, our main task is to derive an accurate estimator of the conditional risk $\tilde{\theta}_n$ of ψ_n, as defined in Equation (4), above. A naive risk estimator is the resubstitution risk estimator $\bar{\theta}_n$ of Equation (5). However, this estimator can be severely *biased downward* due to over-fitting. Cross-validation, described next, provides a general and accurate approach for risk estimation.

2.4 Cross-validation

2.4.1 Cross-validated risk estimation

The main idea in *cross-validation* (CV) is to divide the available learning set into two sets: a *training set* and a *validation set*. Observations in the training set are used to compute (or *train*) the estimator(s) and the validation set is used to assess the risk of (or *validate*) this estimator(s).

To derive a general representation for cross-validation, we introduce a binary random n–vector, or *split vector*, $B_n = (B_n(i) : i = 1, \ldots, n) \in \{0, 1\}^n$, independent of the empirical distribution P_n. A realization of B_n defines a particular split of the learning set of n observations into a training set and a validation set,

$$B_n(i) \equiv \begin{cases} 0, & i\text{th observation } X_i \text{ is in the } \textit{training} \text{ set,} \\ 1, & i\text{th observation } X_i \text{ is in the } \textit{validation} \text{ set.} \end{cases} \tag{7}$$

Let P^0_{n,B_n} and P^1_{n,B_n} denote, respectively, the empirical distributions of the training and validation sets, and let $n_1 \equiv \sum_i B_n(i)$ and $p = p_n \equiv \sum_i B_n(i)/n$ denote, respectively, the number and proportion of observations in the validation sets. A general definition of the *cross-validated risk estimator* for

$\psi_n = \hat{\Psi}(P_n)$ is

$$\begin{aligned} \hat{\theta}_{p_n,n} &\equiv E_{B_n}\Theta(\hat{\Psi}(P^0_{n,B_n}), P^1_{n,B_n}) \\ &= E_{B_n}\int L(x, \hat{\Psi}(P^0_{n,B_n}))dP^1_{n,B_n}(x) \\ &= E_{B_n}\frac{1}{n_1}\sum_{\{i:B_n(i)=1\}} L(X_i, \hat{\Psi}(P_{n,B^0_n})), \end{aligned} \tag{8}$$

where $\hat{\Psi}(P^0_{n,B_n})$ denotes the estimator of the parameter ψ based only on the training set.

The particular distribution of the split vector B_n defines the type of cross-validation procedure. This representation covers a broad class of approaches, including the following.

Leave-one-out cross-validation (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 B_n places mass $1/n$ on each of the n binary vectors, $b_n = (b_n(i) : i = 1, \ldots, n)$, such that $\sum_i b_n(i) = 1$. The proportion of observations in the validation sets is $p_n = 1/n$.

V–fold cross-validation. The learning set is randomly partitioned into V mutually exclusive and exhaustive sets of approximately equal size. Each set is then used in turn as a validation set. The corresponding distribution of B_n places mass $1/V$ on each of V binary vectors, $b^v_n = (b^v_n(i) : i = 1, \ldots, n)$, $v = 1, \ldots, V$, such that $\sum_i b^v_n(i) \approx n/V \ \forall\ v$ and $\sum_v b^v_n(i) = 1 \ \forall\ i$. The proportion of observations in the validation sets is $p_n \approx 1/V$.

Monte Carlo cross-validation. The learning set is repeatedly and randomly divided into two sets, a training set of $n_0 = n(1-p_n)$ observations and a validation set of $n_1 = np_n$ observations. The split vectors B_n are drawn at random with replacement from a distribution that places mass $1/\binom{n}{np_n}$ on each binary vector such that $\sum_i b_n(i) = np_n$.

Bootstrap-based cross-validation. The training sets are based on bootstrap samples of the learning set and the validation sets on the corresponding left-out samples (cf. the *.632 bootstrap estimator*, Efron (1983)). The proportion p_n of observations in the validation sets is a random variable, such that $E[p_n] = E[\sum_i B_n(i)/n] = (1-1/n)^n \approx e^{-1} \approx .368$.

2.4.2 Cross-validated estimator selection

As mentioned above, **Step 3** of our road map (Section 2.1), and many other statistical inference problems, involve *selecting an optimal estimator* (in terms of risk) among a collection of K_n candidate estimators, $\{\psi_{k,n} = \hat{\Psi}_k(P_n) : k = 1, \ldots, K_n\}$, for a parameter $\psi = \Psi(P)$. For instance, k could index the number of variables in a regression model, the number of terminal nodes in a classification or regression tree, or the bandwidth for a kernel density estimator.

Specifically, the estimator selection problem involves choosing a data-adaptive $k_n = K(P_n)$, so that the *risk difference* or *distance*,

$$\Theta(\psi_{k_n,n}, P) - \Theta(\psi, P) = \int (L(x, \psi_{k_n,n}) - L(x, \psi))\, dP(x) \tag{9}$$

converges to zero at asymptotically optimal rate. Ideally, one would like to obtain the *optimal benchmark* or *oracle selector*, $\tilde{k}_n = \tilde{K}(P_n)$, which minimizes this distance, i.e., which minimizes risk with respect to the true data generating distribution P,

$$\tilde{k}_n = \tilde{K}(P_n) \equiv \text{argmin}_{k \in \{1,\ldots,K_n\}} \Theta(\psi_{k,n}, P). \tag{10}$$

However, P is usually unknown and the selection problem involves estimating the *conditional risk*,

$$\tilde{\theta}_n(k) \equiv \Theta(\psi_{k,n}, P) = \int L(x, \psi_{k,n}) dP(x), \tag{11}$$

for each candidate estimator $\psi_{k,n} = \hat{\Psi}_k(P_n)$, $k = 1, \ldots, K_n$. A selector $k_n = K(P_n)$ is said to be *asymptotically equivalent* with the optimal benchmark $\tilde{k}_n$, if the ratio of risk differences with the optimal risk $\theta = \Theta(\psi, P)$ converges to one in probability, i.e.,

$$\frac{\tilde{\theta}_n(k_n) - \theta}{\tilde{\theta}_n(\tilde{k}_n) - \theta} = \frac{\Theta(\psi_{k_n,n}, P) - \Theta(\psi, P)}{\Theta(\psi_{\tilde{k}_n,n}, P) - \Theta(\psi, P)} \xrightarrow{P} 1 \text{ as } n \to \infty. \tag{12}$$

In particular, then k_n is said to be *asymptotically optimal*.

As detailed in Sections 3 and 4, below, cross-validation is a general approach for risk estimation and estimator selection that yields optimal selectors. Specifically, cross-validation provides the following risk estimators for

the candidates $\psi_{k,n}$,

$$\hat{\theta}_{p_n,n}(k) \equiv E_{B_n}\Theta(\hat{\Psi}_k(P^0_{n,B_n}), P^1_{n,B_n}) = E_{B_n}\int L(x, \hat{\Psi}_k(P^0_{n,B_n}))dP^1_{n,B_n}(x). \tag{13}$$

The *cross-validated selector* $\hat{k}_{p_n,n} = \hat{K}_{p_n}(P_n)$ corresponds to the estimator $\psi_{k,n}$ with minimum cross-validated risk,

$$\hat{k}_{p_n,n} = \hat{K}_{p_n}(P_n) \equiv \text{argmin}_{k\in\{1,\ldots,K_n\}}\hat{\theta}_{p_n,n}(k). \tag{14}$$

That is, the cross-validated estimator $\psi_{\hat{k}_{p_n,n},n}$ is chosen to have the best performance on the validation sets.

To obtain a *commensurate optimal benchmark* for the cross-validated selector $\hat{k}_{p_n,n}$, we also define the conditional risks (averaged over B_n) of estimators $\hat{\Psi}_k(P^0_{n,B_n})$, based on cross-validation training sets of size $n(1-p_n)$,

$$\tilde{\theta}_{p_n,n}(k) \equiv E_{B_n}\Theta(\hat{\Psi}_k(P^0_{n,B_n}), P) = E_{B_n}\int L(x, \hat{\Psi}_k(P^0_{n,B_n}))dP(x), \tag{15}$$

and corresponding minimizer,

$$\tilde{k}_{p_n,n} = \tilde{K}_{p_n}(P_n) \equiv \text{argmin}_{k\in\{1,\ldots,K_n\}}\tilde{\theta}_{p_n,n}(k). \tag{16}$$

To simplify notation, as in the proofs of Theorems 1 and 2, we may drop the subscripts n and/or p_n, and use the shorter notation $p = p_n$, $\hat{k} = \hat{k}_{p,n} = \hat{k}_{p_n,n}$, and $\tilde{k} = \tilde{k}_{p,n} = \tilde{k}_{p_n,n}$. When needed, however, we distinguish between $\tilde{k}_{p_n,n}$ and $\tilde{k}_n = \tilde{k}_{0,n}$, the minimizers of the conditional risks $\tilde{\theta}_{p_n,n}(k)$ and $\tilde{\theta}_n(k) = \tilde{\theta}_{0,n}(k)$, for estimators based on training sets of size $n(1-p_n)$ and on the entire learning set of size n (special case $p_n = 0$), respectively (cf. Corollary 1).

The remainder of this article is concerned with deriving optimality results for cross-validation in the context of estimator selection (Section 3) and performance assessment (Section 4).

3 Results for estimator selection

We derive two main results concerning the asymptotic optimality of the cross-validated estimator selectors described in Section 2.4.2. In this context, the

centered conditional risk for the cross-validated selector $\hat{k} = \hat{k}_{p_n,n}$ is compared to the centered conditional risk for the optimal oracle selector $k = \tilde{k}_{p_n,n}$, that is, one compares the two risk differences $\tilde{\theta}_{p_n,n}(\hat{k}) - \theta$ and $\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta$. Finite sample bounds are obtained for the expected value of the *predictive loss*, i.e., the expected value of the difference, $\tilde{\theta}_{p_n,n}(\hat{k}) - \tilde{\theta}_{p_n,n}(\tilde{k})$, between the conditional risks for the cross-validated and oracle selectors. These bounds imply convergence to zero in expectation and in probability of the predictive loss, at rate $O(\log(K_n)/np_n)$ for quadratic loss functions (Theorem 1) and at the slower rate $O(\log(K_n)/\sqrt{np_n})$ for general loss functions (Theorem 2). Consequently, if $E[\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta]$ converges to zero slower than at these rates, then $E[\tilde{\theta}_{p_n,n}(\hat{k}) - \tilde{\theta}_{p_n,n}(\tilde{k})]/E[\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta]$ converges to zero and the ratio of expected risk differences, $E[\tilde{\theta}_{p_n,n}(\hat{k}) - \theta]/E[\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta]$, converges to one. Convergence in probability of the ratio of risk differences follows from Lemma 2, below.

Theorem 1 in van der Laan and Dudoit (2003) provides similar results as the present Theorem 1, for general quadratic loss functions that may depend on a nuisance parameter to handle censored data. The reader is referred to van der Laan et al. (2004a) for a sharper $O(\log(K_n)/np_n)$ bound, as in Theorem 1, for likelihood-based cross-validation.

Note that Theorem 2 applies to general full data loss functions, including the (truncated) absolute error loss function commonly-used for regression with continuous outcomes, indicator and matrix loss functions for classification with polychotomous outcomes, and the negative log density loss function used in density estimation.

Both Theorems 1 and 2 consider general distributions for the split vector B_n, i.e., general cross-validation procedures with an arbitrary proportion p_n of observations included in the validation sets. While the finite sample results hold for any p_n, the asymptotic results are derived under the assumption that the size np_n of the validation sets converges to infinity and hence do not cover leave-one-out cross-validation.

An important and practical issue is the impact of the validation set proportion p_n on risk estimation and estimator selection. In practice, we have found that averaging over split vectors B_n can significantly reduce the sensitivity of the cross-validated selector $\hat{k}_{p_n,n}$ to the choice of p_n, compared to single-split validation (Keleş et al., 2004; van der Laan et al., 2004a). Section 4.4, below, presents a theoretical justification for this behavior in the case of non-quadratic loss functions.

The proofs of Theorems 1, 2, and 3, rely on Bernstein's Inequality, which we state here as Lemma 1 for ease of reference. A proof is given in Györfi et al. (2002a), Lemma A.2, p. 594.

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 $\epsilon > 0$,*

$$Pr\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - E[Z_i]) > \epsilon\right) \leq \exp\left(-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + \epsilon(b-a)/3}\right). \qquad (17)$$

This implies

$$Pr\left(\frac{1}{n}\left|\sum_{i=1}^{n}(Z_i - E[Z_i])\right| > \epsilon\right) \leq 2\exp\left(-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + \epsilon(b-a)/3}\right). \qquad (18)$$

Theorems 1 and 2 first establish *convergence in expectation* of the conditional risk of the cross-validated selector to that of the oracle selector. *Convergence in probability* follows from Lemma 2, below.

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

Proof of Lemma 2. We wish to show that $\forall\ \epsilon > 0$, $\exists\ N$ and $B > 0$, such that $Pr(|Z_n|/g(n) > B) < \epsilon$, $\forall\ n \geq N$. 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 $B = C/\epsilon$ and appealing to Markov's Inequality, one has

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

□

3.1 Quadratic loss function

In this section, we prove that for the full data quadratic (i.e., squared error or L_2) loss function, the ratio of expected risk differences, $E[\tilde{\theta}_{p_n,n}(\hat{k}) -$

$\theta]/E[\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta]$, converges to one at rate $O(\log(K_n)/np_n)$. The proof is based on that of Theorem 7.1, p. 101, in Györfi et al. (2002a), which concerns *single-split* validation (i.e., no averaging over split vectors B_n). However, we modify the proof of Györfi et al. (2002a) do deal with the fact that their 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 results in Theorem 1, below, are more general than the single-split results of Györfi et al. (2002a), as we consider risk estimators based on *multiple random splits* of the learning set based on random vectors B_n. An extra term similar to $T_{1,n}$ is introduced to account for the expected value over B_n. Finally, we point out that a finite sample result similar to that in Equation (7.5) of Györfi et al. (2002a) implies the asymptotic optimality of the cross-validated selector $\hat{k}$ under appropriate conditions.

Theorem 1 *Let $X_1, \ldots, X_n$ be a random sample from a data generating distribution P, where each $X_i = (W_i, Y_i)$ consists of two components, a J–dimensional covariate vector $W_i \in \mathbb{R}^J$ and a univariate outcome $Y_i \in \mathbb{R}$. Let $\{\psi_{k,n} = \hat{\Psi}_k(P_n) : k = 1, \ldots, K_n\}$ denote a sequence of K_n candidate estimators for the conditional mean parameter, $\psi(W) = E[Y|W]$, which is the risk minimizer for the quadratic loss function, $L(X, \psi) = (Y - \psi(W))^2$. Consider the following three risk quantities: the optimal risk θ, corresponding to the parameter of interest ψ,*

$$\theta \equiv \min_{\psi' \in \mathbf{\Psi}} \int L(x, \psi') dP(x),$$

the conditional risk $\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n})$, for the cross-validated selector $\hat{k} = \hat{k}_{p_n,n}$,

$$\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) \equiv E_{B_n} \int L(x, \hat{\Psi}_{\hat{k}_{p_n,n}}(P^0_{n,B_n})) dP(x),$$

and the conditional risk $\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n})$, for the optimal oracle selector $\tilde{k} = \tilde{k}_{p_n,n}$,

$$\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) \equiv \min_{k \in \{1,\ldots,K_n\}} E_{B_n} \int L(x, \hat{\Psi}_k(P^0_{n,B_n})) dP(x).$$

Assumptions. *Suppose that $|Y| \leq M < \infty$ a.s. and $\sup_{W,\psi \in \mathbf{\Psi}} |\psi(W)| \leq M < \infty$ a.s., where the supremum is over a support of the distribution of W.*
Finite sample result. *Let $M_1 \equiv 8M^2$, $M_2 \equiv 16M^2$, and*

$$c(M, \delta) \equiv 2(1 + \delta)^2 \left(\frac{M_1}{3} + \frac{M_2}{\delta} \right). \tag{19}$$

Then, for any $\delta > 0$, one has

$$0 \leq E[\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta] \leq (1+2\delta)E[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta] + 2c(M,\delta)\frac{1+\log(K_n)}{np_n}. \tag{20}$$

Asymptotic results. *The finite sample result in Equation (20) has the following asymptotic implications. If $\frac{\log(K_n)}{(np_n)E[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n})-\theta]} \to 0$ as $n \to \infty$, then*

$$\frac{E[\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta]}{E[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta]} \to 1. \tag{21}$$

Similarly, if $\frac{\log(K_n)}{(np_n)(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n})-\theta)} \xrightarrow{P} 0$ as $n \to \infty$, then

$$\frac{\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta}{\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta} \xrightarrow{P} 1. \tag{22}$$

The proof of Theorem 1 relies on Bernstein's Inequality and the following special property of random variables $Z_k = L(X, \hat{\Psi}_k(P^0_{n,B_n})) - L(X, \Psi(P))$ for the quadratic loss function.

Lemma 3 *Consider the same set-up and assumptions as in Theorem 1. Conditional on the training set empirical distribution P^0_{n,B_n} and split vector B_n, define random variables $Z_k \equiv L(X, \hat{\Psi}_k(P^0_{n,B_n})) - L(X, \Psi(P))$, where L denotes the quadratic loss function, $L(X,\psi) = (Y - \psi(W))^2$. Then,*

$$Var[Z_k | P^0_{n,B_n}, B_n] \leq M_2 E[Z_k | P^0_{n,B_n}, B_n]. \tag{23}$$

Proof of Lemma 3. For the quadratic loss function, $\psi(W) = E[Y|W]$ and

$$Z_k = \left(\Psi(P)(W) - \hat{\Psi}_k(P^0_{n,B_n})(W)\right)\left(2Y - \hat{\Psi}_k(P^0_{n,B_n})(W) - \Psi(P)(W)\right).$$

Thus,

$$\begin{aligned} E[Z_k | P^0_{n,B_n}, B_n] &= E\left[E[Z_k | P^0_{n,B_n}, B_n, W] | P^0_{n,B_n}, B_n\right] \\ &= E\Bigg[\left(\Psi(P)(W) - \hat{\Psi}_k(P^0_{n,B_n})(W)\right) \\ &\qquad \left(2E[Y|W] - \hat{\Psi}_k(P^0_{n,B_n})(W) - \Psi(P)(W)\right) \Bigg| P^0_{n,B_n}, B_n\Bigg] \\ &= E\left[\left(\Psi(P)(W) - \hat{\Psi}_k(P^0_{n,B_n})(W)\right)^2 \Bigg| P^0_{n,B_n}, B_n\right]. \end{aligned}$$

Hence, using the fact that $|2Y - \hat{\Psi}_k(P^0_{n,B_n})(W) - \Psi(P)(W)| \leq 4M$ a.s. and letting $M_2 = (4M)^2$, one has

$$\begin{aligned} Var[Z_k|P^0_{n,B_n}, B_n] &\leq E[Z_k^2|P^0_{n,B_n}, B_n] \\ &\leq (4M)^2 E\left[\left(\Psi(P)(W) - \hat{\Psi}_k(P^0_{n,B_n})(W)\right)^2 \middle| P^0_{n,B_n}, B_n\right] \\ &= M_2 E[Z_k|P^0_{n,B_n}, B_n]. \end{aligned}$$

□

Proof of Theorem 1. Adopt the shorter notation $\hat{k} = \hat{k}_{p_n,n}$ and $\tilde{k} = \tilde{k}_{p_n,n}$.
Finite sample result. We have

$$\begin{aligned} 0 &\leq \tilde{\theta}_{p_n,n}(\hat{k}) - \theta \qquad (24) \\ &= E_{B_n}\int \left(L(x, \hat{\Psi}_{\hat{k}}(P^0_{n,B_n})) - L(x, \Psi(P))\right) dP(x) \\ &\quad - (1+\delta)E_{B_n}\int \left(L(x, \hat{\Psi}_{\hat{k}}(P^0_{n,B_n})) - L(x, \Psi(P))\right) dP^1_{n,B_n}(x) \\ &\quad + (1+\delta)E_{B_n}\int \left(L(x, \hat{\Psi}_{\hat{k}}(P^0_{n,B_n})) - L(x, \Psi(P))\right) dP^1_{n,B_n}(x) \\ &\leq E_{B_n}\int \left(L(x, \hat{\Psi}_{\hat{k}}(P^0_{n,B_n})) - L(x, \Psi(P))\right) dP(x) \\ &\quad - (1+\delta)E_{B_n}\int \left(L(x, \hat{\Psi}_{\hat{k}}(P^0_{n,B_n})) - L(x, \Psi(P))\right) dP^1_{n,B_n}(x) \\ &\quad + (1+\delta)E_{B_n}\int \left(L(x, \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) - L(x, \Psi(P))\right) dP^1_{n,B_n}(x) \\ &= E_{B_n}\int \left(L(x, \hat{\Psi}_{\hat{k}}(P^0_{n,B_n})) - L(x, \Psi(P))\right) dP(x) \\ &\quad - (1+\delta)E_{B_n}\int \left(L(x, \hat{\Psi}_{\hat{k}}(P^0_{n,B_n})) - L(x, \Psi(P))\right) dP^1_{n,B_n}(x) \\ &\quad + (1+\delta)E_{B_n}\int \left(L(x, \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) - L(x, \Psi(P))\right) dP^1_{n,B_n}(x) \\ &\quad - (1+2\delta)E_{B_n}\int \left(L(x, \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) - L(x, \Psi(P))\right) dP(x) \\ &\quad + (1+2\delta)E_{B_n}\int \left(L(x, \hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) - L(x, \Psi(P))\right) dP(x), \end{aligned}$$

where the first inequality follows by definition of the optimal risk θ and the second by definition of $\hat{k}$, such that $\hat{\theta}_{p_n,n}(\hat{k}) \leq \hat{\theta}_{p_n,n}(k)$, $\forall\ k$. Denote the first two terms in the last expression of Equation (24) by $R_{\hat{k},n}$ and the third and fourth terms by $T_{\tilde{k},n}$; the last term is the benchmark risk difference $(1+2\delta)(\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta)$. Hence,

$$0 \leq \tilde{\theta}_{p_n,n}(\hat{k}) - \theta \leq (1+2\delta)(\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta) + R_{\hat{k},n} + T_{\tilde{k},n}. \tag{25}$$

In the sequel, we show that $E[R_{\hat{k},n} + T_{\tilde{k},n}] \leq 2c(M,\delta)(1+\log(K_n))/np_n$. For convenience, introduce the following notation,

$$\begin{aligned}
\tilde{H}_k &\equiv \int \left(L(x, \hat{\Psi}_k(P^0_{n,B_n})) - L(x, \Psi(P)) \right) dP(x), \qquad (26)\\
\hat{H}_k &\equiv \int \left(L(x, \hat{\Psi}_k(P^0_{n,B_n})) - L(x, \Psi(P)) \right) dP^1_{n,B_n}(x),\\
R_{k,n}(B_n) &\equiv (1+\delta)(\tilde{H}_k - \hat{H}_k) - \delta\tilde{H}_k,\\
T_{k,n}(B_n) &\equiv (1+\delta)(\hat{H}_k - \tilde{H}_k) - \delta\tilde{H}_k,
\end{aligned}$$

where, by definition of $\psi = \Psi(P)$ as the risk minimizer, $\tilde{H}_k \geq 0$, $\forall\ k$. One can then rewrite $R_{k,n}$ and $T_{k,n}$ as $R_{k,n} = E_{B_n}[R_{k,n}(B_n)]$ and $T_{k,n} = E_{B_n}[T_{k,n}(B_n)]$, respectively. Note that

$$\begin{aligned}
&Pr(R_{\hat{k},n}(B_n) > s \big| P^0_{n,B_n}, B_n)\\
&\quad = Pr\left(\tilde{H}_{\hat{k}} - \hat{H}_{\hat{k}} > \frac{1}{1+\delta}(s + \delta\tilde{H}_{\hat{k}}) \middle| P^0_{n,B_n}, B_n \right)\\
&\quad \leq K_n \max_{k\in\{1,\ldots,K_n\}} Pr\left(\tilde{H}_k - \hat{H}_k > \frac{1}{1+\delta}(s + \delta\tilde{H}_k) \middle| P^0_{n,B_n}, B_n \right).
\end{aligned}$$

Conditional on P^0_{n,B_n} and B_n, consider the random variables

$$Z_k \equiv L(X, \hat{\Psi}_k(P^0_{n,B_n})) - L(X, \Psi(P)).$$

Let $Z_{k,i}$, $i = 1, \ldots, np_n$, denote the np_n i.i.d. copies of Z_k corresponding with the validation set, i.e., with $\{X_i : B_n(i) = 1\}$. Note that $\hat{H}_k = \sum_{i=1}^{np_n} Z_{k,i}/np_n$ and $\tilde{H}_k = E[Z_k|P^0_{n,B_n}, B_n]$, so that $\tilde{H}_k - \hat{H}_k = E[Z_k|P^0_{n,B_n}, B_n] - \sum_{i=1}^{np_n} Z_{k,i}/np_n$ represents an empirical mean of np_n centered i.i.d. random variables. For the quadratic loss function, the random variables Z_k are bounded, with $|Z_k| \leq 4M^2$ a.s.

Next, we apply Bernstein's Inequality (Lemma 1) to the centered empirical mean $\tilde{H}_k - \hat{H}_k$ and exploit the special property of Z_k derived in Lemma 3 for the quadratic loss function, to obtain an $\exp(-(np_n)s/c)$ bound for tail probabilities of $R_{k,n}(B_n)$ and $T_{k,n}(B_n)$, instead of the usual $\exp(-(np_n)s^2/(a+bs))$, for some constants a, b, $c < \infty$. This shows that the risk differences converge at $O(\log(K_n)/np_n)$ rate instead of the slower $O(\log(K_n)/\sqrt{np_n})$ rate derived in Theorem 2, below, for general loss functions. Specifically, from Lemma 3,

$$\sigma_k^2 \equiv Var[Z_k|P^0_{n,B_n}, B_n] \leq M_2 E[Z_k|P^0_{n,B_n}, B_n] = M_2\tilde{H}_k.$$

For $s > 0$ and $M_1 = 8M^2$, Bernstein's Inequality then yields

$$\begin{aligned} &Pr(R_{k,n}(B_n) > s\big|P^0_{n,B_n}, B_n) \\ &\quad = \; Pr\left(\tilde{H}_k - \hat{H}_k > \frac{1}{1+\delta}(s + \delta\tilde{H}_k)\middle|P^0_{n,B_n}, B_n\right) \\ &\quad \leq \; Pr\left(\tilde{H}_k - \hat{H}_k > \frac{1}{1+\delta}(s + \delta\sigma_k^2/M_2)\middle|P^0_{n,B_n}, B_n\right) \\ &\quad \leq \; \exp\left(-\frac{np_n}{2(1+\delta)^2}\frac{(s+\delta\sigma_k^2/M_2)^2}{\sigma_k^2 + \frac{M_1}{3(1+\delta)}(s+\delta\sigma_k^2/M_2)}\right). \end{aligned}$$

Note that

$$\frac{(s+\delta\sigma_k^2/M_2)^2}{\sigma_k^2 + \frac{M_1}{3(1+\delta)}(s+\delta\sigma_k^2/M_2)} = \frac{(s+\delta\sigma_k^2/M_2)}{\frac{\sigma_k^2}{s+\delta\sigma_k^2/M_2} + \frac{M_1}{3(1+\delta)}} \geq \frac{(s+\delta\sigma_k^2/M_2)}{\frac{M_2}{\delta} + \frac{M_1}{3}} \geq \frac{s}{\frac{M_2}{\delta} + \frac{M_1}{3}}.$$

This shows that, for $s > 0$,

$$Pr(R_{\hat{k},n}(B_n) > s|P^0_{n,B_n}, B_n) \leq K_n \exp\left(-\frac{np_n}{c(M,\delta)}s\right),$$

where $c(M,\delta) = 2(1+\delta)^2\left(\frac{M_1}{3} + \frac{M_2}{\delta}\right)$, with $M_1 = 8M^2$ and $M_2 = 16M^2$. The same bound applies to the marginal probabilities $Pr(R_{\hat{k},n}(B_n) > s)$.

For each $u > 0$, we have

$$E[R_{\hat{k},n}] \leq u + \int_u^\infty K_n \exp\left(-\frac{np_n}{c(M,\delta)}s\right) ds.$$

The function of u on the right-hand side of the above inequality achieves a minimum value of $c(M,\delta)(1+\log(K_n))/(np_n)$ at $u_n = c(M,\delta)\log(K_n)/(np_n)$. Thus,

$$E[R_{\hat{k},n}] \leq c(M,\delta)\frac{1+\log(K_n)}{np_n}.$$

The same bound applies to $E[T_{\tilde{k},n}]$. Taking the expected values of the quantities in Equation (25) yields the finite sample result in Equation (20).

Asymptotic results. Convergence to one of the ratio of expected risk differences, in Equation (21), follows trivially from the finite sample result of Equation (20). Convergence in probability, as in Equation (22), follows from convergence in expectation by Lemma 2.

□

Theorem 1 provides a finite sample bound, $2c(M,\delta)(1+\log(K_n))/(np_n)$, for comparing the performance of the cross-validated selector $\hat{k}_{p_n,n}$ to that of the commensurate optimal oracle selector $\tilde{k}_{p_n,n}$ based on $n(1-p_n)$ training observations. However, one would like the cross-validated selector $\hat{k}_{p_n,n}$ to perform as well as an oracle selector $\tilde{k}_n = \tilde{k}_{0,n}$ based on the *entire learning set of size* n, rather than smaller training sets of size $n(1-p_n)$ as above. The following is a corollary of Theorem 1, which relates the conditional risk of the cross-validated selector, $\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n})$, to that of the oracle selector based on n observations, $\tilde{\theta}_n(\tilde{k}_n)$.

Corollary 1 *Denote the conditional risk of an estimator $\hat{\Psi}_k(P_n)$, based on the entire learning set of size n, by*

$$\tilde{\theta}_n(k) \equiv \int L(x, \hat{\Psi}_k(P_n))dP(x)$$

and let

$$\tilde{k}_n \equiv argmin_{k\in\{1,\ldots,K_n\}}\tilde{\theta}_n(k)$$

denote the corresponding risk minimizer. As before, let $\tilde{k}_{p_n,n}$ denote the minimizer for the conditional risk $\tilde{\theta}_{p_n,n}(k)$ of estimators $\hat{\Psi}_k(P^0_{n,B_n})$, based on cross-validation training sets of size $n(1-p_n)$.

Assumptions. *As in Theorem 1, suppose that $|Y| \leq M < \infty$ a.s. and $\sup_{W,\psi\in\mathbf{\Psi}}|\psi(W)| \leq M < \infty$ a.s., where the supremum is over a support of the*

distribution of W. *Further assume that, as* $n \to \infty$, $p_n \to 0$, $\log(K_n)/(np_n) \to 0$, $\frac{\log(K_n)}{(np_n)(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n})-\theta)} \xrightarrow{P} 0$, *and*

$$\frac{\tilde{\theta}_n(\tilde{k}_n) - \theta}{\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta} \xrightarrow{P} 1. \tag{27}$$

Asymptotic result. *Then,*

$$\frac{\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta}{\tilde{\theta}_n(\tilde{k}_n) - \theta} \xrightarrow{P} 1. \tag{28}$$

Sufficient condition. *A sufficient condition for Equation (27) is that there exists* $\gamma > 0$ *such that*

$$\left(n^\gamma \left(\tilde{\theta}_n(\tilde{k}_n) - \theta\right),\ (n(1-p_n))^\gamma \left(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta\right)\right) \xRightarrow{\mathcal{L}} (Z, Z), \tag{29}$$

for a random variable Z *with* $Pr(Z > a) = 1$ *for some* $a > 0$. *In particular, for single-split validation, where* $Pr(B_n = b) = 1$ *for some* $b \in \{0,1\}^n$, *it suffices to assume that there exists* $\gamma > 0$ *such that* $n^\gamma \left(\tilde{\theta}_n(\tilde{k}_n) - \theta\right) \xRightarrow{\mathcal{L}} Z$, *for a random variable* Z *with* $Pr(Z > a) = 1$ *for some* $a > 0$.

Proof of Corollary 1.
Asymptotic result. The main statement of the corollary, in Equation (28), follows immediately from Theorem 1 by noting that

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

Sufficient condition. We now show that the assumption in Equation (27) holds under the first sufficient condition in Equation (29). Define $Z_{1,n} \equiv n^\gamma(\tilde{\theta}_n(\tilde{k}_n) - \theta)$ and $Z_{2,n} \equiv (n(1-p_n))^\gamma(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta)$. If $(Z_{1,n}, Z_{2,n}) \xRightarrow{\mathcal{L}} (Z, Z)$, then, by the Continuous Mapping Theorem, $Z_{1,n}/Z_{2,n} \xRightarrow{\mathcal{L}} 1$. However, note that

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

which yields Equation (27) if $p_n \to 0$. For single-split validation, i.e., $Pr(B_n = b) = 1$ for some $b \in \{0,1\}^n$, then $Z_{1,n} \stackrel{\mathcal{L}}{=} Z_{2,\frac{n}{1-p_n}}$, and hence $Z_{1,n} \xRightarrow{\mathcal{L}} Z$ implies

$(Z_{1,n}, Z_{2,n}) \stackrel{\mathcal{L}}{\Rightarrow} (Z, Z)$. This completes the proof of Corollary 1.

□

Previous results for the quadratic loss function. In Corollaries 7.1 – 7.3, Györfi et al. (2002a) apply their Theorem 7.1 to kernel, partitioning, and nearest neighbor estimators, respectively. Finite sample bounds are derived for the expected risk difference $E[\tilde{\theta}_{p_n,n}(\hat{k}) - \theta]$ of the single-split selector $\hat{k}$, based on previously derived estimator-specific bounds for the benchmark $E[\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta]$.

Devroye et al. (2003) establish rate of convergence results for estimators of the optimal risk θ for the quadratic loss function. Building on results in Antos et al. (1999), they show that one cannot estimate θ with guaranteed rate of convergence (Theorem 2.1). However, by imposing certain conditions on the data generating distribution P, such as Lipschitz continuity of the regression function $\psi(W) = E[Y|W]$, they derive non-trivial rates of convergence for two classes of estimators, based on single-split validation (Theorem 3.1) and first nearest neighbor cross-validation (Theorem 4.1).

3.2 General loss function

In this section, we derive an analog of Theorem 1 for general full data loss functions, whereby the ratio of expected risk differences, $E[\tilde{\theta}_{p_n,n}(\hat{k}) - \theta]/E[\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta]$, is shown to converge to one at rate $O(\log(K_n)/\sqrt{np_n})$ rather than the faster $O(\log(K_n)/np_n)$ applicable to quadratic loss functions.

Theorem 2 *Let $X_1, \ldots, X_n$ be a random sample from a data generating distribution P. Let $\{\psi_{k,n} = \hat{\Psi}_k(P_n) : k = 1, \ldots, K_n\}$ denote a sequence of K_n candidate estimators for the parameter $\psi = \Psi(P)$, where ψ is a risk minimizer for the loss function $L(X, \psi)$. Consider the same three risk quantities, θ, $\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n})$, and $\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n})$, as in Theorem 1.*
Assumptions. *Suppose that $\sup_{X,\psi\in\Psi} L(X, \psi) \leq M < \infty$ a.s., where the supremum is over a support of the distribution of X.*
Finite sample result. *Let $m \equiv 2M$ and $v \equiv M^2$, and define*

$$f(M, K_n, np_n) \equiv 2\left[u_n + \int_{u_n}^{\infty} K_n \exp\left(-\frac{1}{2}\frac{(np_n)x^2}{v + mx/3}\right) dx\right], \qquad (30)$$

where

$$u_n \equiv \frac{m\log(K_n)/3 + \sqrt{(m\log(K_n)/3)^2 + 2(np_n)v\log(K_n)}}{np_n}. \tag{31}$$

Then, one has the following finite sample result

$$0 \le E[\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta] \le E[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta] + f(M, K_n, np_n). \tag{32}$$

Asymptotic results. *Suppose that* $\log(K_n)/\sqrt{np_n} \to 0$ *as* $n \to \infty$. *Then,* $f(M, K_n, np_n) = O(\log(K_n)/\sqrt{np_n})$ *and hence* $E[\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta] = E[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta] + O(\log(K_n)/\sqrt{np_n})$. *In particular,* $\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta = \tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta + O_P(\log(K_n)/\sqrt{np_n})$. *If* $\frac{\log(K_n)}{\sqrt{np_n}E[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta]} \to 0$ *as* $n \to \infty$, *then*

$$\frac{E[\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta]}{E[\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta]} \to 1. \tag{33}$$

Similarly, if $\frac{\log(K_n)}{\sqrt{np_n}(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta)} \xrightarrow{P} 0$ *as* $n \to \infty$, *then*

$$\frac{\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta}{\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta} \xrightarrow{P} 1. \tag{34}$$

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

Proof of Theorem 2. As in the proof of Theorem 1, adopt the shorter notation $\hat{k} = \hat{k}_{p_n,n}$ and $\tilde{k} = \tilde{k}_{p_n,n}$.

Finite sample result. We have

$$\begin{aligned} 0 &\leq \tilde{\theta}_{p_n,n}(\hat{k}) - \theta \qquad (35)\\ &= E_{B_n}\int \left(L(x,\hat{\Psi}_{\hat{k}}(P^0_{n,B_n})) - L(x,\Psi(P))\right) dP(x)\\ &\quad - E_{B_n}\int \left(L(x,\hat{\Psi}_{\hat{k}}(P^0_{n,B_n})) - L(x,\Psi(P))\right) dP^1_{n,B_n}(x)\\ &\quad + E_{B_n}\int \left(L(x,\hat{\Psi}_{\hat{k}}(P^0_{n,B_n})) - L(x,\Psi(P))\right) dP^1_{n,B_n}(x)\\ &\leq E_{B_n}\int \left(L(x,\hat{\Psi}_{\hat{k}}(P^0_{n,B_n})) - L(x,\Psi(P))\right) d(P - P^1_{n,B_n})(x)\\ &\quad + E_{B_n}\int \left(L(x,\hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) - L(x,\Psi(P))\right) d(P^1_{n,B_n} - P)(x)\\ &\quad + E_{B_n}\int \left(L(x,\hat{\Psi}_{\tilde{k}}(P^0_{n,B_n})) - L(x,\Psi(P))\right) dP(x), \end{aligned}$$

where the first inequality follows by definition of the optimal risk θ and the second by definition of $\hat{k}$, such that $\hat{\theta}_{p_n,n}(\hat{k}) \leq \hat{\theta}_{p_n,n}(k)$, $\forall\ k$. For convenience, introduce the following notation,

$$\begin{aligned} \tilde{H}_k &\equiv \int \left(L(x,\hat{\Psi}_k(P^0_{n,B_n})) - L(x,\Psi(P))\right) dP(x), \qquad (36)\\ \hat{H}_k &\equiv \int \left(L(x,\hat{\Psi}_k(P^0_{n,B_n})) - L(x,\Psi(P))\right) dP^1_{n,B_n}(x),\\ T_{k,n}(B_n) &\equiv \tilde{H}_k - \hat{H}_k,\\ T_{k,n} &\equiv E_{B_n}[T_{k,n}(B_n)]. \end{aligned}$$

Then, the first two terms in the last expression of Equation (35) are $T_{\hat{k},n}$ and $-T_{\tilde{k},n}$, respectively; the last term is the benchmark risk difference $(\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta)$. Hence,

$$0 \leq \tilde{\theta}_{p_n,n}(\hat{k}) - \theta \leq \tilde{\theta}_{p_n,n}(\tilde{k}) - \theta + T_{\hat{k},n} - T_{\tilde{k},n}. \qquad (37)$$

In the sequel, we show that both $|ET_{\hat{k},n}|$ and $|ET_{\tilde{k},n}|$ are bounded by $f(M,K_n,np_n)/2$, so that $E[T_{\hat{k},n} - T_{\tilde{k},n}] \leq f(M,K_n,np_n)$. Note that

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

Conditional on P^0_{n,B_n} and B_n, consider the random variables

$$Z_k \equiv L(X, \hat{\Psi}_k(P^0_{n,B_n})) - L(X, \Psi(P)).$$

Let $Z_{k,i}$, $i = 1, \ldots, np_n$, denote the np_n i.i.d. copies of Z_k corresponding with the validation set, i.e., with $\{X_i : B_n(i) = 1\}$. Note that $\hat{H}_k = \sum_{i=1}^{np_n} Z_{k,i}/np_n$ and $\tilde{H}_k = E[Z_k|P^0_{n,B_n}, B_n]$, so that $\tilde{H}_k - \hat{H}_k = E[Z_k|P^0_{n,B_n}, B_n] - \sum_{i=1}^{np_n} Z_{k,i}/np_n$ represents an empirical mean of np_n centered i.i.d. random variables. The random variables Z_k are bounded, with $|Z_k| \leq M$ a.s. and $Var[Z_k|P^0_{n,B_n}, B_n] \leq v = M^2$. Thus, from Bernstein's Inequality (Lemma 1), for $s > 0$,

$$Pr\left(\tilde{H}_k - \hat{H}_k > s \middle| P^0_{n,B_n}, B_n\right) \leq \exp\left(-\frac{1}{2}\frac{(np_n)s^2}{v + ms/3}\right),$$

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

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

The same bound applies to the marginal probabilities $Pr(T_{\hat{k},n}(B_n) > s)$.

Now note that, for any random variable Z,

$$E[Z] \leq E[\mathrm{I}(Z > 0)Z] = \int_0^\infty Pr(Z > z)dz.$$

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

$$\begin{aligned} E[T_{\hat{k},n}] &= E[E_{B_n}[T_{\hat{k},n}(B_n)]] \leq \int_0^\infty Pr(T_{\hat{k},n}(B_n) > x)dx \qquad (38) \\ &\leq u + \int_u^\infty K_n \exp\left(-\frac{1}{2}\frac{(np_n)x^2}{v + mx/3}\right)dx. \end{aligned}$$

The quantity u_n in Equation (31) corresponds with the minimizer of the function of u on the right-hand side of Equation (38). In particular, u_n is a solution of the equation obtained by setting the derivative of this term with respect to u equal to zero. The same bound can be derived for $-E[T_{\hat{k},n}]$, by applying Bernstein's Inequality to $\hat{H}_k - \tilde{H}_k$. Hence, $|ET_{\hat{k},n}| \leq f(M, K_n, np_n)/2$. A similar proof as above shows that $|ET_{\tilde{k},n}| \leq f(M, K_n, np_n)/2$. Thus, taking the expected values of the quantities in Equation (37) yields the finite sample result

$$0 \leq E[\tilde{\theta}_{p_n,n}(\hat{k}) - \theta] \leq E[\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta] + f(M, K_n, np_n),$$

where

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

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

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

The integrand in the definition of $f(M, K_n, np_n)$, in Equation (30), is a decreasing function of x for $x > 0$, which achieves a value of one at $x = u_n$ (by definition of u_n) and tends to zero as x approaches ∞. Hence, for $y > (\sqrt{np_n}/\log(K_n))u_n$, the integrand in the last expression is bounded above by one. Since $u_n = O(\sqrt{\log(K_n)/np_n})$, then $\exists\ N_1 > 0$ and some constant $1 < A < \infty$, such that $(\sqrt{np_n}/\log(K_n))u_n \leq A$, $\forall\ n \geq N_1$. Thus, for $n \geq N_1$,

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

Consider now the second term in the above expression. Since $\log(K_n)/\sqrt{np_n} \to 0$ as $n \to \infty$, then $\exists\ N_2 > 0$ such that $m\log(K_n)/3\sqrt{np_n} < \epsilon$, $\forall\ n \geq N_2$.

Hence, for $n \geq N_2$ and $1 < A < y$,

$$\frac{y^2}{v + \frac{m\log(K_n)}{3\sqrt{np_n}}y} > \frac{y^2}{v + \epsilon y} > \frac{y}{v+\epsilon}.$$

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

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

The above expression will be $O(\log(K_n)/\sqrt{np_n})$, 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 $K_n > K_{n'} \geq 1$ and note that $0 < g(A) < g(y)$ for $y > A$. Then,

$$\begin{aligned}
\frac{K_{n'}^{1-\frac{1}{2}g(y)\log(K_{n'})}}{K_n^{1-\frac{1}{2}g(y)\log(K_n)}} &\geq \left(\frac{K_{n'}}{K_n}\right)^{1-\frac{1}{2}g(y)\log(K_n)} \\
&\geq \left(\frac{K_n}{K_{n'}}\right)^{\frac{1}{2}g(A)\log(K_n)-1},
\end{aligned}$$

which is greater than one for each $y > A$, when A is chosen so that $\frac{1}{2}g(A)\log(K_0) > 1$ for a constant $K_0 < K_n$. It remains to show that the integral is finite for some K_n, which is immediate from

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

Thus, $f(M, K_n, np_n) = O(\log(K_n)/\sqrt{np_n})$. By definition of $\tilde{k}$, $\tilde{\theta}_{p_n,n}(\tilde{k}) \leq \tilde{\theta}_{p_n,n}(\hat{k})$, and from the finite sample result in Equation (32), it follows that

$$E[\tilde{\theta}_{p_n,n}(\hat{k}) - \theta] = E[\tilde{\theta}_{p_n,n}(\tilde{k}) - \theta] + O\left(\frac{\log(K_n)}{\sqrt{np_n}}\right).$$

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

□

Bound for $f(M, K_n, np_n)$. We now present a simpler bound for $f(M, K_n, np_n)$. Let $c \equiv 2(v + m/3)$ 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 $(np_n)x^2/(v+mx/3) \geq (np_n)x/(v+m/3)$ for $x \geq 1$ and $(np_n)x^2/(v+mx/3) \geq (np_n)x^2/(v+m/3)$ for $x \leq 1$, $f(M, K_n, np_n)$ can be bounded by the following analytical expression,

$$\begin{aligned} f(M, K_n, np_n) &\leq 2\Bigg[u_n + K_n \frac{c}{np_n} \exp\left(-\frac{np_n}{c} \max(1, u_n)\right) \\ &+ \mathrm{I}(u_n \leq 1) K_n \sqrt{\frac{c\pi}{np_n}} \left(\Phi\left(\sqrt{\frac{2np_n}{c}}\right) - \Phi\left(\sqrt{\frac{2np_n}{c}} u_n\right)\right)\Bigg]. \end{aligned} \tag{39}$$

Although we are concerned with risk estimation for the purpose of estimator selection for *arbitrary* loss functions and estimators $\psi_{k,n}$, we refer below to previous work on risk estimation for specific classes of loss functions and estimators.

Previous results for the indicator loss function. The special case of the indicator loss function used in classification is treated in detail in Devroye et al. (1996), Chapters 8, 22, 23, 24, and 31. For binary classification, Antos et al. (1999) show that the Bayes risk (i.e., the optimal risk θ for the indicator loss function) cannot be estimated with guaranteed rate of convergence. Specifically, for any estimator θ_n of the Bayes risk θ and any sequence of positive numbers $\{a_n\}$ converging to zero, one can find a data generating distribution P such that $E|\theta_n - \theta| \geq a_n$ infinitely often, i.e., $\forall\ m > 0$, $\exists\ n \geq m$, such that $E|\theta_n - \theta| \geq a_n$.

Previous results for the L_1 and L_2 norms. Györfi et al. (2002b) derive relative stability results for certain classes of density and regression estimators. Given an estimator ψ_n of the parameter ψ, the estimation error is measured by $\|\psi_n - \psi\|$, where $\|\cdot\|$ is the L_1 norm in density estimation and the L_2 norm in regression. For histogram and kernel density estimators,

Györfi et al. (2002b) show that the L_1 error $\|\psi_n - \psi\|$ is *relatively stable*, in the sense that the ratio $\|\psi_n - \psi\|/E[\|\psi_n - \psi\|]$ converges to one in probability. Similar results are derived for the L_2 error of partitioning, kernel, and nearest neighbor regression estimators.

4 Results for performance assessment

4.1 Asymptotic linearity of the cross-validated risk estimator

We first derive a consistency and asymptotic linearity result for the cross-validated estimator $\hat{\theta}_{p_n,n}$ of the conditional risk $\tilde{\theta}_{p_n,n}$ for estimators $\hat{\Psi}(P^0_{n,B_n})$ based on cross-validation training sets of size $n(1-p_n)$.

Theorem 3 *Let $X_1, \ldots, X_n$ be a random sample from a data generating distribution P. Let $\psi_n = \hat{\Psi}(P_n)$ denote an estimator for the parameter $\psi = \Psi(P)$, where ψ is a risk minimizer for the loss function $L(X, \psi)$. Consider the following three risk quantities: the optimal risk θ, corresponding to the parameter of interest ψ,*

$$\theta \equiv \min_{\psi' \in \mathbf{\Psi}} \int L(x, \psi') dP(x),$$

the conditional risk $\tilde{\theta}_{p_n,n}$, for the estimator mapping $\hat{\Psi}$ applied to cross-validation training sets of size $n(1-p_n)$,

$$\tilde{\theta}_{p_n,n} \equiv E_{B_n} \int L(x, \hat{\Psi}(P^0_{n,B_n})) dP(x),$$

and the cross-validated risk estimator,

$$\hat{\theta}_{p_n,n} \equiv E_{B_n} \int L(x, \hat{\Psi}(P^0_{n,B_n})) dP^1_{n,B_n}(x).$$

Assumptions. *Suppose that $\sup_{X,\psi \in \mathbf{\Psi}} L(X, \psi) \leq M < \infty$ a.s., where the supremum is over a support of the distribution of X, $1/(p_n\sqrt{n}) \to 0$, and*

$$\frac{1}{\sqrt{p_n}} E_{B_n} \sqrt{\int \left(L(x, \hat{\Psi}(P^0_{n,B_n})) - L(x, \Psi(P)) \right)^2 dP(x)} = o_P(1). \qquad (40)$$

Asymptotic linearity result. *Then,*

$$\hat{\theta}_{p_n,n} - \tilde{\theta}_{p_n,n} = \frac{1}{n}\sum_{i=1}^{n}\left(L(X_i, \Psi(P)) - \theta\right) + o_P(1/\sqrt{n}). \tag{41}$$

Proof of Theorem 3. We have

$$\begin{aligned}
&\hat{\theta}_{p_n,n} - \tilde{\theta}_{p_n,n} \\
&\quad = E_{B_n}\int \left(L(x, \hat{\Psi}(P^0_{n,B_n})) - L(x, \Psi(P))\right) d(P^1_{n,B_n} - P)(x) \\
&\qquad + E_{B_n}\int L(x, \Psi(P)) d(P^1_{n,B_n} - P)(x).
\end{aligned}$$

By virtue of the expected value w.r.t. the split vector B_n, the second term equals

$$\int L(x, \Psi(P)) d(P_n - P)(x) = \frac{1}{n}\sum_{i=1}^{n}(L(X_i, \Psi(P)) - \theta).$$

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

$$Z_n \equiv L(X, \hat{\Psi}(P^0_{n,B_n})) - L(X, \Psi(P)),$$

and let $Z_{n,i}$, $i = 1, \ldots, np_n$, denote the np_n i.i.d. copies of Z_n corresponding with the validation set, i.e., with $\{X_i : B_n(i) = 1\}$. Then, $T_n(B_n)$ can be written as an empirical mean, $\sum_{i=1}^{np_n} Z_{n,i}/np_n - E[Z_n|P^0_{n,B_n}, B_n]$, of np_n centered random variables, with $|Z_n| < M$ a.s. Let

$$\sigma_n^2(B_n) \equiv \max\left\{(np_n)^{-1}, E[Z_n^2|P^0_{n,B_n}, B_n]\right\}.$$

Then, $Var[Z_n|P^0_{n,B_n}, B_n] \leq \sigma_n^2(B_n)$. From Bernstein's Inequality, with $W = 2M$,

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

Thus,

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

where we carried out the substitution $x = (\sigma_n(B_n)/\sqrt{np_n})y$. Since $\sigma_n(B_n) \geq 1/\sqrt{np_n}$, 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 the assumption in Equation (40),

$$\frac{1}{\sqrt{p_n}} E_{B_n} \sqrt{E[Z_n^2 | P^0_{n,B_n}, B_n]} = o_P(1),$$

and since $1/(p_n\sqrt{n}) \to 0$, then

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

□

Note that by Jensen's Inequality, applied to the concave square root function, a sufficient condition for the assumption in Equation (40) is

$$\frac{1}{p_n} E_{B_n} \int \left(L(x, \hat{\Psi}(P^0_{n,B_n})) - L(x, \Psi(P))\right)^2 dP(x) = o_P(1). \tag{42}$$

The argument in Section 4.4, concerning the impact of the validation set proportion p_n for non-quadratic loss functions, suggests that the risk difference $\tilde{\theta}_{p_n,n} - \tilde{\theta}_n$ is zero in first order. Thus, by virtue of the expectation w.r.t. to B_n, one expects the cross-validated risk estimator $\hat{\theta}_{p_n,n}$ to also be a decent approximation of the conditional risk $\tilde{\theta}_n$, for a fixed validation proportion $p_n \in (0, 1)$.

4.2 Asymptotic linearity of the resubstitution risk estimator

The previous asymptotic linearity result for the cross-validated risk estimator $\hat{\theta}_{p_n,n}$ also applies to the resubstitution estimator $\bar{\theta}_n$, but under different conditions. In addition, 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 *Let $X_1, \ldots, X_n$ be a random sample from a data generating distribution P. Let $\psi_n = \hat{\Psi}(P_n)$ denote an estimator for the parameter $\psi = \Psi(P)$, where ψ is a risk minimizer for the loss function $L(X, \psi)$. Consider the following three risk quantities: the optimal risk θ, corresponding to the parameter of interest ψ,*

$$\theta \equiv \min_{\psi' \in \mathbf{\Psi}} \int L(x, \psi') dP(x),$$

the conditional risk $\tilde{\theta}_n$, for the estimator mapping $\hat{\Psi}$ applied to the entire learning set of size n,

$$\tilde{\theta}_n \equiv \int L(x, \hat{\Psi}(P_n)) dP(x),$$

and the resubstitution risk estimator,

$$\bar{\theta}_n \equiv \int L(x, \hat{\Psi}(P_n)) dP_n(x).$$

Assumptions. *Suppose $\mathcal{C}$ is a class of functions of X so that $Pr(\psi_n \in \mathcal{C}) \to 1$, $\int (L(x, \psi_n) - L(x, \psi))^2 dP(x) = o_P(1)$, and $\mathcal{F} \equiv \{x \to L(x, \psi') - L(x, \psi) : \psi' \in \mathcal{C}\}$ is a P-Donsker class.*
Asymptotic linearity result. *Then,*

$$\bar{\theta}_n - \tilde{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} (L(X_i, \Psi(P)) - \theta) + o_P(1/\sqrt{n}). \tag{43}$$

Proof of Theorem 4. We have

$$\begin{aligned} \bar{\theta}_n - \tilde{\theta}_n &= \int (L(x, \psi_n) - L(x, \psi))\, d(P_n - P)(x) \\ &\quad + \int L(x, \psi) d(P_n - P)(x). \end{aligned} \tag{44}$$

The second term is the desired linear component, $\sum_{i=1}^{n} (L(X_i, \psi) - \theta)/n$. 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

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

indexed by $f \in \mathcal{F}$, where $\mathcal{F}$ is the assumed P-Donsker class, $\mathcal{F} = \{x \to L(x, \psi') - L(x, \psi) : \psi' \in \mathcal{C}\}$. 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 $f_n(X) \equiv L(X, \psi_n) - L(X, \psi)$. By assumption, we have

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

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 that

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

hence, as required, the first term in Equation (44) is indeed $o_P(1/\sqrt{n})$.

□

Previous results for the indicator loss function. Györfi and Horváth (1998) and Pintér (2002) provide asymptotic results for the resubstitution risk estimator in the special case of the indicator loss function for binary partitioning classification rules. In particular, Györfi and Horváth (1998) show that the difference between the resubstitution risk estimator and the Bayes risk restricted to the partition (i.e., conditional on the partition induced by the classifier, where the Bayes rule is applied within each set in the partition) is asymptotically normal. For rectangular partitions, Pintér (2002) establishes asymptotic normality of the difference between the resubstitution risk estimator $\bar{\theta}_n$ and the conditional risk $\tilde{\theta}_n$ (Theorem 2). She also proves that the difference between the expected value of the resubstitution risk estimator and the Bayes risk restricted to the partition converges to zero at rate faster than $O(1/\sqrt{n})$ (Theorem 3). Similar results are provided for leave-one-out cross-validation.

In spite of the good asymptotic behavior of resubstitution risk estimators under the conditions of Theorem 4 and in the above articles, a number of caveats are in order. Firstly, resubstitution risk 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 analog in Theorem 3 for the cross-validated risk estimator. Finally, resubstitution estimators tend to perform poorly in model selection due to over-fitting.

4.3 Risk confidence intervals

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

$$IC(X|P) \equiv L(X, \Psi(P)) - \theta.$$

Then, $E[IC(X|P)] = 0$ and $\sigma^2 = Var[IC(X|P)] = \int IC^2(x|P)dP(x)$. From the Central Limit Theorem and Theorem 3, as $n \to \infty$, $\sqrt{n}(\hat{\theta}_{p_n,n} - \tilde{\theta}_{p_n,n})/\sigma$ converges in distribution to a standard normal random variable. Consider the following resubstitution estimators for $IC(X|P)$ and its variance σ^2,

$$IC(X|P_n) \equiv L(X, \hat{\Psi}(P_n)) - \bar{\theta}_n \qquad \text{and} \qquad \sigma_n^2 \equiv \int IC^2(x|P_n)dP_n(x).$$

Then, an approximate *asymptotic* $(1-\alpha)100\%$ *confidence interval* for the conditional risk $\tilde{\theta}_{p_n,n}$ is given by

$$\hat{\theta}_{p_n,n} \pm z_{1-\alpha/2}\frac{\sigma_n}{\sqrt{n}}, \tag{45}$$

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

4.4 Impact of validation set proportion

An important and practical issue is the impact of the validation set proportion p_n on risk estimation and estimator selection. The following discussion provides some intuition regarding the behavior of the conditional risk $\tilde{\theta}_{p_n,n}$, of estimators $\hat{\Psi}(P^0_{n,B_n})$ based on cross-validation training sets of size $n(1-p_n)$,

compared to the conditional risk $\tilde{\theta}_n = \tilde{\theta}_{0,n}$, of estimators $\hat{\Psi}(P_n)$ based on the entire learning set of size n.

For non-quadratic loss functions, we argue below that the expected value with respect to split vectors B_n in the definition of $\tilde{\theta}_{p_n,n}$ (i.e., the multiple splits) results, to a first order linear approximation, in a risk difference $\tilde{\theta}_{p_n,n} - \tilde{\theta}_n$ of zero, for each fixed $p_n \in (0,1)$. Suppose that

$$\tilde{\theta}_n - \theta = \frac{1}{n}\sum_{i=1}^{n} IC(X_i|P) + R(P_n, P), \tag{46}$$

for some function $IC(\cdot|P)$ of X and remainder term $R(P_n, P)$. Then,

$$\tilde{\theta}_{p_n,n} - \theta = E_{B_n}\frac{1}{n(1-p_n)}\sum_{i=1}^{n} \mathrm{I}(B_n(i) = 0)\, IC(X_i|P) + E_{B_n} R(P^0_{n,B_n}, P).$$

By virtue of the expected value w.r.t. B_n, the first term actually equals $\sum_{i=1}^{n} IC(X_i|P)/n$. Consequently,

$$\tilde{\theta}_n - \tilde{\theta}_{p_n,n} = R(P_n, P) - E_{B_n} R(P^0_{n,B_n}, P). \tag{47}$$

Thus, for a fixed validation proportion $p_n \in (0,1)$, $\tilde{\theta}_{p_n,n}$ can be viewed as a decent approximation of $\tilde{\theta}_n = \tilde{\theta}_{0,n}$. This suggests that averaging over split vectors B_n significantly reduces the sensitivity of the cross-validated risk estimators $\hat{\theta}_{p_n,n}(k)$ and corresponding selector $\hat{k}_{p_n,n}$ to the choice of p_n, compared to single-split validation.

Note that the preceding argument is only interesting for non-quadratic loss functions, i.e., for loss functions for which one has a first order term in the expansion for $\tilde{\theta}_n - \theta$. For the quadratic loss function $L(X,\psi) = (Y - \psi(W))^2$, one has $\tilde{\theta}_n - \theta = \int (\psi_n(W) - \psi(W))^2 dP(W)$, hence the influence curve can trivially be chosen as zero, $IC(X|P) \equiv 0$.

5 Discussion

The present article derived distributional properties of cross-validated risk estimators in the context of estimator selection and performance assessment. We stress that, unlike previously published results, the theorems derived in this and our related articles apply to general data generating distributions, loss functions (i.e., parameters), estimators (e.g., linear and non-linear

predictors, parametric and non-parametric density estimators), and cross-validation procedures (e.g., V-fold and Monte Carlo cross-validation). The reader is referred to van der Laan and Dudoit (2003) for a detailed discussion of the general loss-based estimation framework introduced in Section 2 and for extensions of the results to estimation based on censored data.

For estimator selection, the asymptotic optimality of cross-validation procedures is established, in the sense that a selector based on cross-validated risk estimators performs asymptotically as well as an optimal oracle selector based on the risk under the true, unknown data generating distribution. That is, for a fixed validation set proportion $p_n \in (0,1)$, the ratio of conditional risk differences comparing the cross-validated selector $\hat{k}_{p_n,n}$ to the optimal oracle selector $\tilde{k}_{p_n,n}$, $(\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta)/(\tilde{\theta}_{p_n,n}(\tilde{k}_{p_n,n}) - \theta)$, converges to one in probability (Theorems 1 and 2). For a sequence p_n converging to zero slowly enough with the sample size n, Corollary 1 proves asymptotic equivalence of the cross-validated selector $\hat{k}_{p_n,n}$ and the absolutely optimal oracle selector $\tilde{k}_n$, based on the entire empirical distribution P_n, that is, $(\tilde{\theta}_{p_n,n}(\hat{k}_{p_n,n}) - \theta)/(\tilde{\theta}_n(\tilde{k}_n) - \theta)$ converges to one in probability. In the special case of the quadratic loss function, Theorem 1 provides a stronger convergence result than Theorem 2: for the L_2 loss function, the rate of convergence is shown to be $O(\log(K_n)/np_n)$ rather than the slower $O(\log(K_n)/\sqrt{np_n})$ applicable to general loss functions. Note that while the finite sample results hold for any p_n, the asymptotic results are derived under the assumption that the size np_n of the validation sets converges to infinity and hence do not cover leave-one-out cross-validation. However, one could possibly derive similar results for LOOCV under different conditions.

For performance assessment, cross-validated risk estimators are shown, under certain conditions, to be consistent and asymptotically linear for the conditional risk $\tilde{\theta}_{p_n,n}$ (Theorem 3). The asymptotic linearity result allows us to derive confidence intervals for $\tilde{\theta}_{p_n,n}$ (Section 4.3).

An important and practical issue is the impact of the validation set proportion p_n on risk estimation and estimator selection. Preliminary sensitivity analysis results are discussed in related articles on likelihood-based cross-validation (van der Laan et al., 2004a) and cross-validation selection for regression with censored outcomes (Keleş et al., 2004). In practice, we have found that averaging over split vectors B_n can significantly reduce the sensitivity of the cross-validated selector $\hat{k}_{p_n,n}$ to the choice of p_n, compared to single-split validation. For non-quadratic loss functions, Section 4.4 pro-

vides some intuition regarding the impact of the validation set proportion p_n. It suggests, in particular, that the risk difference $\tilde{\theta}_{p_n,n} - \tilde{\theta}_n$ is zero in first order.

References

A. Antos, L. Devroye, and L. Györfi. Lower bounds for Bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7):643–645, 1999.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer-Verlag, New York, 1997.

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, University of California, Berkeley, 1996b.

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.

P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76:503–514, 1989.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

L. Devroye, D. Schäfer, L. Györfi, and H. Walk. The estimation problem of minimum mean squared error. *Statistics and Decisions*, 21:15–28, 2003.

S. Dudoit, M. J. van der Laan, S. Keleş, A. M. Molinaro, S. E. Sinisi, and S. L. Teng. Loss-based estimation with cross-validation: Applications to microarray data analysis. In G. Piatetsky-Shapiro and P. Tamayo, editors, *Microarray Data Mining*, volume 5 of *SIGKDD Explorations*, pages 56–68. ACM, 2003.

B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983.

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

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320–328, 1975.

G. Gong. Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *Journal of the American Statistical Association*, 81:108–113, 1986.

L. Györfi and M. Horváth. On the asymptotic normality of a resubstitution error estimate. In A. Rizzi, M. Vichi, and H. H. Bock, editors, *Advances in Data Science and Classification*, pages 197–204. Springer, 1998.

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

L. Györfi, D. Schäfer, and H. Walk. Relative stability of global errors of nonparametric function estimators. *IEEE Transactions on Information Theory*, 48(8):2230–2242, 2002b.

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 with right censored outcomes. *Bernoulli*, 10(6):1011–1037, 2004.

K-C Li. Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics*, 15:958–975, 1987.

P. J. McCarthy. The use of balanced half-sample replication in cross-validation studies. *Journal of the American Statistical Association*, 71: 596–604, 1976.

A. M. Molinaro, S. Dudoit, and M. J. van der Laan. Tree-based multivariate regression and density estimation with right-censored data. In S. Dudoit, R. C. Gentleman, and M. J. van der Laan, editors, *Multivariate Methods in Genomic Data Analysis*, volume 90 of *Journal of Multivariate Analysis*, pages 154–177. Elsevier, 2004.

A. M. Molinaro and M. J. van der Laan. Deletion/Substitution/Addition algorithm for partitioning the covariate space in prediction. Technical Report 162, Division of Biostatistics, University of California, Berkeley, 2004. URL `www.bepress.com/ucbbiostat/paper162`.

R. R. Picard and R. D. Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79:575–583, 1984.

M. Pintér. On the rate of convergence of error estimates for the partitioning classification rule. *Theoretical Computer Science*, 284(1):181–196, 2002.

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

J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88:486–494, 1993.

S. E. Sinisi and M. J. van der Laan. Deletion/substitution/addition algorithm in learning with applications in genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 18, 2004. URL `www.bepress.com/sagmb/vol3/iss1/art18`.

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 and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive ϵ-net estimator: Finite sample oracle inequalities and examples. Technical Report 130, Division of Biostatistics, University of California, Berkeley, 2003. URL `www.bepress.com/ucbbiostat/paper130`.

M. J. van der Laan, S. Dudoit, and S. Keleş. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 4, 2004a. URL `www.bepress.com/sagmb/vol3/iss1/art4`.

M. J. van der Laan, S. Dudoit, and A. W. van de Vaart. The cross-validated adaptive ϵ-net estimator. Technical Report 142, Division of Biostatistics, University of California, Berkeley, 2004b. URL `www.bepress.com/ucbbiostat/paper142`.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York, 2003.

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

P. Zhang. Model selection via multifold cross-validation. *Annals of Statistics*, 21:299–313, 1993.

Table 1: *Loss functions.* Examples of full data loss functions, $L(X, \psi)$, for different estimation problems.

Data	**Parameter**	**Loss function**
X	ψ	$L(X, \psi)$
Univariate prediction, continuous outcome, $Y \in I\!R$		
$X = (W, Y)$	Conditional mean:	Squared error (L_2):
	$\psi(W) = E[Y\|W]$	$L(X, \psi) = (Y - \psi(W))^2$
	Conditional median:	Absolute error (L_1):
	$\psi(W) = \text{Median}[Y\|W]$	$L(X, \psi) = \|Y - \psi(W)\|$
Univariate prediction, polychotomous outcome, $Y \in \{1, \ldots, K\}$		
$X = (W, Y)$	Class with max posterior probability:	Indicator:
	$\psi(W) = \text{argmax}_y \; Pr(y \mid W)$	$L(X, \psi) = \text{I}(Y \neq \psi(W))$
		(risk = classification error rate)
	Posterior class probabilities:	Negative log density:
	$\psi(X) = Pr(Y \mid W)$	$L(X, \psi) = -\log \psi(X)$
		(risk = entropy)
	Class with max posterior probability:	Gini:
	$\psi(X) = \text{I}\big(Y = \text{argmax}_y Pr(y \mid W)\big)$	$L(X, \psi) = 1 - \psi(X)$
Multivariate prediction, continuous outcome, $Y = (Y(m) : m = 1, \ldots, M) \in I\!R^M$		
$X = (W, Y)$	Conditional mean vector:	Quadratic (L_2):
	$\psi(W) = (E[Y(m) \mid W] : m = 1, \ldots, M)$	$L(X, \psi) = (Y - \psi(W))^\top \Omega(W)(Y - \psi(W))$
Density estimation		
X	Density (for c.d.f. F):	Negative log density:
	$\psi(X) = \frac{d}{dX} F(X)$	$L(X, \psi) = -\log \psi(X)$
Hazard function estimation		
$X = (W, T)$	Hazard function for T given W:	Negative log density for $g(T\|W)$:
	$\psi(W, T) = \lambda(T\|W) = -\frac{d}{dT} \log(\bar{G}(T\|W))$	$L(X, \psi) = -\log \psi(W, T) + \int_0^T \psi(W, u) du$

Survivor function: $\bar{G}(t|w) = 1 - G(t|w) = Pr(T > t|W = w)$;
Density: $g(t|w) = \frac{d}{dt} G(t|w) = \lambda(t|w) \exp(-\int_0^t \lambda(u|w) du)$.