Data-adaptive Loss-based Estimation with Cross-validation

Sandrine Dudoit and Mark J. van der Laan
Division of Biostatistics, UC Berkeley
www.stat.berkeley.edu/~sandrine
www.stat.berkeley.edu/~laan

PH243A – Fall 2003
Multivariate Statistical Methods in Genomics
Version: October 6, 2003

Acknowledgments

Joint work with

Sündüz Keleş, Division of Biostatistics, UC Berkeley.

Annette Molinaro, Division of Biostatistics, UC Berkeley.

Matthieu Cornec, INSEE, France.

Thanks to Joe Gray, Dan Moore, and Fred Waldman
(Comprehensive Cancer Center, UCSF) for providing the CGH breast cancer dataset.
Outline

- Estimation road map.
- Loss function.
- Estimator selection using cross-validation: finite sample results and asymptotic optimality.
- Estimator performance assessment using cross-validation: risk confidence intervals.
- Examples.
- Application 1: Likelihood cross-validation for the identification of regulatory motifs.
- Application 2: Tree-based prediction of survival based on microarray data.
Problem 1. **Prediction of biological and clinical outcomes using microarray measures of transcript levels or DNA copy number.**

Cells respond to various treatments/conditions by activating or repressing the expression of particular genes. **DNA microarrays** are high-throughput biological assays that can be used to measure **gene expression levels** on a genomic scale.

E.g. In cancer research, microarrays are used to measure transcript levels (i.e., mRNA levels) and DNA copy number in tumor samples for tens of thousands of genes at a time.

**Statistical question.** Relate microarray measures to biological and clinical outcomes.
Motivation: Microarray experiments

- Outcomes (phenotypes): tumor class, response to treatment, patient survival, affectedness/unaectedness — polychotomous or continuous; censored or uncensored.
- Explanatory variables (genotypes): measures of transcript (i.e., mRNA) levels for thousands of genes, DNA copy number for thousands of genes, age, sex, treatment, clinical predictors — polychotomous or continuous.

Small n, large p.

Motivation: Microarray experiments

- Selecting a good predictor: linear discriminant analysis (LDA), trees, support vector machines (SVMs), neural networks, other?
- Selecting a good subset of marker genes: How many genes? Which genes?
- Assessing the performance of the resulting predictor.

“Clinical outcome X for cancer Y can be predicted accurately based on gene expression measures.”
Motivation: Sequence analysis

Problem 2. Identification of regulatory motifs in DNA sequences.

Transcription factors (TF) are proteins that selectively bind to DNA to regulate gene expression.

Transcription factor binding sites, or regulatory motifs, are short DNA sequences (5–25 base pairs) in the upstream control region (UCR) of genes, i.e., in regions roughly 600–1,000 base pairs from the gene start site (in lower eukaryotes such as yeast).

E.g. GAL4 binding sites for different yeast genes (from SCPD).

>YBR019C TCGCGATACCTTCACCG
>YBR020W CGGGCGACGATTACCCG
>YLR081W TATCGGAGCGTAGGGCGCGAAAC
>YML051W CGGCATCCTACATGCGG
>YOR120W TCGTTCAGACAGGTCCCG
Motivation: Sequence analysis

From unaligned DNA sequence data, estimate motif start sites and base composition, i.e., position specific weight matrix (PWM).

- **Likelihood estimation** for DNA sequence data.

- **Prediction** of gene expression levels based on sequence features.
  E.g. Keles et al. (2002).

- Selecting a *good* model for transcription factor binding sites:
  Distribution of bases in motif? Distribution of bases in background sequence? Constraints on PWM? Motif length? Number of motifs per sequence?

- Assessing the performance of the resulting estimators.

Motivation: Genetic mapping

**Problem 3.** *Identification of genes associated with complex phenotypes.*

- Outcomes (phenotypes): affectedness/unaffectedness, quantitative trait, response to treatment, patient survival — polychotomous or continuous; *censored* or uncensored.

- Explanatory variables (genotypes): thousands of SNP genotypes, IBD status, age, sex — usually polychotomous.
General estimation road map

Our proposed unified strategy for estimator construction, selection, and performance assessment is driven by the choice of a loss function corresponding to the parameter of interest for the full, uncensored data structure.

The term estimator is used in a broad sense, to provide a common treatment of multivariate outcome prediction and density estimation problems based on censored data. Each of these problems can be dealt with by the choice of a suitable loss function.

Special cases and applications: Dudoit & van der Laan (2003), Keles et al. (2003a,2003b), van der Laan et al. (2003), Molinaro et al. (2003).

Estimation road map: Step 1

Step 1. Definition of the parameter of interest in terms of a loss function for the observed data.

- Full, uncensored data: define the parameter of interest as the minimizer of the expected loss, or risk, for a loss function chosen to represent the desired measure of performance.

- Observed, censored data: apply the general estimating function methodology of van der Laan & Robins (2002) to map the full, uncensored data loss function into an observed, censored data loss function having the same expected value and leading to an efficient estimator of this risk based on censored data.
Estimation road map: Step 1

Table 1: Examples of loss functions for different estimation problems.

<table>
<thead>
<tr>
<th>Estimation problem</th>
<th>Parameter</th>
<th>Loss function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>Conditional mean of an outcome given covariates</td>
<td>Squared error (L2)</td>
</tr>
<tr>
<td>Regression</td>
<td>Conditional median of an outcome given covariates</td>
<td>Absolute error (L1)</td>
</tr>
<tr>
<td>Classification</td>
<td>Posterior class probabilities</td>
<td>Indicator, Gini, negative log-likelihood</td>
</tr>
<tr>
<td>Density estimation</td>
<td>Density</td>
<td>Negative log-likelihood (deviance, Kullback-Leibler)</td>
</tr>
</tbody>
</table>

Estimation road map: Step 2

Step 2. Construction of candidate estimators based on a loss function for the observed data.

- Generate 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 defined as the minimizer of the empirical risk based on the observed data loss function.

E.g. stepwise variable selection; recursive binary partitioning of the covariate space in tree-based estimation; addition/deletion/substitution algorithm (van der Laan & Dudoit, 2003; Sinisi & van der Laan, 2003).
Estimation road map: Step 3

Step 3. Cross-validation estimator selection and performance assessment based on a loss function for the observed data.

Use cross-validation to estimate risk based on the observed data loss function and to select an optimal estimator among the candidates in Step 2.

van der Laan & Dudoit (2003): unified cross-validation methodology for selection among estimators, finite sample and asymptotic optimality results for the cross-validation selector for general data generating distributions, loss functions (possibly depending on a nuisance parameter), and estimators.

Full data structure

The full data structure is defined as a multivariate stochastic process

\[ X \equiv \mathcal{X}(T) = \{X(t) : 0 \leq t \leq T\}, \]

where \( T \) denotes a possibly random endpoint.

- \( W \): time-independent, or baseline, covariates.
- \( Z \equiv \log T \): log survival time.
- \( Z(t), t \in \{t_0 = 0, \ldots, t_{m-1} = T\}, T \) fixed: an outcome process of interest, included in \( X(t) \).

Denote the distribution of the full data structure \( X \) by \( F_{X,0} \).

In many applications, \( X = (W, Z) \).
Observed data structure

The observed data structure is

\[ O \equiv \left( \tilde{T} = \min(T, C), \Delta = I(T \leq C), \bar{X}(\tilde{T}) \right), \]

for a censoring variable \( C \) with conditional distribution \( G_0(\cdot | X) \), given the full data structure \( X \).

By convention, if \( T < C \), let \( C = \infty \). One can then rewrite the observed data structure as \( O = (\bar{X}(C), C) \).

The distribution, \( P_0 = P_{F_X,0,G_0} \), of the observed data structure \( O \) is indexed by the full data distribution \( F_{X,0} \) and the conditional distribution \( G_0(\cdot | X) \) of the censoring variable \( C \).

Coarsening at random (CAR) is assumed for the censoring mechanism \( C \)

\[ Pr_0(C = t \mid C \geq t, X(T)) = Pr_0(C = t \mid C \geq t, \bar{X}(t)), \quad \text{for } t < T. \]

If \( X \) does not include time-dependent covariates (e.g., \( X = (W, Z) \)), then, under CAR, the censoring time \( C \) is conditionally independent of the survival time \( T \), given baseline covariates \( W \).
Full data loss function

The parameter of interest, $\psi_0$, is a mapping, $\psi : S \rightarrow \mathbb{R}$, from a covariate space $S$ into the real line $\mathbb{R}$. Denote the parameter space by $\Psi$.

The parameter $\psi_0$ is defined in terms of a loss function, $L(X, \psi)$, as (one of) the minimizer(s) of the expected loss, or risk,

$$\int L(x, \psi_0) dF_{X,0}(x) \equiv \min_{\psi \in \Psi} \int L(x, \psi) dF_{X,0}(x).$$

Note that we do not require uniqueness of the risk minimizer, rather, we simply assume that there is a loss function such that the parameter of interest $\psi_0$ achieves the minimum risk.

---

Full data loss function

- **Univariate prediction**, $X = (W, Z)$.
  - Conditional mean: $\psi_0(W) = E_0[Z \mid W]$.
    Quadratic ($L_2$), or squared error, loss function:
    $$L(X, \psi) = (Z - \psi(W))^2.$$  
  - Conditional median: $\psi_0(W) = \text{Median}_0[Z \mid W]$.
    Absolute error ($L_1$) loss function: $L(X, \psi) = |Z - \psi(W)|$.

- **Multivariate prediction**, $X = (W, (Z(t_0), \ldots, Z(t_{m-1})))$.
  Conditional mean vector: $\psi_0(t, W) = E_0[Z(t) \mid W]$, $t \in \{t_0 = 0, \ldots, t_{m-1} = T\}$.
  Quadratic loss function: For a symmetric matrix function $\Omega(W)_{m \times m}$, $L(X, \psi) = (Z(\cdot) - \psi(\cdot, W))^\top \Omega(W)(Z(\cdot) - \psi(\cdot, W))$.

- **Density estimation**, $X = (T, W)$.
  Density: $\psi_0(T, W) = f_0(T, W)$.
  Negative log-likelihood loss function: $L(X, \psi) = -\log \psi(T, W)$. 

Observed data loss function

The general estimating function methodology of van der Laan & Robins (2002) maps the full data loss function $L(X, \psi)$ into an observed data loss function $L(O, \psi | \eta_0)$ with the same risk

$$\int_{\text{Observed data}} \frac{L(o, \psi | \eta_0)}{dP_0(o)} dP_0(o) = \int_{\text{Full data}} L(x, \psi) dF_{X,0}(x).$$

Here, $\eta_0$ denotes nuisance parameters $G_0$ and possibly $Q_0$, where $G_0$ identifies the conditional distribution of the censoring variable $C$ given $X$ and $Q_0 = Q(F_{X,0})$ identifies the $F_X$-part of the observed data density under the CAR assumption.

IPCW loss function

The inverse probability of censoring weighted (IPCW) loss function corresponding to the full data loss function $L(X, \psi)$ is

$$L(O, \psi | G) \equiv L(X, \psi) \frac{\Delta}{G(T|X)},$$

where $\bar{G}$ is a conditional survival function for $C$ given $X$ and $\Delta = I(T \leq C)$ is the censoring indicator.

Under CAR, $\bar{G}(T|X) = \bar{G}(T|W)$. 

IPCW loss function

In regression, $X = (W, Z)$ and the parameter of interest is the conditional mean: $\psi_0(W) = E_0[Z \mid W]$.

Full data loss function:

$$L(X, \psi) = (Z - \psi(W))^2.$$  

IPCW observed data loss function:

$$L(O, \psi \mid G) = (Z - \psi(W))^2 \frac{\Delta}{G(T \mid W)}.$$  

DR-IPCW loss function

The doubly robust inverse probability of censoring weighted (DR-IPCW) loss function is

$$L(O, \psi \mid Q, G) = \frac{L(X, \psi)\Delta}{G(T \mid X)} + \int E_{G, Q} \left( \frac{L(X, \psi)\Delta}{G(T \mid X)} \mid \bar{X}(u), \bar{T} \geq u \right) dM_G(u),$$

where

$$dM_G(u) = I(\bar{T} \in du, \Delta = 0) - I(\bar{T} \geq u)\lambda_c(u \mid X)du$$

and $Q = Q(F_X)$ refers to the $F_X$-part of the density for the observed data, $O = (\bar{X}(C), C)$, under the CAR assumption.
DR-IPCW loss function

Double robustness: The loss functions satisfy
\[
\int L(o, \psi \mid Q, G)dP_0(o) = \int L(x, \psi)dF_{X,0}(x),
\]
if either \( G = G_0 \) or \( Q = Q_0 \).

The estimator selection problem

- Suppose we have a learning set of \( n \) independent and identically distributed (i.i.d.) observations, \( O_1, \ldots, O_n \), with \( O_i \sim P_0 \). Let \( P_n \) be the empirical distribution of \( O_1, \ldots, O_n \).
- Let \( \hat{\psi}_k(\cdot) = \psi_k(\cdot \mid P_n) \in \Psi, k = 1, \ldots, K_n \), be a collection of candidate estimators of the parameter \( \psi_0(\cdot) \).

E.g. In tree-based estimation, the \( \hat{\psi}_k \) are obtained by recursive binary partitioning of the covariate space using one of the above observed data loss functions; \( k \) corresponds to tree size.
The estimator selection problem

The selection problem. Choose a data adaptive \( \hat{k} = k(P_n) \) so that the distance, or risk difference,

\[
d_n(\hat{\psi}_k, \psi_0) = \int \left\{ L(o, \hat{\psi}_k | \eta_0) - L(o, \psi_0 | \eta_0) \right\} dP_0(o)
\]

(observed data loss function)

\[
d_n(\hat{\psi}_k, \psi_0) = \int \left\{ L(x, \hat{\psi}_k) - L(x, \psi_0) \right\} dF_{X,0}(x)
\]

(full data loss function)

\[\longrightarrow 0 \quad \text{at asymptotically optimal rate.}\]

• For the squared error loss function, 
  \[L(X, \psi) = L_2(X, \psi) = (Z - \psi(W))^2,\]
  the risk difference simplifies to

\[
d_n(\hat{\psi}_k, \psi_0) = \int \left\{ L_2(x, \hat{\psi}_k) - L_2(x, \psi_0) \right\} dF_{X,0}(x)
\]

\[
= \int \left( \hat{\psi}_k(w) - \psi_0(w) \right)^2 dF_{W,0}(w).
\]

• For the negative log-likelihood loss function, the risk difference is the Kullback-Leibler divergence between \( \hat{\psi}_k \) and \( \psi_0 \)

\[
d_n(\hat{\psi}_k, \psi_0) = -\int \log \left( \frac{\hat{\psi}_k(x)}{\psi_0(x)} \right) \psi_0(x) d\mu(x).
\]
The estimator selection problem

The optimal benchmark selector. Let

\[ \tilde{k}_n \equiv \arg\min_k d_n(\hat{\psi}_k, \psi_0) \]

denote the minimizer of the distance \( d_n(\hat{\psi}_k, \psi_0) \). This optimal
benchmark selector depends on the unknown data generating
distribution \( P_0 \).

A selector \( \hat{k} = k(P_n) \) is asymptotically equivalent with the optimal
benchmark \( \tilde{k}_n \) if

\[ \frac{d_n(\hat{\psi}_{\hat{k}}, \psi_0)}{d_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)} \xrightarrow{\text{in probability}} 1 \text{ as } n \to \infty. \]

In particular, then it is asymptotically optimal.

van der Laan & Dudoit (2003): finite sample and asymptotic
optimality results for the cross-validation selector.

The estimator selection problem

The selection problem involves estimating the conditional risk

\[ \tilde{\theta}_n(k) \equiv \int L(o, \psi_k(\cdot | P_n) | \eta_0)dP_0(o) \]

for each candidate estimator \( \hat{\psi}_k(\cdot) = \psi_k(\cdot | P_n) \in \Psi, k = 1, \ldots, K_n. \)

Cross-validation is a general approach for risk estimation and
estimator selection.
General framework for cross-validation

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 performance of (or validate) this estimator(s).

The cross-validation estimator $\hat{\psi}_k$ is chosen to have the best performance on the validation set.

To derive a general representation for the cross-validation selector $\hat{k}$, we introduce a binary random $n$-vector, or split 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 learning sample of $n$ observations into a training set and validation set:

$$S_{n,i} = \begin{cases} 
0, & \text{ith observation is in the training sample,} \\
1, & \text{ith observation is in the validation sample.}
\end{cases}$$

The particular distribution of $S_n$ defines the type of cross-validation procedure.
General framework for cross-validation

Let \( P_{n,S_n}^0 \) and \( P_{n,S_n}^1 \) denote the empirical distributions of the training and validation sets, respectively, and let \( p = p_n = n_1/n \) be the proportion of observations in the validation set.

A general definition of the cross-validation selector is

\[
\hat{k} \equiv \arg\min_k E_{S_n} \int L(o, \psi_k(\cdot \mid P_{n,S_n}^0, \eta_{n,S_n}^0)) dP_{n,S_n}^1(o)
\]

\[= \arg\min_k E_{S_n} \sum_{\{i: S_n, i = 1\}} L(O_i, \psi_k(\cdot \mid P_{n,S_n}^0, \eta_{n,S_n}^0)).\]

Here, \( \psi_k(\cdot \mid P_{n,S_n}^0) \) and \( \eta_{n,S_n}^0 \) denote, respectively, estimators for the parameter of interest \( \psi_0 \) and the nuisance parameter \( \eta_0 \), using only the training set.

---

**Figure 1:** *Five-fold cross-validation. S_n has 5 realizations.*
General framework for cross-validation

The particular distribution of the split vector $S_n$ defines the type of cross-validation procedure. This representation covers many types of CV procedures.

- **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 $S_n$ places mass $1/n$ on each the $n$ binary vectors $s_n = (s_{n,1}, \ldots, s_{n,n})$ such that $\sum_i s_{n,i} = 1 (p_n = 1/n)$.

- **V-fold cross-validation.** The learning set is randomly divided into $V$ mutually exclusive and exhaustive sets, each used in turn as the validation sets. The corresponding distribution of $S_n$ places mass $1/V$ on each of $V$ binary vectors $s_n^v = (s_{n,1}^v, \ldots, s_{n,n}^v)$, $v = 1, \ldots, V$, such that $\sum_i s_{n,i}^v \approx n/V$ and $\sum_v s_{n,i}^v = 1 (p_n = 1/V)$.

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

- **Bootstrap-based cross-validation.** The training sets are based on bootstrap samples and the validation sets on the corresponding left-out samples. $E[p_n] = E[\sum_i S_{n,i}/n] = (1 - 1/n)^n \approx e^{-1} \approx .368$. E.g. $.632$ bootstrap estimator (Efron, 83).
Honest cross-validation

Prediction error rates, or related measures, are usually reported to

- compare the performance of different predictors;
- support statements such as "Clinical outcome X for cancer Y can be predicted accurately based on microarray gene expression measures."

It is common practice in microarray experiments to screen genes and fine-tune predictor parameters (e.g., number of neighbors $k$ in nearest neighbor classification, kernel in SVMs) using all the learning set and then perform cross-validation only on the predictor building portion of the process.

$\implies$ The reported error rates are usually biased downward and give an overly optimistic view of the predictive power of microarray expression measures.

$\implies$ Predictors are not compared on an equal footing.
Honest cross-validation

Prediction error rates (risk) can be estimated by cross-validation (CV), but...

- These estimates relate only to the experiment that was cross-validated.
- It is essential to perform cross-validation on the entire predictor training process, including feature selection and other training decisions (e.g., choice for the number of neighbors in $k$-NN, kernel in SVMs).
- Otherwise, risk estimates can be severely biased downward, i.e., overly optimistic.


---

Honest cross-validation

**Resubstitution estimation.** The entire learning set is used to perform feature selection, build the classifier, and estimate classification error.

**Internal cross-validation.** Feature selection is done on the entire learning set, CV is applied only to the classifier building process.

**External cross-validation.** CV is applied to the feature selection AND the classifier building process.
### Figure 2: Estimates of classification error by leave-one-out cross-validation

**Breast tumor nodal dataset, 25 nodal+ and 24 nodal- tumors (West et al., 2001).**

<table>
<thead>
<tr>
<th></th>
<th>DLDA</th>
<th>1-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resubstitution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internal LOOCV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>External LOOCV</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Graph showing classification error for DLDA and 1-NN methods](image)
Estimator selection using cross-validation

Define the distance, or risk difference, for estimators based on training samples of size \( n(1 - p) \) as

\[
d_{n(1-p)}(\hat{\psi}_k, \psi_0) \equiv \mathbb{E}_{S_n} \int \{ L(o, \psi_k(\cdot | P_{n,S_n}^0) | \eta_0) - L(o, \psi_0(\cdot | \eta_0) \} dP_0(o). \]

The selector \( \hat{k} \) aims to minimize this unknown distance.

Denote the unknown minimizer, i.e., the comparable optimal benchmark selector for \( n(1 - p) \) observations by

\[
\tilde{k}_{n(1-p)} \equiv \arg \min_k d_{n(1-p)}(\hat{\psi}_k, \psi_0).
\]

Theorem 1. (Stated in special case of known \( \eta_0 \), \( L(O, \psi | \eta_0) = L(O, \psi) \)).

Suppose that

- **A1.** the loss function \( L(O, \psi) \) is uniformly bounded by \( M_1 \), and
- **A2.** there exists an \( 0 \leq M_2 < \infty \) so that for all \( k \)

\[
\int \{ L(o, \psi_k(\cdot | P_{n,S_n}^0) - L(o, \psi_0(\cdot)) \}^2 dP_0(o) \leq M_2 \int \{ L(o, \psi_k(\cdot | P_{n,S_n}^0) - L(o, \psi_0(\cdot)) \} dP_0(o) \ a.s.
\]

 Finite sample result. For any \( \delta > 0 \) and constant \( C(M_1, M_2, \delta) \)

\[
0 \leq \mathbb{E}d_{n(1-p)}(\hat{\psi}_{\tilde{k}}, \psi_0) \leq (1 + 2\delta) \mathbb{E}d_{n(1-p)}(\hat{\psi}_{\tilde{k}_{n(1-p)}}, \psi_0) + C(M_1, M_2, \delta) \frac{1 + \log(K_n)}{np}.
\]
Estimator selection using cross-validation

**Asymptotic optimality.** If

\[
\frac{\log(K_n)}{(np) Ed_n(1-p)(\hat{\psi}_{k_n(1-p)}, \psi_0)} \longrightarrow 0, \quad \text{as } n \to \infty,
\]

then

\[
\frac{Ed_n(1-p)(\hat{\psi}_k, \psi_0)}{Ed_n(1-p)(\hat{\psi}_{k_n(1-p)}, \psi_0)} \longrightarrow 1, \quad \text{as } n \to \infty.
\]

**Corollary.** In addition to the conditions of Theorem 1, suppose that, as \( n \to \infty, \ p = p_n \to 0 \) slowly enough that

\[
\frac{\log(K_n)}{(np) Ed_n(1-p)(\hat{\psi}_{k_n(1-p)}, \psi_0)} \longrightarrow 0,
\]

and

\[
\frac{Ed_n(\hat{\psi}_k, \psi_0)}{Ed_n(1-p)(\hat{\psi}_{k_n(1-p)}, \psi_0)} \longrightarrow 1.
\]

Then,

\[
\frac{Ed_n(1-p)(\hat{\psi}_k, \psi_0)}{Ed_n(\hat{\psi}_k, \psi_0)} \longrightarrow 1, \quad \text{as } n \to \infty.
\]

That is, the data adaptive CV selector \( \hat{k} \) is asymptotically optimal.
Estimator selection using cross-validation

- The corresponding convergence in probability of the ratios of risk differences follows by noting that $E|Z_n| = O(g(n))$ implies $Z_n = O_P(g(n))$, for a positive function $g(n)$.
- A more general version of Theorem 1 was derived for loss functions that depend on a nuisance parameter $\eta_0$.
- An analog of Theorem 1, which does not require Assumption A2, was derived. In this case, convergence is shown to be $O(\log(K_n)/\sqrt{np})$ rather than $O(\log(K_n)/np)$.

van der Laan & Dudoit (2003)

Estimator selection using cross-validation

- 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.
- 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 theorems apply to general distributions $P_0$, general loss functions $L(O, \psi \mid \eta_0)$, and general estimators $\psi(\cdot \mid P_n)$. 
**Estimator performance assessment**

Consider a particular estimator $\hat{\psi}(\cdot) = \psi(\cdot | P_n)$ and loss function $L(O, \psi | \eta_0) = L(O, \psi)$ with known $\eta_0$.

**Cross-validation risk estimator** (observable random variable)

$$\hat{\theta}_{n(1-p)} \equiv ES_n \int L(o, \psi(\cdot | P^0_{n,S_n}))dP^1_{n,S_n}(o).$$

**Conditional risk, $n(1-p)$ observations** (unknown random variable)

$$\check{\theta}_{n(1-p)} \equiv ES_n \int L(o, \psi(\cdot | P^0_{n,S_n}))dP_0(o).$$

**Conditional risk, $n$ observations** (unknown random variable)

$$\tilde{\theta}_n \equiv \int L(o, \psi(\cdot | P_n))dP_0(o).$$

**Asymptotic risk** (unknown parameter)

$$\theta \equiv \int L(o, \psi(\cdot | P_0))dP_0(o).$$

---

**Asymptotic linearity of CV risk estimator**

**Theorem.** Suppose the loss function $L(O, \psi)$ is uniformly bounded by $M_1$ and

$$ES_n \left[ \frac{\int \left\{ L(o, \psi(\cdot | P^0_{n,S_n})) - L(o, \psi(\cdot | P_0)) \right\}^2 dP_0(o)}{p_n} \right] = o_P(1).$$

Then

$$\hat{\theta}_{n(1-p)} - \check{\theta}_{n(1-p)} = \frac{1}{n} \sum_{i=1}^{n} \left\{ L(O_i, \psi(\cdot | P_0)) - \theta \right\} + o_P(1/\sqrt{n}).$$
Risk confidence intervals

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

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

where

\[ \hat{\sigma}^2 = \int (IC(o \mid P_n))^2 dP_n(o), \]

\[ IC(o \mid P_n) = L(o, \psi(\cdot \mid P_n)) - \int L(o, \psi(\cdot \mid P_n))dP_n(o), \]

and $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ for the standard normal cumulative distribution function $\Phi(\cdot)$.
Simulation study: Risk confidence intervals

Figure 3: Convergence to zero of $\hat{\theta} - \hat{\theta}_0$.

$X_j \sim N(Y_1^2, 1)$,

$Y \sim B(1/2)$,

$D(A)$, $LDA$, $rpart$, two- and ten-fold CV, 200 simulations.
van de L, Y. Dudoit (2003)

5. Prediction of multivariate outcomes.
4. Survival function estimation.
3. Predictor based on right-censored outcomes.
2. Density estimation.
1. Prediction of polychotomous and continuous outcomes.

A framework by choosing a suitable loss function.
The above results hold for general cross-validation procedures and apply to general distributions $P_0$, general loss functions $L(O; \theta)$, and general estimators $\hat{\theta}(T)$. The following problems can be addressed within our general estimation framework by choosing a suitable loss function.

1. Prediction of polychotomous and continuous outcomes.
2. Density estimation.
3. Predictor based on right-censored outcomes.
4. Survival function estimation.
5. Prediction of multivariate outcomes.

Applications of estimation:

- Risk confidence intervals.
- Confidence intervals.
- $X \sim N(\mu, \Sigma)$, $Y \sim N(\mu, \Sigma)$.
- $n = 100, 200, 500, 1000, \text{ten-fold CV}$.
- True $\theta$ and $\hat{\theta}$.
Example 1: Predictor selection

Suppose we have a learning set of \( n \) i.i.d. observations
\[ O = (W, Z) \sim P_0, \]
where \( Z \) is an outcome of interest and \( W \) a vector of explanatory variables.

Consider the quadratic loss function
\[ L(O, \psi) = (Z - \psi(W))^2. \]

The parameter of interest, which minimizes the risk
\[ E_0[L(O, \psi)] = \int (z - \psi(w))^2 dP_0(o), \]
is the conditional expectation \( \psi_0(W) = E_0[Z \mid W] \).

Example 1: Predictor selection

Given candidate predictors \( \hat{\psi}_k = \psi_k(\cdot \mid P_n) \), the risk difference for the quadratic loss simplifies to
\[ d_n(\hat{\psi}_k, \psi_0) = \int (\psi_k(w \mid P_n) - \psi_0(w))^2 dF_{W,0}(w). \]

The cross-validation selector is given by
\[ \hat{k} = \arg\min_k E_{S_n} \int (z - \psi_k(w \mid P_{n,S_n})^2 dP_{1,n,S_n}(o) \]
\[ = \arg\min_k E_{S_n} \sum_{\{i: S_n, i=1\}} (Z_i - \psi_k(W_i \mid P_{n,S_n}^0))^2. \]
Example 1: Predictor selection

Prediction of biological and clinical outcomes using microarray gene expression measures or SNP marker genotypes.

- Outcomes (phenotypes), $Z$: tumor class, response to treatment, patient survival, affectedness/ unaffectedness — polychotomous or continuous; censored (see Example 3, below) or uncensored.

- Explanatory variables (genotypes), $W$: measures of transcript (i.e., mRNA) levels for thousands of genes, DNA copy number for thousands of genes, SNP haplotypes, age, sex, treatment, clinical predictors — polychotomous or continuous.

Example 1: Predictor selection

Prediction of gene expression levels using DNA sequence data to identify transcription factor binding sites.

- Outcomes (phenotypes), $Z$: microarray gene expression measures — multivariate outcomes.

- Explanatory variables (genotypes), $W$: DNA sequence in upstream control region of genes.

Keleș et al. (2002). Bioinformatics.
Example 2: Density estimator selection

Suppose we have a learning set of \( n \) i.i.d. observations \( O \sim f_0 \equiv \frac{dP_0}{d\mu} \). Consider the log-likelihood loss function (a.k.a. cross-entropy loss, deviance)

\[
L(O, f) = -\log(f(O)).
\]

The parameter of interest, which minimizes the risk

\[
E_0[-L(O, f)] = -\int \log f(o) f_0(o) d\mu(o),
\]

is the density itself, \( \psi_0 = f_0 \).

---

Example 2: Density estimator selection

Given candidate density estimators, \( \hat{\psi}_k = f_k(\cdot \mid P_n) \), of \( \psi_0 = f_0 \), the risk difference is the Kullback-Leibler divergence between \( f_k(\cdot \mid P_n) \) and \( f_0 \)

\[
d_n(\hat{\psi}_k, \psi_0) = -\int \log \left( \frac{f_k(o \mid P_n)}{f_0(o)} \right) f_0(o) d\mu(o).
\]

The cross-validation selector is given by

\[
\hat{k} = \arg\min_k E_{S_n} \int \log f_k(o \mid P_n) dP_{n,S_n}^1(o) = \arg\min_k E_{S_n} \sum_{\{i: S_{n,i}=1\}} \log f_k(O_i \mid P_{n,S_n}^0).
\]
Example 2: Density estimator selection

Consider the special case when \( O = (W, Z) \), with \( Z|W \sim N(\psi_0(W), \sigma^2) \), \( \psi_0(W) = E_0[Z|W] \), and known variance \( \sigma^2 \).

The conditional density of \( Z \) given \( W \), corresponding to a candidate estimator \( \psi_k(\cdot|P_n) \), is denoted by \( f_k(z; w | P_n) \).

Then, the risk for the log-likelihood loss function is equal to the risk based on the squared error loss (up to + and \( \times \) constants)

\[
- \int \log f_k(z; w | P_n) f_0(o) d\mu(o) \\
= - \int \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{1}{2\sigma^2} (z - \psi_k(w|P_n))^2 \right) \right\} f_0(o) d\mu(o) \\
= \int (z - \psi_k(w|P_n))^2 f_0(o) d\mu(o).
\]

Example 2: Density estimator selection

Likelihood-based cross-validation for bandwidth selection in kernel density estimation.

- The true density \( f_0 \) is standard normal with compact support in the interval \([-2, 2]\).
- \( B = 20 \) replicate datasets were generated from \( f_0 \) for six different sample sizes, \( n = 50, 100, 200, 400, 800, 1600 \).
- The Gaussian kernel density estimator, \( \hat{f}_k(\cdot) = f_k(\cdot | P_n) \), for a learning set \( x_1, \ldots, x_n \) is given by

\[
\hat{f}_k(x) = \frac{1}{nk} \sum_{i=1}^{n} \phi \left( \frac{x - x_i}{k} \right),
\]

where \( \phi(.) \) is the standard normal density function and \( k \) is the bandwidth. \( K_n = 100 \) different bandwidth values \( k \) were considered from the interval \([0.02, 2]\), so that the difference between any two consecutive bandwidth values is 0.02.
Example 2: Density estimator selection

Figure 5: $\frac{d_{n(1-p)}(\hat{\psi}_L, \psi_0)}{d_{n(1-p)}(\hat{\psi}_{k_n(1-p)}, \psi_0)}$ vs. $n$, for $p = 1/10$. The bandwidth $\hat{k}$ was selected using ten-fold CV ($p = 1/10$), for 20 replicate datasets at each of six sample sizes, $n$.

Table 2: $\frac{\hat{E}d_{n(1-p)}(\hat{\psi}_L, \psi_0)}{Ed_{n(1-p)}(\hat{\psi}_{k_n(1-p)}, \psi_0)}$ vs. $n$, for $p = 1/10$. The estimated distance ratios are based on 20 replicate datasets at each of the six different sample sizes $n$. The bandwidth $\hat{k}$ was selected using ten-fold CV ($p = 1/10$).
Example 2: Density estimator selection

![Graph showing density estimates for different sample sizes and bandwidths.]

Figure 6: Cross-validation density estimates $\hat{f}_k$ and true density $f_0$. The cross-validation kernel density estimate $f_k(\cdot | P_n)$ is shown for six sample sizes, $n = 50, 100, 200, 400, 800, 1600$, for one simulated dataset. The bandwidth $\hat{k}$ was selected using ten-fold CV ($p = 1/10$).

---

Table 3: $V$-fold likelihood cross-validation: $\frac{\hat{E}_{d_n}(\hat{\psi}_k(p), \psi_0)}{\hat{E}_{d_n}(\psi_k^* n, \psi_0)}$ vs. $n$ and $p$.

<table>
<thead>
<tr>
<th>$p$</th>
<th>0.05</th>
<th>0.1</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>1.493594</td>
<td>1.465201</td>
<td>1.168274</td>
<td>1.115338</td>
<td>1.089441</td>
<td>1.047685</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.531736</td>
<td>1.391971</td>
<td>1.144236</td>
<td>1.136916</td>
<td>1.075563</td>
<td>1.048454</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>1.577241</td>
<td>1.473550</td>
<td>1.118831</td>
<td>1.117599</td>
<td>1.076197</td>
<td>1.061919</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>1.518429</td>
<td>1.417260</td>
<td>1.120498</td>
<td>1.120069</td>
<td>1.065835</td>
<td>1.064060</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>1.302580</td>
<td>1.443560</td>
<td>1.111674</td>
<td>1.182325</td>
<td>1.060759</td>
<td>1.100572</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>1.430726</td>
<td>1.388704</td>
<td>1.148916</td>
<td>1.119423</td>
<td>1.080356</td>
<td>1.083632</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Estimated distance ratios are based on 20 replicate datasets at six different sample sizes $n$ and for ten different validation set proportions $p = 1/V$. 

---
Example 2: Density estimator selection

<table>
<thead>
<tr>
<th>( p )</th>
<th>0.05</th>
<th>0.1</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.30</th>
<th>0.35</th>
<th>0.40</th>
<th>0.45</th>
<th>0.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>50</td>
<td>100</td>
<td>200</td>
<td>400</td>
<td>800</td>
<td>1600</td>
<td>20</td>
<td>30</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>0.05</td>
<td>21.778985</td>
<td>30.591547</td>
<td>5.366258</td>
<td>3.488738</td>
<td>2.147304</td>
<td>1.287172</td>
<td>2.495359</td>
<td>4.262036</td>
<td>1.246727</td>
<td>1.232388</td>
</tr>
<tr>
<td>0.1</td>
<td>4.969151</td>
<td>8.139912</td>
<td>3.709904</td>
<td>2.105173</td>
<td>1.948626</td>
<td>1.291611</td>
<td>2.260952</td>
<td>3.862468</td>
<td>1.511450</td>
<td>1.111143</td>
</tr>
<tr>
<td>0.15</td>
<td>1.972465</td>
<td>5.234631</td>
<td>2.283455</td>
<td>1.831317</td>
<td>1.628340</td>
<td>1.153562</td>
<td>1.446852</td>
<td>1.615702</td>
<td>1.276998</td>
<td>1.123451</td>
</tr>
<tr>
<td>0.20</td>
<td>1.836114</td>
<td>10.036376</td>
<td>2.465654</td>
<td>1.377272</td>
<td>1.370639</td>
<td>1.093183</td>
<td>1.583617</td>
<td>1.757668</td>
<td>1.263186</td>
<td>1.170124</td>
</tr>
<tr>
<td>0.25</td>
<td>2.495359</td>
<td>4.262036</td>
<td>1.246727</td>
<td>1.232388</td>
<td>1.209813</td>
<td>1.092931</td>
<td>1.335555</td>
<td>2.193936</td>
<td>1.258745</td>
<td>1.164263</td>
</tr>
<tr>
<td>0.30</td>
<td>2.260952</td>
<td>4.298054</td>
<td>1.410498</td>
<td>1.149826</td>
<td>1.215430</td>
<td>1.123646</td>
<td>1.583617</td>
<td>1.757668</td>
<td>1.263186</td>
<td>1.170124</td>
</tr>
<tr>
<td>0.35</td>
<td>1.553013</td>
<td>3.862468</td>
<td>1.511450</td>
<td>1.111143</td>
<td>1.165148</td>
<td>1.151871</td>
<td>1.446852</td>
<td>1.615702</td>
<td>1.276998</td>
<td>1.123451</td>
</tr>
<tr>
<td>0.40</td>
<td>1.446852</td>
<td>1.615702</td>
<td>1.276998</td>
<td>1.123451</td>
<td>1.146859</td>
<td>1.113719</td>
<td>1.583617</td>
<td>1.757668</td>
<td>1.263186</td>
<td>1.170124</td>
</tr>
<tr>
<td>0.45</td>
<td>1.446852</td>
<td>1.615702</td>
<td>1.276998</td>
<td>1.123451</td>
<td>1.146859</td>
<td>1.113719</td>
<td>1.583617</td>
<td>1.757668</td>
<td>1.263186</td>
<td>1.170124</td>
</tr>
<tr>
<td>0.50</td>
<td>1.335555</td>
<td>2.193936</td>
<td>1.258745</td>
<td>1.164263</td>
<td>1.149889</td>
<td>1.175700</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Single-split likelihood cross-validation: \( \frac{Ed_n(\hat{\psi}_{k(p)},\psi_0)}{Ed_n(\psi_{kn},\psi_0)} \) vs. \( n \) and \( p \). Estimated distance ratios are based on 20 replicate datasets at six different sample sizes \( n \) and for ten different validation set proportions \( p \).
Figure 7: *V*-fold vs. single split CV: $\hat{E}_{dn}(\psi_{k(p)}, \psi_0)$ vs. $n$. Estimated distance ratios are based on 20 replicate datasets at six different sample sizes $n$ and for ten different validation set proportions $p$.

Application 1: Identification of regulatory motifs

Incorporating biological knowledge in the identification of regulatory motifs in DNA sequences.

- Palindromic binding sites.
  E.g. CACGTG with reverse complement CACGTG.

- Binding sites with gaps.
  E.g. GCGNNNNNNNNNNNNNTAG.

- Information content profile of the binding site PWM. The information content (IC) of the PWM $(p_{wj})$ at position $w$ is

$$IC(w) = 2 + \sum_{j=1}^{4} p_{wj} \log_2 p_{wj} = 2 - \text{Entropy}(w) \in [0, 2].$$

The information content profile of a PWM is a measure of a site’s tolerance for substitution: high IC, low tolerance.
Application 1: Identification of regulatory motifs

- Direct relationship between the structural footprint of a protein on DNA and the information content profile of the PWM (Mirny & Gelfand, 2002).
- Transcription factors that have similar structures bind to sites with similar information content profiles (Eisen, 2002).
- The specific nature of TF–DNA interactions imposes constraints on the types of sequences that are likely to be TF binding sites (Eisen, 2002).

E.g. GAL4 binding sites for different yeast genes (from SCPD).

>YBR019C TCGCGATACCTTCACCG
>YBR020W CGGGCGACGATTACCGC
>YLR081W TATCGGAGCGTAGGCAGCGCAGC
>YML051W CGGCATCTACATGGCC
>YOR120W TCGGTGCACAGGTCCCG
Application 1: Identification of regulatory motifs

Figure 8: GAL4 binding sites sequence logo. [Link to sequence logo](http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html)

Figure 9: GAL4 binding. From [www.cryst.bbk.ac.uk/PPS2/](http://www.cryst.bbk.ac.uk/PPS2/)
Application 1: Identification of regulatory motifs

![Graphs showing information content profiles for GAL4, CRP, ABF1, and PURR.]

Figure 10: *Information content profiles. GAL4, CRP, ABF1, PURR.*

Application 1: COMODE


- Unaligned DNA sequences are distributed according to independent mixtures of multinomials at each position.

- Specific structural constraints on the motifs are enforced as constraints on the entropy/information content profile and/or individual entries of their position specific weight matrix (PWM).

- Estimation of motif start site and PWM involves constrained maximum likelihood estimation for a multinomial mixture model.
Application 1: COMODE

- Selecting a good model for regulatory motifs: Distribution of bases in motif? Distribution of bases in background sequence? Constraints on PWM? Motif length? Number of motifs per sequence?
- Assessing the performance of the resulting estimators.

\[ \Rightarrow \text{likelihood-based cross-validation.} \]

Examples of constraints on PWM.

- Constraints on the information content profile. E.g. parametric model such as
  \[ IC(w; \phi_1, \phi_2, w^*) = \phi_1 - |w - w^*| \tan \phi_2, \quad w = 1, \ldots, W. \]

  Structured motifs refer to binding sites with constraints on the IC of the PWM.

- Constraints on the information content of specific positions. E.g. \( IC(w) > q \) for a given \( q \) and \( w \).

- Constraints on specific nucleotide frequencies at a particular position. E.g. \( p_{w1} > 0.8 \Rightarrow \text{preference for nucleotide A at position } w \).
Application 1: COMODE

![Diagram showing parameterization for the IC profile of a motif PWM.](image)

\[ IC(w; \theta_1, \theta_2, w^*) = \theta_1 - \frac{1}{2} w - w^* \tan \theta_2, \quad w = 1, ..., W \]

Figure 11: Example of parameterization for the IC profile of a motif PWM.

Input Output

- \( n \) unaligned sequences
- Motif length \( W \)
- PWM constraint functions

- Estimated PWM
- Predicted start site for each input sequence

Available from Sündüz Keleş, [www.stat.berkeley.edu/~sunduz](http://www.stat.berkeley.edu/~sunduz)
Application 1: COMODE

$B = 100$ datasets, each comprising $n = 30$ sequences of length $L = 100$, were generated using an i.i.d. background model, with an instance of the weak motif inserted in a varying percentage ($F = 100\%, 75\%, 50\%, 25\%$) of the sequences.

Application 1: COMODE

Three different types of constraints for the motif IC profile were supplied to COMODE.

- **c.zoops-I**: piecewise linear IC profile, V-shaped (two additional parameters $\theta_1$ and $\theta_2$).
- **c.zoops-II**: ordered IC profile, first and last three positions have equal high IC, middle positions have equal low IC, HHHLLLLLLHhhh.
- **c.zoops-III**: piecewise linear IC profile, hat-shaped, mirror image of c.zoops-I (two additional parameters $\theta_1$ and $\theta_2$).

Profiles used for c.zoops-I and c.zoops-II roughly match the true IC profile, the profile for c.zoops-III is misspecified.
Application 1: COMODE

A sensitivity measure was computed as follows for each method in each of the $B$ simulated datasets
\[
\text{sens}_b = \frac{|K_b \cap \hat{K}_b|}{|K_b|},
\]
where
$K_b = \{\text{set of true motif sites in dataset } b\}$,
$\hat{K}_b = \{\text{set of predicted motif sites in dataset } b\}$.
Application 1: Likelihood CV for motif structure selection

We have applied two-fold likelihood-based cross-validation to choose among these 4 models at $F = 100\%$.

Out of the $B = 100$ datasets, c.zoops-I was selected 61 times and c.zoops-II was selected 39 times.
Application 1: Likelihood CV for motif width selection

$B = 200$ datasets, each comprising $n = 20,100$ sequences of length $L = 600$, were generated using an i.i.d. background model. A motif of width 10 was inserted in each of the sequences.

Motif start sites and PWM were estimated using COMODE with no constraints on PWM, for motif widths ranging from 6 to 15bp.

Two-fold ($p = 0.5$) and five-fold ($p = 0.2$) cross-validation were used to select motif width.

Table 5: Likelihood CV for motif width selection. Number of simulations (out of $B = 200$) each motif width was selected, for sample sizes $n = 20,100$ and using two- and five-fold CV. The true motif width is 10.
Application 2: Tree-based estimation with censored data

Tree-based estimation procedures, such as the Classification and Regression Trees (CART) of Breiman et al. (1984), can be formulated in terms of the three main steps of our roadmap and correspond to a particular choice of candidates in Step 2.

**Step 1. Loss-based definition of parameter of interest.**
The parameter of interest is defined as the risk minimizer for a particular loss function.
E.g. **Regression trees**: conditional expected value of an outcome given covariates → squared error loss function.
**Classification trees**: posterior class probabilities → indicator loss function, also Gini and negative log-likelihood (entropy).

**Step 2. Node splitting and tree pruning.**
- The sieve of candidate estimators is generated by recursive binary partitioning of a suitably defined covariate space into *nodes*, using a *loss-based node splitting rule*.  
  E.g. MSE, Gini, entropy.
- A *loss-based pruning algorithm* (minimal cost-complexity) is applied to yield a nested decreasing sequence of subtrees.  
  (Cf. *forward* selection followed by *backward* deletion.)
- For each candidate tree, an estimator is returned for each set in the resulting partition (i.e., each *terminal node*, or *leaf*) by minimizing the empirical risk.

**Step 3. Cross-validation estimator selection.**
Selection of a ’right-sized’ tree by cross-validation.
Application 2: Tree-based estimation with censored data

The outcome is a right-censored survival time.

Parameters of interest include

- conditional expected value of (log) survival time given covariates \(\rightarrow\) squared error loss function;
- conditional median of (log) survival time given covariates \(\rightarrow\) absolute error loss function;
- conditional density (survival or hazard function) of survival time given covariates \(\rightarrow\) negative log-likelihood loss function.

Problem. How to evaluate the loss function with censored data?

Common approaches for tree-based regression and density estimation bypass the risk estimation problem for censored outcomes by altering the node splitting, tree pruning, and performance assessment criteria in manners that are specific to right-censored survival times.
Application 2: Tree-based estimation with censored data

Within-node homogeneity.

- Pittman et al. (2003). Bayesian tree prediction, node splitting rule based on Bayes' factors for Weibull models. On transformed data, use exponential survival distribution and conjugate Gamma priors.

Between-node heterogeneity.

Ciampi et al. (1986) and Segal (1988) employ two-sample log-rank test statistics for between-node heterogeneity measures.

Abandoning the notion of risk (within-node homogeneity) leads to significant deviations from the standard CART framework for node splitting and tree pruning.
Application 2: Tree-based estimation with censored data

Using a loss function that is specific to the parameter of interest. One may be interested in other parameters than the conditional survival distribution, such as the conditional mean or median survival time.

In such cases, gains in accuracy may be achieved by employing a loss function that is specific to the parameter of interest (e.g., L2 or L1 loss).

Risk estimation for performance assessment. Existing methods do not provide means for assessing risk for arbitrary loss functions. Current approaches typically rely on the negative log-likelihood loss function or ignore censored observations altogether.

For any choice of full data loss function $L(X, \psi)$, one can use the above IPCW or DR-IPCW observed data loss functions $L(O, \psi | \eta_0)$ for node splitting, tree pruning, and performance assessment by cross-validation.

Note that in the absence of censoring, i.e., when $\Delta = 1$, then $L(O, \psi | \eta_0) = L(X, \psi)$ for both the IPCW and the DR-IPCW loss functions.

This ensures that the censored and full data estimators coincide when there is no censoring.
Application 2: Tree-based estimation with censored data

Estimation road map

- **Step 1.** Specify a full data loss function $L(X, \psi)$ for the parameter of interest; obtain the corresponding IPCW observed data loss function $L(O, \psi | \eta_0)$.
- **Step 2.** Apply standard node splitting and tree pruning procedures with the new IPCW loss function.
- **Step 3.** Use cross-validation with the IPCW loss function to select the right-sized tree.

Possibly bagging or boosting.

The proposed tree-based estimation procedures with the IPCW loss function can be implemented using the R `rpart` package (Therneau & Atkinson, 1997), by supplying the IPCW to the `weights` argument of the `rpart` function.

The IPCW and DR-IPCW loss functions can be used for any type of prediction method, including standard linear regression, logic regression, and bagging and boosting procedures.
Application 2: Simulation study

Comparison of survival trees built using two different loss functions for node splitting and tree pruning

- \textit{NLL\_PH}: negative log-likelihood loss function for Cox proportional hazards model (LeBlanc & Crowley, 1992), \texttt{rpart} default for survival data, \texttt{method='exp'};

- \textit{square\_IPCW}: IPCW squared error loss function, \texttt{rpart} with \texttt{method='anova'}, \texttt{weights=IPCW}.

For each loss function, obtain a final partition of the covariate space by five-fold cross-validation. Consider two within-node survival estimation methods

- IPCW mean, squared error loss function;

- Kaplan-Meier (KM) median, absolute error loss function.

Application 2: Simulation study

- Full data structure, $X = (W, Z)$: log-survival time $Z = \log T = W^2 + \epsilon$, where $W \sim U(0, 1)$, $\epsilon \sim N(0, \sigma^2)$, $\sigma^2 = 0.25$.

- Censoring variable, $C$: from uniform distributions.

- One hundred simulated learning samples were generated from an observed data distribution with 20% censoring, for sample sizes $n = 250, 600, 1250, \text{and} \ 6000$. Risk estimates are based on test samples of size $N = 5000$ generated from the full data distribution.
Application 2: Simulation study

Table 6: Ratios of average test sample risk for the *square IPCW* loss function to the *NLL PH* loss function, for two different within-node survival estimation methods.

<table>
<thead>
<tr>
<th>Sample size, n</th>
<th>Survival estimation method</th>
<th>IPCW mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>0.9422</td>
<td>0.8838</td>
</tr>
<tr>
<td>600</td>
<td>0.9524</td>
<td>0.9062</td>
</tr>
<tr>
<td>1250</td>
<td>0.9629</td>
<td>0.9244</td>
</tr>
<tr>
<td>6000</td>
<td>0.9767</td>
<td>0.9533</td>
</tr>
</tbody>
</table>

**N. B.** Ratios less than one correspond to improved accuracy for trees based on *IPCW* loss function — Risk *square IPCW*/Risk *NLL PH*.

Application 2: Breast cancer survival and CGH copy number

Comparative genomic hybridization (CGH) is a microarray-based technique for measuring genome-wide DNA copy number.

DNA copy number alterations have been linked to a number of cancers: gains can over-express oncogenes, losses can inactivate tumor suppressor genes.

In cancer research, CGH analysis produces thousands of DNA copy number measurements for each patient, in addition to epidemiological, histological, and pathological variables.

*Predict clinical outcome from thousands of explanatory variables.*
Application 2: Breast cancer survival and CGH copy number

CGH study of breast cancer patients (Waldman et al., in preparation).

- 152 patients, all with initial occurrences of breast cancer (invasive ductal carcinoma).
- **Outcome**: Time to recurrence (in months) — 52 patients recurred, censoring percentage of 66%.
- **Explanatory variables**:
  - epidemiological variables (e.g., age at diagnosis, race),
  - histopathological variables (e.g., tumor stage, grade),
  - and DNA copy number measures from a CGH array with 2,254 bacterial artificial chromosomes (BAC).

The 152 observations were split at random into a learning set and a test set of 128 and 24 observations, respectively.

- Trees were grown using the learning set with the IPCW squared error loss function.
- Five-fold cross-validation was used to select the 'best' tree.
- The survival function $\tilde{G}_0$ in the IPCW loss function was estimated separately for each training sample by fitting a Cox proportional hazards model to the epidemiological and histopathological variables (\texttt{coxph} function).
- Overall performance was assessed on the test sample.
Application 2: Breast cancer survival and CGH copy number

The selected two-split tree is based on BACs that fall in chromosomal regions known to contain genes related to breast cancer.

This tree suggests that copy number gains in both regions are associated with longer survival.

Improved prediction accuracy and more information on chromosomal regions related to breast cancer survival may be obtained from aggregation methods such as bagging and boosting and from more aggressive strategies for generating candidate estimators.

Figure 13: *Breast cancer survival and CGH copy number dataset.* Learning set survival tree, IPCW mean log survival time (in months).
Application 2: Summary

- The choice of loss function for node splitting, tree pruning, and within node estimation can have a large impact on accuracy.
- Gains in accuracy are obtained by using a loss function that is specific to the parameter of interest.

Ongoing work

- More extensive study of the properties of different loss functions for multivariate outcome prediction and density estimation (Step 1).
- Loss-based variable importance statistics.
- R package.
References
www.bepress.com/ucbbiostat/
www.bepress.com/sagmb/