

Unified Cross-validation Methodology for Estimator Selection and Applications to Genomics

Sandrine Dudoit

Division of Biostatistics, UC Berkeley

`www.stat.berkeley.edu/~sandrine`

Statistics and Genomics Seminar, UC Berkeley

April 10, 2003

©Copyright 2003, all rights reserved

Acknowledgments

Joint work with

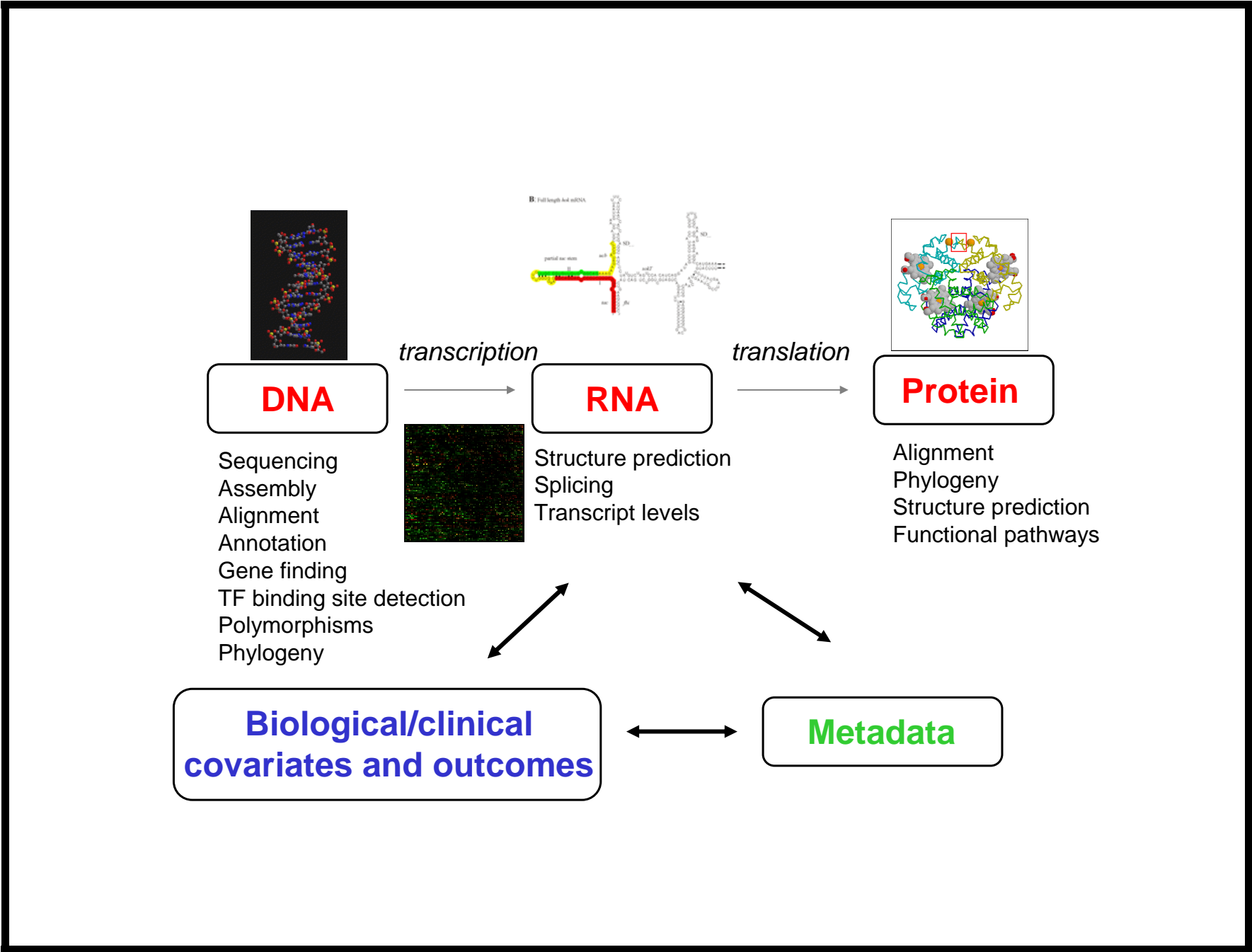
Mark van der Laan, Division of Biostatistics, UC Berkeley.

Sündüz Keleş, Division of Biostatistics, UC Berkeley.

Matthieu Cornec, Department of Statistics, UC Berkeley.

Outline

- Motivation: estimator selection and performance assessment in genomic data analysis.
- The estimator selection problem.
- Unified framework for cross-validation.
- Three examples.
- Properties of the cross-validation selector: finite sample and asymptotic optimality results.
- Asymptotic linearity of cross-validation risk estimator and risk confidence intervals.
- Simulation studies.
- Ongoing work: software development and applications.



Motivation: estimator selection/assessment in genomics

Problem 1. *Prediction of biological and clinical outcomes using microarray gene expression measures.*

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** (DNA or RNA abundance) on a genomic scale.

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

Statistical question. Relate microarray gene expression measures to biological and clinical outcomes.

Motivation: estimator selection/assessment in genomics

- Outcomes (phenotypes): tumor class, response to treatment, patient survival, affectedness/unaffectedness, etc. — polychotomous or continuous; censored or uncensored.
- Explanatory variables (genotypes): gene expression measures, age, sex, treatment, etc. — polychotomous or continuous.
- Selecting a *good* predictor: LDA, CART, SVMs, neural networks?
- Selecting a *good* subset of marker genes for building this predictor: 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: estimator selection/assessment in genomics

Problem 2. *Identification of regulatory motifs in DNA sequences.*

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

The 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 to 1,000 base pairs from the gene start site (in lower eukaryotes, e.g., yeast).

Motivation: estimator selection/assessment in genomics

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

>YBR019C TCGGCGATACCTTCA~~CCG~~

>YBR020W ~~CGGG~~CGACGATTAC~~CCG~~

>YLR081W TAT~~CGG~~AGCGTAGGCGG~~CCG~~AAC

>YML051W ~~CGG~~CATCCTACATG~~CCG~~

>YOR120W T~~CGG~~TTCAGACAGGT~~CCG~~G

Motivation: estimator selection/assessment in genomics

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

- Common approaches are based on estimating the likelihood of (posterior probabilities given) the DNA sequence data under certain models for the distribution of bases in the binding sites and background sequence.

E.g. Bailey & Elkan (1994), Kechris et al. (2002), Keleş et al. (2002), Lawrence & Reilly (1990).

- Selecting a *good* model for binding sites: Length of motifs? Constraints on PWM?
- Assessing the performance of the resulting estimators.

The estimator selection problem

The main ingredients

- An unknown **population distribution**, P_0 .
- An unknown **parameter** of interest, $\psi_0(\cdot) = \psi_0(\cdot | P_0)$ (i.e., function of P_0), with parameter space Ψ .
- A **loss function**, $L(O, \psi | \eta_0)$, for a candidate ψ , possibly depending on a nuisance parameter $\eta_0 = \eta_0(P_0)$.

The expected loss, or **risk**, is **minimized by the optimal ψ_0**

$$\begin{aligned}\psi_0 &\equiv \operatorname{argmin}_{\psi \in \Psi} \int L(o, \psi | \eta_0) dP_0(o) \\ &= \operatorname{argmin}_{\psi \in \Psi} E_0[L(O, \psi | \eta_0)].\end{aligned}$$

E.g. Squared error loss, ψ_0 is a population mean.

The estimator selection problem

- A **sample**, or **learning set**, of n independent and identically distributed (i.i.d.) observations O_1, \dots, O_n , with $O_i \sim P_0$.
- Let P_n be the **empirical distribution** of O_1, \dots, O_n .
- Let $\hat{\psi}_k(\cdot) = \psi_k(\cdot | P_n) \in \Psi$, $k = 1, \dots, K_n$, be a collection of **estimators** of $\psi_0(\cdot)$, i.e., functions/algorithms one can apply to data.

The estimator selection problem

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

$$\begin{aligned} d_n(\hat{\psi}_{\hat{k}}, \psi_0) &\equiv \int \{L(o, \psi_{\hat{k}}(\cdot | P_n) | \eta_0) - L(o, \psi_0(\cdot) | \eta_0)\} dP_0(o) \\ &\longrightarrow 0 \end{aligned}$$

at asymptotically optimal rate.

The estimator selection problem

The optimal benchmark selector. Let

$$\begin{aligned}\tilde{k}_n &\equiv \operatorname{argmin}_k d_n(\hat{\psi}_k, \psi_0) \\ &= \operatorname{argmin}_k \int L(o, \psi_k(\cdot | P_n) | \eta_0) dP_0(o).\end{aligned}$$

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 $\hat{k} = \hat{k}(P_n)$ is **asymptotically optimal** if

$$\frac{d_n(\hat{\psi}_{\hat{k}}, \psi_0)}{d_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)} \longrightarrow 1 \text{ in probability as } n \longrightarrow \infty.$$

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 $\hat{\psi}_k(\cdot) = \psi_k(\cdot | P_n) \in \Psi$, $k = 1, \dots, K_n$.

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 $\hat{\psi}_{\hat{k}}$ is chosen to have the best performance on the validation set.

General framework for cross-validation

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 $S_n = (S_{n,1}, \dots, S_{n,n})$ defines a particular split of the learning sample of n observations into a training set and validation set

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

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.

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

$$\hat{k} \equiv \operatorname{argmin}_k E_{S_n} \int L(o, \underbrace{\psi_k(\cdot \mid P_{n,S_n}^0)}_{\text{Training}} \mid \underbrace{\eta_{n,S_n}^0}_{\text{Validation}}) dP_{n,S_n}^1(o).$$

Here, $\psi_k(\cdot \mid P_{n,S_n}^0)$ and η_{n,S_n}^0 denote, respectively, estimators for the parameter of interest ψ_0 and the nuisance parameter η_0 , using only the training set.

General framework for cross-validation

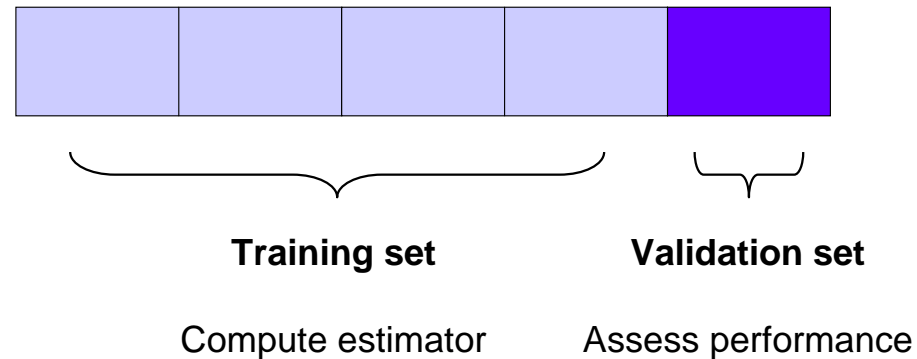


Figure 1: *Five-fold cross-validation*. S_n has 5 realizations.

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 $s_n = (s_{n,1}, \dots, s_{n,n})$ 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 $s_n^v = (s_{n,1}^v, \dots, s_{n,n}^v)$, $v = 1, \dots, V$, such that $\sum_i s_{n,i}^v \approx n/V$ and $\sum_v s_{n,i}^v = 1$ ($p_n = 1/V$).

General framework for cross-validation

- **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 $1/\binom{n}{n_1}$ 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 in the microarray literature to

compare the performance of different predictors of biological and clinical outcomes;

support statements such as

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

Honest cross-validation

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) using **all the learning set** and then perform cross-validation only on the predictor building portion of the process.

- ⇒ The reported error rates are usually **biased downward** and give an overly optimistic view of the predictive power of microarray expression measures.
- ⇒ Predictors are not compared on an equal footing.

Honest cross-validation

Estimates of prediction error (risk) from cross-validation (or other procedures) relate **only** to the experiment that was cross-validated.

Cross-validation should be done on the **entire predictor training process**, including gene screening and predictor parameter selection.

Ref. Ambroise & McLachlan (2002), Dudoit & Fridlyand (2003), West et al. (2001).

Examples

Our general framework and results for estimator selection using cross-validation allow us to handle a broad range of problems that have traditionally been treated separately in the statistical literature.

1. Predictor selection.
2. Density estimator selection.
3. Predictor selection based on right-censored outcomes.
4. Survival function estimator selection.
5. Predictor selection for multivariate outcomes.
6. Counterfactual predictor selection in causal inference.

van der Laan & Dudoit (2003)

Example 1: Predictor selection

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

Consider the **quadratic loss function**

$$L(O, \psi) = (Y - \psi(W))^2.$$

The parameter of interest, which minimizes the risk

$$E_0[L(O, \psi)] = \int (y - \psi(w))^2 dP_0(o),$$

is the conditional expectation $\psi_0(W) = E_0[Y | W]$.

Example 1: Predictor selection

Given candidate predictors $\hat{\psi}_k = \psi_k(\cdot | P_n)$, the risk difference for the quadratic loss simplifies to

$$d_n(\hat{\psi}_k, \psi_0) = \int (\psi_k(w | P_n) - \psi_0(w))^2 dF_{W,0}(w).$$

The cross-validation selector is given by

$$\begin{aligned} \hat{k} &= \operatorname{argmin}_k E_{S_n} \int (y - \psi_k(w | P_{n,S_n}^0))^2 dP_{n,S_n}^1(o) \\ &= \operatorname{argmin}_k E_{S_n} \sum_{\{i: S_{n,i}=1\}} (Y_i - \psi_k(W_i | P_{n,S_n}^0))^2. \end{aligned}$$

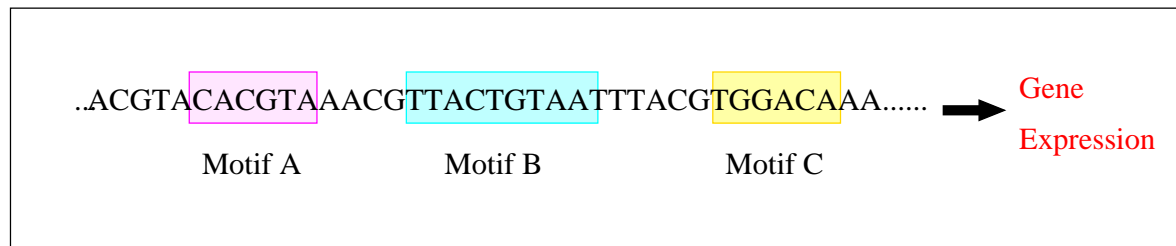
Example 1: Predictor selection

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

- Outcomes (phenotypes), Y : tumor class, response to treatment, patient survival, affectedness/unaffectedness, etc. — polychotomous or continuous; see Example 3, below, for censored outcomes.
- Explanatory variables (genotypes), W : microarray gene expression measures, SNP haplotypes, age, sex, treatment, etc. — polychotomous or continuous.
- Which predictor: linear models, generalized linear models, classification and regression trees, support vector machines, neural networks?
- How many genes? Which genes?

Example 1: Predictor selection

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



- Outcomes (phenotypes), Y : microarray gene expression measures — multivariate outcomes.
- Explanatory variables (genotypes), W : DNA sequence in upstream control region of genes.

Keleş 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, $\hat{\psi}_k = f_k(\cdot | P_n)$, of $\psi_0 = f_0$, the risk difference is the **Kullback-Leibler divergence** between $f_k(\cdot | P_n)$ and f_0

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

The cross-validation selector is given by

$$\begin{aligned} \hat{k} &= \operatorname{argmin}_k - E_{S_n} \int \log f_k(o | P_{n,S_n}^0) dP_{n,S_n}^1(o) \\ &= \operatorname{argmin}_k - E_{S_n} \sum_{\{i: S_{n,i}=1\}} \log f_k(O_i | P_{n,S_n}^0). \end{aligned}$$

Example 2: Density estimator selection

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

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

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)

$$\begin{aligned} & - \int \log f_k(y; w | P_n) f_0(o) d\mu(o) \\ &= - \int \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (y - \psi_k(w|P_n))^2\right) \right\} f_0(o) d\mu(o) \\ &= \int (y - \psi_k(w|P_n))^2 f_0(o) d\mu(o). \end{aligned}$$

Example 2: Density estimator selection

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 at position w is

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

The information content profile of a PWM is a measure of a site's tolerance for substitution: high IC, low tolerance.

Example 2: Density estimator selection

- 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).

Example 2: Density estimator selection

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

>YBR019C TCGGCGATACCTTCA~~CCG~~

>YBR020W ~~CGGG~~CGACGATTAC~~CCG~~

>YLR081W TAT~~CGG~~AGCGTAGGGCGG~~CCG~~AAC

>YML051W ~~CGG~~CATCCTACATG~~CCG~~

>YOR120W T~~CGG~~TTCAGACAGGT~~CCG~~G

Example 2: Density estimator selection

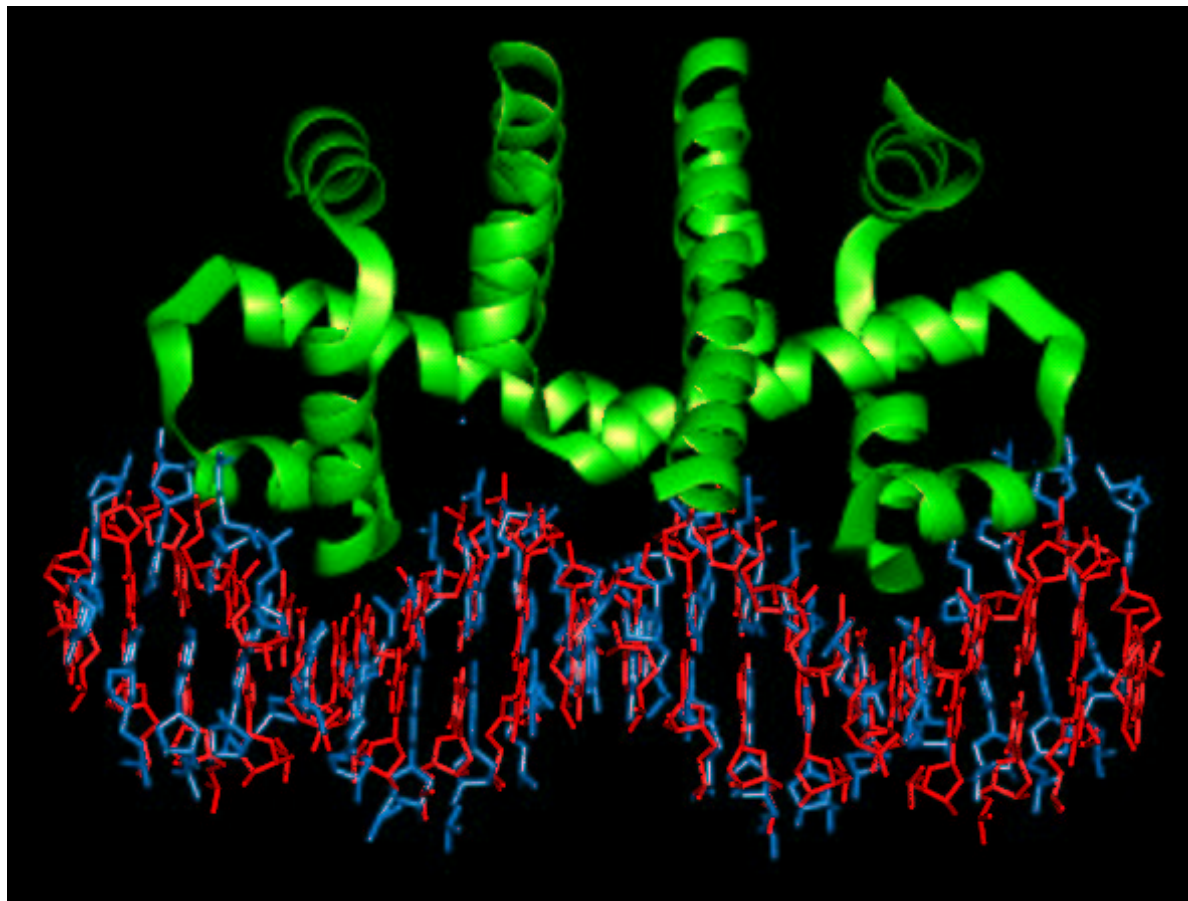


Figure 2: *GAL4 binding*. From www.cryst.bbk.ac.uk/PPS2/.

Example 2: Density estimator selection

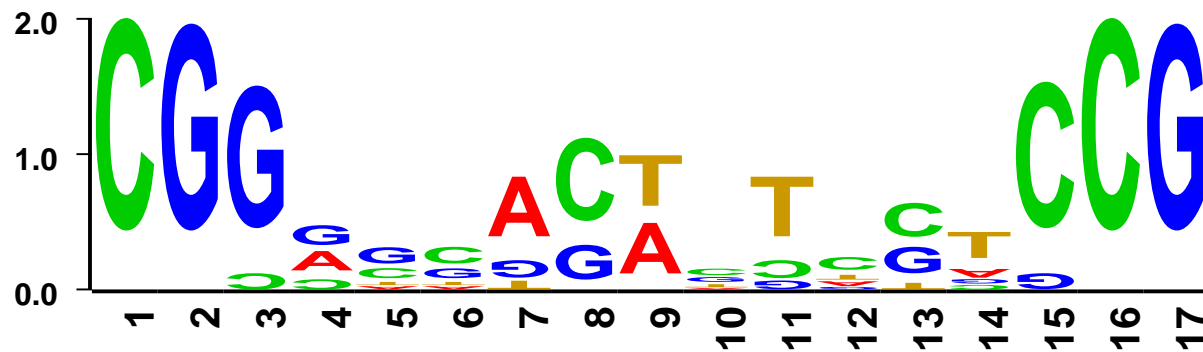


Figure 3: *GAL4* sequence logo. From Schneider et al. (1990)

Example 2: Density estimator selection

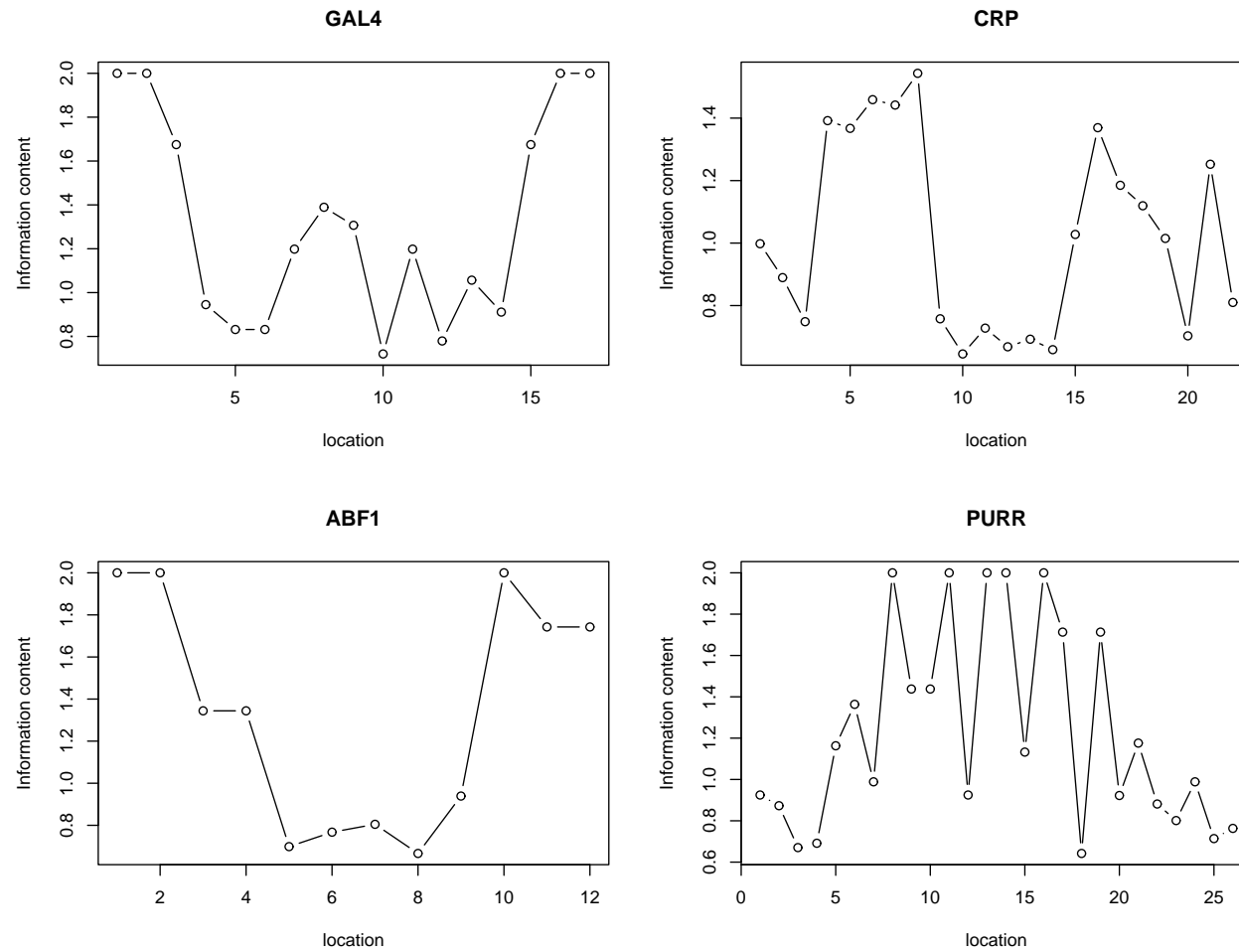


Figure 4: *Information content profiles.* GAL4, CRP, ABF1, PURR.

Example 2: Density estimator selection

Keleş et al. (2002). Likelihood-based method for detecting structured regulatory motifs in biological sequences.

- Unaligned DNA sequences are distributed according to independent mixtures of multinomials at each base.
- Specific structural constraints on the motifs are enforced as constraints on the entropy/information content profile of their position specific weight matrix (PWM).
- Estimation of motif start site and PWM involves constrained maximum likelihood estimation for a mixture of multinomials model.
- Selecting a *good* model for binding sites: Length of motifs?
Constraint on the IC profile of PWM?
⇒ likelihood-based cross-validation.

Example 2: Density estimator selection

INPUT

N sequences
Motif length W
Constraint function

OUTPUT

PWM
Start site prediction
on each sequence

Example 3: Predictor selection for right-censored outcomes

Let $X = (Y, W) \sim F_{X,0}$ be the full data structure of interest, where $Y = \log(T)$ is a log survival time and W a vector of explanatory variables (covariates).

Let C be a right-censoring time, with conditional distribution $G_0(\cdot | X)$. Assume $C \perp Y$, given W .

Suppose we have a learning set of n i.i.d. observations of the right-censored data structure

$$O = \left(\min(Y, C), \Delta = I(Y \leq C), W \right) \sim P_0 = P_{F_{X,0}, G_0}.$$

Example 3: Predictor selection for right-censored outcomes

Consider the quadratic loss function

$$L(X, \psi) = L_2(X, \psi) = (Y - \psi(W))^2.$$

The parameter of interest, which minimizes the risk for this loss function, is the conditional expectation $\psi_0(W) = E_0[Y | W]$.

$$\begin{aligned}\psi_0 &= \operatorname{argmin}_{\psi} \int L_2(x, \psi) dF_{X,0}(x) \\ &= \operatorname{argmin}_{\psi} E_{F_{X,0}}(Y - \psi(W))^2 \\ &= \operatorname{argmin}_{\psi} E_{P_0} \left\{ L_2(X, \psi) \frac{\Delta}{\bar{G}_0(Y | X)} \right\}.\end{aligned}$$

Example 3: Predictor selection for right-censored outcomes

General problem. The loss function is a function of the full data structure $X = (Y, W)$ — unobservable.

Solution. The general [estimating function methodology](#) for censored data of van der Laan & Robins (2002) maps full data estimating functions $D(X)$ into observed data estimating functions $IC(O | Q(F_X), G, D)$, indexed by nuisance parameters G and (possibly) $Q(F_X)$. The estimating functions satisfy

$$E_{P_0} IC(O | Q, G, D) = E_{F_{X,0}} D(X) \quad \text{if } G = G_0 \text{ or } Q = Q_0.$$

Thus, we can choose the following loss function for the observable right-censored data structure O

$$L(O, \psi | \eta_0 = (Q_0, G_0)) = IC(O | Q_0, G_0, L_2(\cdot, \psi)).$$

Example 3: Predictor selection for right-censored outcomes

Inverse probability of censoring weighted (IPCW) estimating function

$$IC(O | G, D) = D(X) \frac{\Delta}{\bar{G}(Y | X)}.$$

For candidate predictors $\hat{\psi}_k = \psi_k(\cdot | P_n)$, the cross-validation selector based on the IPCW estimating function is given by

$$\hat{k} = \operatorname{argmin}_k E_{S_n} \sum_{\{i: S_{n,i}=1\}} (Y_i - \psi_k(W_i | P_{n,S_n}^0))^2 \frac{\Delta_i}{\bar{G}_{n,S_n}^0(Y_i | W_i)}.$$

Example 3: Predictor selection for right-censored outcomes

Given candidate predictors $\hat{\psi}_k = \psi_k(\cdot | P_n)$, the corresponding risk difference for the quadratic loss simplifies to

$$\begin{aligned}d_n(\hat{\psi}_k, \psi_0) &= \int L(o, \hat{\psi}_k | \eta_0) - L(o, \psi_0 | \eta_0) dP_0(o) \\ &= \int L_2(x, \hat{\psi}_k) - L_2(x, \psi_0) dF_{X,0}(x) \\ &= \int (\psi_k(w | P_n) - \psi_0(w))^2 dF_{W,0}(w).\end{aligned}$$

Example 3: Predictor selection for right-censored outcomes

E.g. 1. Predicting survival of cancer patients based on microarray gene expression profile of cancer tissue.

E.g. 2. Predicting survival of AIDS patients from DNA sequence of HIV virus.

Properties of cross-validation selector

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

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

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

$$\tilde{k}_{n(1-p)} \equiv \operatorname{argmin}_k d_{n(1-p)}(\hat{\psi}_k, \psi_0).$$

Properties of cross-validation selector

Theorem 1. (Stated in special case of known η_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

$$\begin{aligned} & \int \{L(o, \psi_k(\cdot | P_{n, S_n}^0)) - L(o, \psi_0(\cdot))\}^2 dP_0(o) \\ & \leq M_2 \int L(o, \psi_k(\cdot | P_{n, S_n}^0)) - L(o, \psi_0(\cdot)) dP_0(o) \text{ a.s.} \end{aligned}$$

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

$$\begin{aligned} 0 \leq E d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0) & \leq (1 + 2\delta) E d_{n(1-p)}(\hat{\psi}_{\tilde{k}_{n(1-p)}}, \psi_0) \\ & + C(M_1, M_2, \delta) \frac{1 + \log(K_n)}{np}. \end{aligned}$$

Properties of cross-validation selector

Asymptotic optimality. If

$$\frac{\log(K_n)}{(np) Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}_{n(1-p)}}, \psi_0)} \longrightarrow 0, \quad \text{as } n \rightarrow \infty,$$

then

$$\frac{Ed_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{Ed_{n(1-p)}(\hat{\psi}_{\tilde{k}_{n(1-p)}}, \psi_0)} \longrightarrow 1, \quad \text{as } n \rightarrow \infty.$$

Properties of cross-validation selector

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

$$\frac{\log(K_n)}{(np) \text{Ed}_{n(1-p)}(\hat{\psi}_{\tilde{k}_{n(1-p)}}, \psi_0)} \longrightarrow 0,$$

and

$$\frac{\text{Ed}_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)}{\text{Ed}_{n(1-p)}(\hat{\psi}_{\tilde{k}_{n(1-p)}}, \psi_0)} \longrightarrow 1.$$

Then,

$$\frac{\text{Ed}_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{\text{Ed}_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)} \longrightarrow 1, \quad \text{as } n \rightarrow \infty.$$

That is, the data adaptive CV selector \hat{k} is **asymptotically optimal**.

Properties of cross-validation selector

- 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 η_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)

Properties of cross-validation selector

- Both theorems consider general distributions of S_n , i.e., **general cross-validation procedures** with an arbitrary proportion p_n of observations included in the validation sets.
- The finite sample results hold for any p_n , while the asymptotic results require that $np_n \rightarrow \infty$; the later condition rules out LOOCV.
- The theorems apply to **general distributions P_0 , general loss functions $L(O, \psi | \eta_0)$, and general estimators $\psi(\cdot | P_n)$** .

Examples

1. Predictor selection.
2. Density estimator selection.
3. Predictor selection based on right-censored outcomes.
4. Survival function estimator selection.
5. Predictor selection for multivariate outcomes.
6. Counterfactual predictor selection in causal inference.

van der Laan & Dudoit (2003)

Asymptotic linearity of CV risk estimator

Consider a particular estimator $\hat{\psi}(\cdot) = \psi(\cdot | P_n)$ and loss function $L(O, \psi | \eta_0) = L(O, \psi)$ with known η_0 .

Cross-validation risk estimator (observable random variable)

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

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

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

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

$$E_{S_n} \sqrt{\frac{\int \left\{ L(o, \psi(\cdot | P_{n, S_n}^0)) - L(o, \psi(\cdot | P_0)) \right\}^2 dP_0(o)}{p_n}} = o_P(1).$$

Then

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

Risk confidence intervals

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

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

where

$$\hat{\sigma}_n^2 = \int (IC(o | P_n))^2 dP_n(o),$$

$$IC(o | P_n) = L(o, \psi(\cdot | P_n)) - \int L(o, \psi(\cdot | P_n)) dP_n(o),$$

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

Simulation study 1: Likelihood cross-validation

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 $[-2, 2]$.
- 20 replicate datasets were generated from f_0 for six different sample sizes, $n = 50, 100, 200, 400, 800, 1600$.
- The Gaussian kernel density estimate, $\hat{f}_k(\cdot) = f_k(\cdot | P_n)$, for a learning set x_1, \dots, 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 $\phi(\cdot)$ 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.

Simulation study 1: Likelihood cross-validation

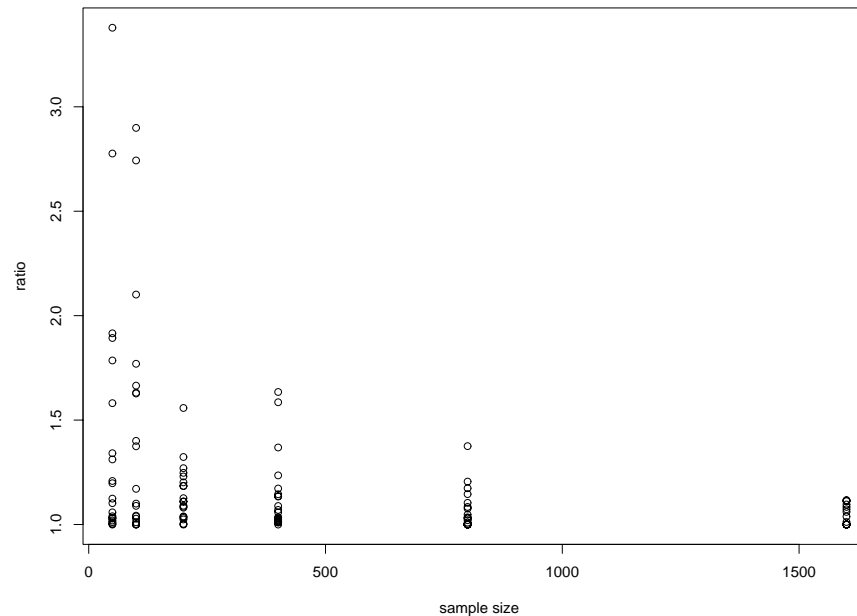


Figure 5: $\frac{d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{d_{n(1-p)}(\hat{\psi}_{\tilde{k}_{n(1-p)}}, \psi_0)}$ vs. n , for $p = 1/10$. The bandwidth \hat{k} was selected using ten-fold CV ($p = 1/10$), for 20 replicate datasets at each of six sample sizes, n .

Simulation study 1: Likelihood cross-validation

n					
50	100	200	400	800	1600
1.542497	1.400015	1.150882	1.139386	1.068780	1.033064

Table 1: $\frac{\hat{E}d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{\hat{E}d_{n(1-p)}(\hat{\psi}_{\hat{k}_{n(1-p)}}, \psi_0)}$ 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 ($p = 1/10$).

Simulation study 1: Likelihood cross-validation

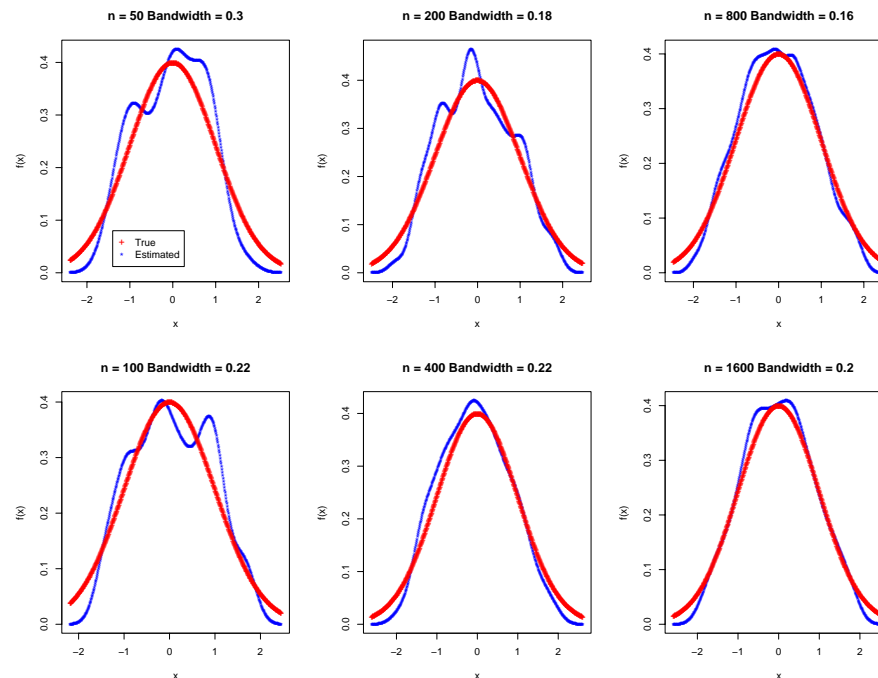


Figure 6: *Cross-validation density estimates $\hat{f}_{\hat{k}}$ and true density f_0 . The cross-validation kernel density estimate $\hat{f}_{\hat{k}}(\cdot | P_n)$ is shown for six sample sizes, $n = 50, 100, 200, 400, 800, 1600$, for one simulated dataset. The bandwidth \hat{k} was selected using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

Simulation study 1: Likelihood cross-validation

		n					
		50	100	200	400	800	1600
p	0.05	1.493594	1.465201	1.168274	1.115338	1.089441	1.047685
	0.1	1.531736	1.391971	1.144236	1.136916	1.075563	1.048454
	0.15	1.577241	1.473550	1.118831	1.117599	1.076197	1.061919
	0.20	1.518429	1.417260	1.120498	1.100698	1.065835	1.064060
	0.25	1.302580	1.443560	1.111674	1.182325	1.060759	1.100572
	0.30	1.430726	1.388704	1.148916	1.119423	1.080356	1.083632
	0.35	1.238741	1.414966	1.076628	1.093445	1.092477	1.112602
	0.40	1.477980	1.617694	1.200306	1.123990	1.091412	1.091008
	0.45	1.411283	1.483116	1.090528	1.142125	1.134810	1.143657
	0.50	1.320979	1.398095	1.099359	1.136470	1.146952	1.167325

Table 2: V -fold likelihood cross-validation: $\frac{\hat{E}d_n(\hat{\psi}_{\hat{k}(p)}, \psi_0)}{\hat{E}d_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)}$ 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 $p = 1/V$.

Simulation study 1: Likelihood cross-validation

		n					
		50	100	200	400	800	1600
	0.05	21.778985	30.591547	5.366258	3.488738	2.147304	1.287172
	0.1	4.969151	8.139912	3.709904	2.105173	1.948626	1.291611
	0.15	1.972465	5.234631	2.283455	1.831317	1.628340	1.153562
	0.20	1.836114	10.036376	2.465654	1.377272	1.370639	1.093183
	0.25	2.495359	4.262036	1.246727	1.232388	1.209813	1.092931
p	0.30	2.260952	4.298054	1.410498	1.149826	1.215430	1.123646
	0.35	1.553013	3.862468	1.511450	1.111143	1.165148	1.151871
	0.40	1.446852	1.615702	1.276998	1.123451	1.146859	1.113719
	0.45	1.583617	1.757668	1.263186	1.170124	1.112150	1.133443
	0.50	1.333555	2.193936	1.258745	1.164263	1.149889	1.175700

Table 3: *Single-split likelihood cross-validation*: $\frac{\hat{E}d_n(\hat{\psi}_{\hat{k}(p)}, \psi_0)}{\hat{E}d_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)}$ 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 .

Simulation study 1: Likelihood cross-validation

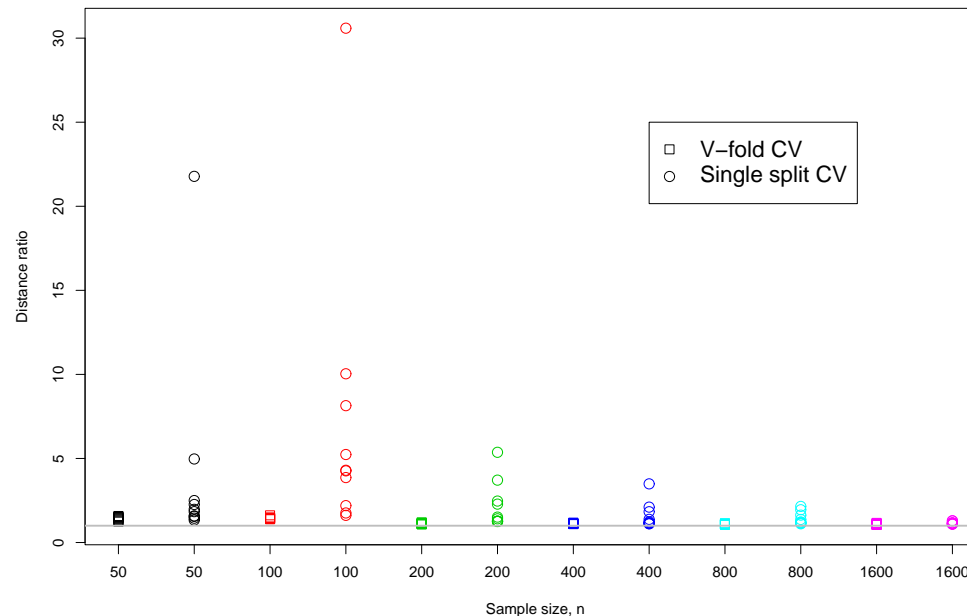


Figure 7: *V-fold vs. single split CV*: $\frac{\hat{E}d_n(\hat{\psi}_{\hat{k}(p)}, \psi_0)}{\hat{E}d_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)}$ vs. n . Estimated distance ratios are based on 20 replicate datasets at six different sample sizes n and for ten different validation set proportions p .

Simulation study 2: Prediction based on censored outcomes

Cross-validation for bin width selection in histogram regression on right-censored outcomes.

- The full data structure is $X = (Y, W)$, where $W \sim U(0, 1)$ and $Y = \log T = W^2 + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, $\sigma^2 = 2$, enforced compact support in the interval $[-10, 10]$.
- Censoring times C are generated from an Exponential(λ) distribution.
- 50 replicate datasets were generated for sample sizes $n = 50, 100, 200, 400, 800, 1600$.
- $K_n = 100$ different bin widths were considered. For $k = 1, \dots, K_n$, the unit interval is divided into k bins with width $1/k$ each.

Simulation study 2: Prediction based on censored outcomes

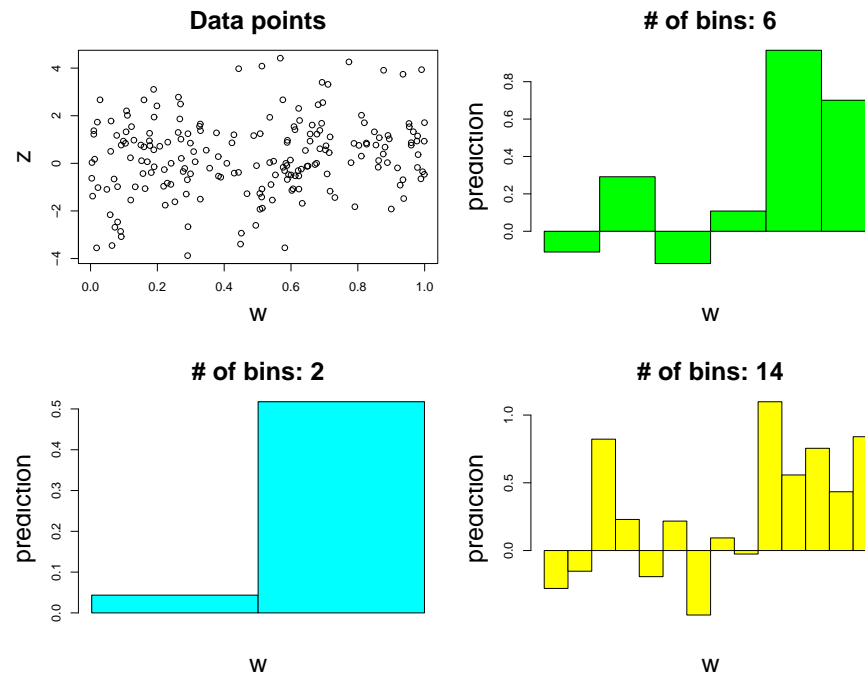


Figure 8: *Histogram regression*. Predictors are indexed by the number of bins and the prediction for a given bin is the mean outcome for observations in that bin.

Simulation study 2: Prediction based on censored outcomes

- For k -bin histogram regression and for a particular training sample P_{n,S_n}^0 , let $B_j(P_{n,S_n}^0)$ denote the set of observations in the j th bin, $[(j-1)/k, j/k)$, $j = 1, \dots, k$,

$$B_j(P_{n,S_n}^0) = \left\{ i : S_{n,i} = 0, W_i \in [(j-1)/k, j/k) \right\}.$$

For $w \in [(j-1)/k, j/k)$, the predicted log survival time is

$$\psi_k(w \mid P_{n,S_n}^0) = \frac{1}{|B_j(P_{n,S_n}^0)|} \sum_{i \in B_j(P_{n,S_n}^0)} \frac{(\log T_i) \Delta_i}{\bar{G}_{n,S_n}^0(T_i \mid W_i)},$$

where $\bar{G}_{n,S_n}^0(\cdot \mid W)$ is the Kaplan-Meier estimator of $\bar{G}(\cdot \mid W)$.

- Bin widths were selected by ten-fold cross-validation ($p = 1/10$).

Simulation study 2: Prediction based on censored outcomes

		Censoring proportion		
		0%	10%	20%
	50	6.578537	7.133457	7.846112
	100	1.100901	1.333004	1.974709
	200	1.022957	1.199649	1.418739
n	400	1.013431	1.137665	1.255642
	800	1.010221	1.119677	1.155544
	1600	1.003344	1.071642	1.107322

Table 4: *Ten-fold cross-validation.* $\frac{d_n(\hat{\psi}_{\hat{k}}, \psi_0)}{d_n(\hat{\psi}_{\tilde{k}_n}, \psi_0)}$ vs. n for different censoring proportions ($\lambda = 0.07$ and 0.15 for 10% and 20% censoring, respectively).

Simulation study 3: Consistency and asymptotic linearity

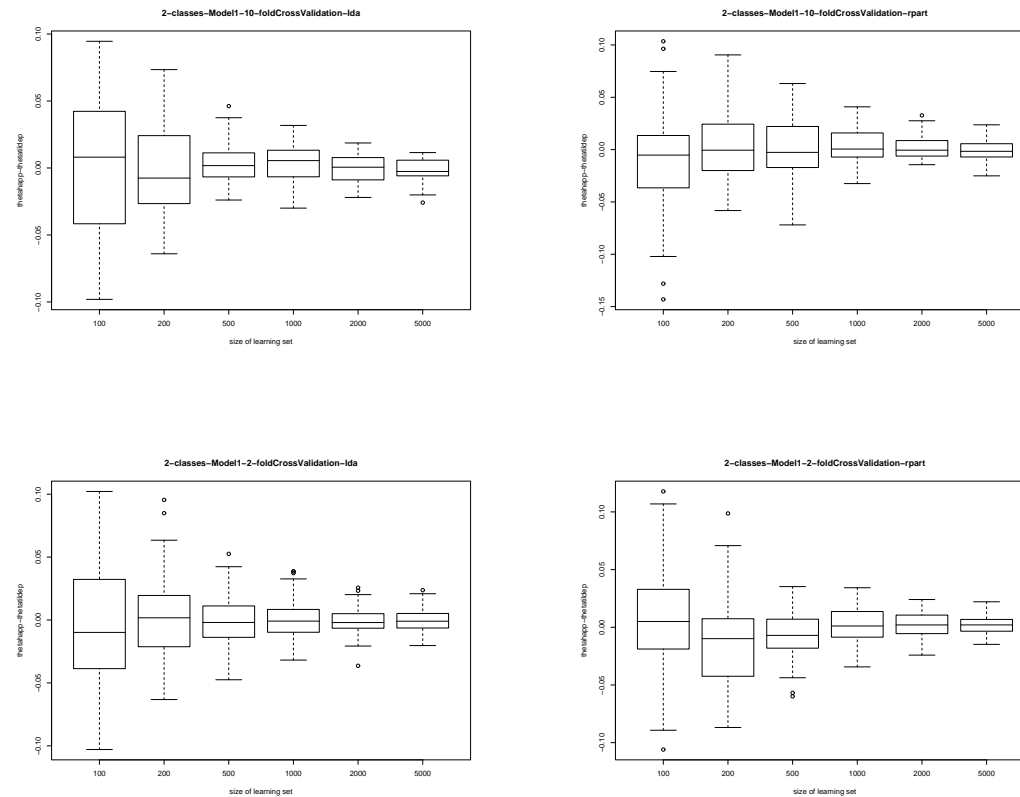


Figure 9: *Convergence to zero of $\hat{\theta}_{n(1-p)} - \tilde{\theta}_{n(1-p)}$. $X|Y \sim N(Y1_2, I_2)$, $Y \sim B(1/2)$, LDA, rpart, two- and ten-fold CV, 200 simulations.*

Simulation study 3: Risk confidence intervals

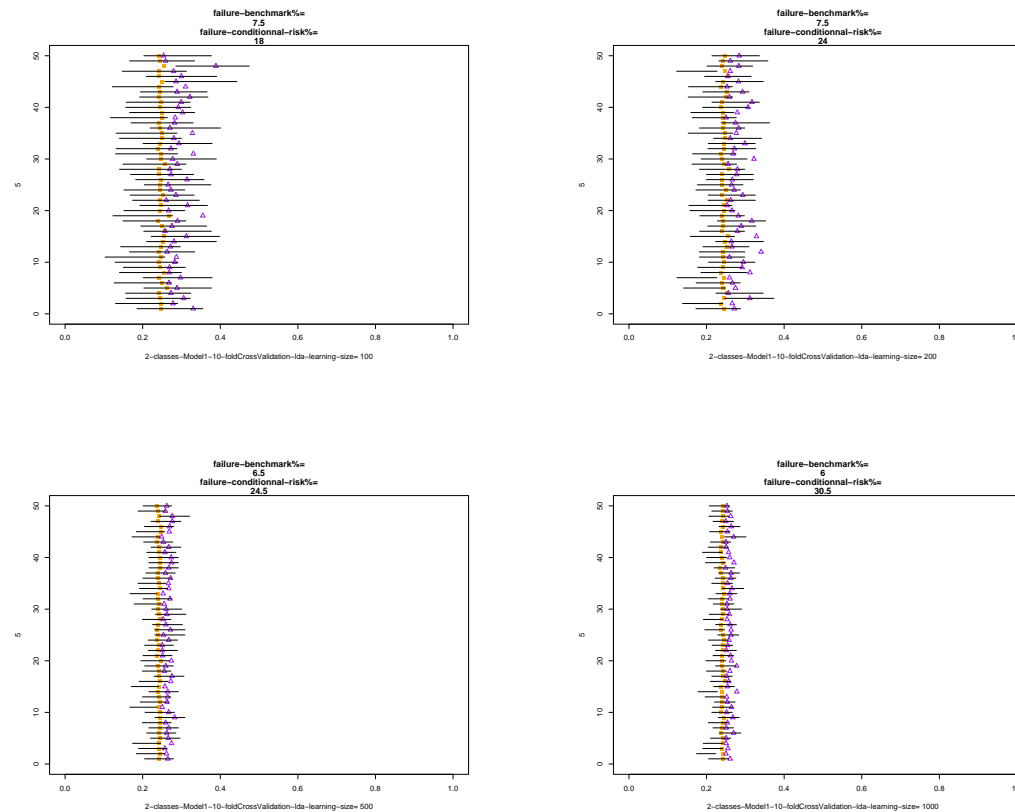


Figure 10: *Risk confidence intervals.* $X|Y \sim N(Y1_2, I_2)$, $Y \sim B(1/2)$, $n = 100, 200, 500, 1000$, LDA, ten-fold CV, $\tilde{\theta}_{n(1-p)}$ and $\tilde{\theta}_n$.

Ongoing work

- R software package for cross-validation risk estimation and estimator selection.

`www.bioconductor.org`.

- Regression trees for right-censored outcomes: general risk estimation methodology (Annette Molinaro).
- Application to prediction and variable selection in microarray experiments.
- Application to the identification of structured motifs in biological sequences (Sündüz Keleş).

References

www.bepress.com/ucbbiostat/

- S. Keleş, M. J. van der Laan, and S. Dudoit (2003). Asymptotically Optimal Model Selection Method for Regression on Censored Outcomes. Division of Biostatistics, UC Berkeley, Technical Report #124.
- M. J. van der Laan, S. Dudoit, and S. Keleş (2003). Asymptotic Optimality of Likelihood Based Cross-validation. Division of Biostatistics, UC Berkeley, Technical Report #125.
- S. Dudoit and M. J. van der Laan (2003). Asymptotics of Cross-Validated Risk Estimation in Model Selection and Performance Assessment. Division of Biostatistics, UC Berkeley, Technical Report #126.
- M. J. van der Laan and S. Dudoit (2003). Unified Cross-validation Methodology for Selection among Estimators: Finite Sample Results, Asymptotic Optimality, and Applications. Division of Biostatistics, UC Berkeley, Technical Report #130.
- S. Keleş, M. J. van der Laan, S. Dudoit, B. Xing, and M. B. Eisen (2002). Detecting regulatory motifs with entropy constraints in DNA sequences.
- S. Keleş, M. J. van der Laan, and M. B. Eisen (2002). Identification of Regulatory Elements Using a Feature Selection Method. *Bioinformatics*, **18**: 1167–1175.