Multivariate Statistical Methods in Genomics

PH 243 A, MW 12-2, 2305 Tolman

Instructor: Mark van der Laan

Office: Haviland Hall 108, tel: 643-9866

website: www.stat.berkeley.edu/ laan

Technical reports at www.bepress.com/ucbbiostat/

email: laan@stat.berkeley.edu

General Topics Covered:

1) Resampling Based Multiple Testing
2) Clustering
3) Cross-validated Selection among Estimators
4) Cross-validated Selection with Censored Data
5) Algorithms for construction of Estimators

Subtopics:

Classification and Regression, Regression on multivariate outcomes, Regression on censored outcomes (Prediction of survival), conditional density and hazard estimation.

Applications in Genomics:

a) Detection of binding sites in gene expression experiments,

b) Regression of single nucleotide polymorphisms (SNP's), gene expressions, comparitive genome hybridization measurements, and epidemiologic variables, on clinical outcomes such as survival or time till recurrence,

c) Clustering protein structures, classifying or predicting protein structures, clustering genes/patients based on gene expression experiments

d) many others.

Resampling based Multiple Testing with Asymptotic Control of Type-I Error: Single Parameter Hypotheses and Single Step Procedures

Mark van der Laan

Division of Biostatistics, UC Berkeley

www.stat.berkeley.edu/~laan

www.bepress.com/ucbbiostat/

Multivariate Statistical Methods in Genomics

PH 243A, 2305 Tolman, MW 12-2

Fall 2003 ©Copyright 2003, all rights reserved

DATA AND NULL HYPOTHESES

Data: $X_1, \ldots, X_n$ i.i.d. observations of a multidimensional vector $X \sim P \in \mathcal{M}$ for a model $\mathcal{M}$.

- gene expression measurements
- gene expression, covariates, and outcomes (*e.g.*: survival)
- SNPs, covariates, and an outcome (*e.g.*: response to treatment)
- occurance of sequence motifs and gene expression.

Parameters: Real valued parameters $\mu_j(P)$, $j = 1, \ldots, p$.

Examples of Parameters:

- location parameters (means, medians, differences in means)
- regression parameters (association between gene j's expression and outcome)
- Survival probabilities.

Null Hypotheses:

$$H_{0,j} : \mu_j(P) = \mu_j^0, \; j = 1, \ldots, p,$$

where μ_j^0 are hypothesized null values.

TEST STATISTICS

Test $H_{0,j}, j = 1, \ldots, p$, with T_{jn} defined by

$$\begin{aligned} T_{jn} &\equiv \mu_{jn} - \mu_j^0 \\ \text{or } T_{jn} &\equiv \frac{\mu_{jn} - \mu_j^0}{\hat{\sigma}(\mu_{jn})}. \end{aligned}$$

We assume that μ_{jn} is an **asymptotically linear** estimator of μ_j, that is,

$$\sqrt{n}(\mu_{jn} - \mu_j) = \frac{1}{n} \sum_{i=1}^{n} IC_j(X_i|P) + o_P(1), \qquad (1)$$

for some function $X \rightarrow IC_j(X|P)$, $j = 1, \ldots, p$. Then, we know that as $n \to \infty$

$$Z_n \equiv \sqrt{n}(\mu_n - \mu(P)) \overset{D}{\Rightarrow} N(0, \Sigma(P)), \tag{2}$$

where

$$\Sigma(P) = E_P(IC(X \mid P)IC(X \mid P)^\top)$$

is the covariance of the vector **influence curve** $IC(X \mid P) = \{IC_j(X \mid P) : j = 1, \ldots, p\}$ of μ_{jn}.

Let

$$Z \sim Q_0(P) \equiv N(0, \Sigma(P)) \tag{3}$$

represent the limit (in distribution) of Z_n.

ERROR RATES

Given a vector c, consider a corresponding **multiple testing procedure** $MT(c)$ defined by:

$$\text{Reject } H_{0,j}, \text{ if } \mid T_{jn} \mid > c_j,\ j = 1, \ldots, p, \tag{4}$$

Let:

- $V_n(c) = \sum_{j=1}^{p} I(\mid T_{jn} \mid > c_j, \mu_j(P) = \mu_j^0)$ be the number of false positives of $MT(c)$,
- For a candidate cdf F of V_n, let $\theta(F) \in (0,1)$ measure a particular type-I-error rate satisfying 1) continuity in F and 2) monotonicity in F in the sense that $\theta(F) \geq \theta(G)$ if $F \leq G$.

Examples of Error rates $\theta(F_{V_n})$:

- $\int x dF_{V_n}(x)/p = E(V_n)/p$: per-comparison error rate (PCER),
- $\int x dF_{V_n}(x) = E(V_n)$: per-family error rate (PFER),
- $1 - F_{V_n}(0) = Pr(V_n \geq 1)$: family-wise error rate (FWER),
- $1 - F_{V_n}(k) = Pr(V_n \geq k)$: Generalized family wise error rate (gFWER).

SINGLE STEP CUT-OFF RULE and ERROR CONTROL

Let $c = c(Q, \alpha)$ denote a vector function cut-off rule such that if $T_n \sim Q$, then $MT(c)$ has the property that $\theta(F_{R_n(c)}) = \alpha$, where

$$R_n(c) = \sum_{j=1}^{p} I(| T_{jn} |> c_j).$$

A sensible cut-off rule: set c_j equal to the 1-δ-quantile of the j-th marginal distribution of Q, where δ is fine-tuned to yield control at level α.

So, $MT(c) = MT(c(Q, \alpha))$ depends critically on the choice of distribution Q under which the error rate is controlled.

We want to choose an estimated distribution Q_n so that $c_n = c(Q_n, \alpha)$ satisfies

$$\limsup_{n\to\infty} \theta(F_{V_n}) \leq \alpha.$$

That is, for large enough sample size, the error rate α_n for a sample of size n is bounded from above by the target error rate α.

NULL DISTRIBUTIONS

Let $Q_n(P)$ be the distribution of the test statistics under $X \sim P$. We seek to control the error rate under a **test statistic distribution** that satisfies the overal null hypotheses and is as close as possible to the true test statistic distribution $Q_n(P)$. Therefore, the correct null distribution is the **projection** of $Q_n(P)$ onto the space of mean zero distributions.

NOTE: Current approach is to choose a null data generating distribution $P_0 \in \mathcal{M}_0 = \{P : \mu(P) = \mu_0\}$, and control error rate under $Q_n(P_0)$.

Let $P_0 = P_0(P) \equiv \Pi(P \mid \mathcal{M}_0)$ be a projection (e.g. Kullback-Leibler) of the true data generating distribution onto $\mathcal{M}_0$. Let $Q_{0n} = Q_{0n}(P) = \Pi(Q_n(P) \mid \mathcal{Q}_0)$ be the projection of the test-statistic distribution $Q_n(P)$ onto the space of mean zero distributions. In general,

$$\lim_{n\to\infty} Q_{0n} = N(0, \Sigma(P)) \neq \lim_{n\to\infty} Q_n(P_0) = N(0, \Sigma(P_0)).$$

RESULTS:

1. Let $Q_0 = N(0, \Sigma(P))$. If $c_0 \equiv c(Q_0, \alpha)$ then $MT(c_0)$ asymptotically controls the error rate at level α.

2. Let Q_{0n} be an estimator of Q_0. Let $c_{0n} \equiv c(Q_{0n}, \alpha)$ and suppose $c_{0n} \overset{P}{\rightarrow} c_0 = c(Q_0, \alpha)$ for $n \rightarrow \infty$. Then

$$\limsup_{n\rightarrow\infty} \theta\left(F_{V_n(c_{0n})}\right) \leq \alpha. \qquad (5)$$

ESTIMATION: Estimate Q_0 with Q_{0n}, *e.g.*:

- $Q_{0n} = N(0, \Sigma_n)$. Provides asymptotic control.
- Bootstrap method. Provides asymptotic control.
- $Q_n(P_{0n})$, where P_{0n} is an estimated data null distribution. Does **not** provide asymptotic control *unless*

$$\Sigma(P_0) = \Sigma(P). \tag{6}$$

Condition (6) is the formal analogue of the **subset pivotality condition** (Westfall and Young, 1993, p.42-43).

BOOTSTRAP ESTIMATED NULL DISTRIBUTION

Suppose $T_n = \sqrt{n}(\mu_n - \mu_0)$. Let

- $\tilde{P}_n$ be an estimator of P according to model $\mathcal{M}$.
- $\tilde{\mu}_n = \mu(\tilde{P}_n)$ be the parameter estimate under $\tilde{P}_n$
- $\mu_n^\#$ be μ_n applied to n i.i.d. copies $X_1^\#, \ldots, X_n^\#$ of $X^\# \sim \tilde{P}_n$
- $Q_{0n}^\#$ be the distribution of $Z_n^\# = \sqrt{n}(\mu_n^\# - \tilde{\mu}_n)$

Estimate Q_0 with $Q_{0n}^\#$. Under weak regularity conditions, it is known that $Z_n^\# \overset{D}{\Rightarrow} Z \sim Q_0$ conditional on $\tilde{P}_n$, and hence Q_{0n} consistently estimates Q_0.

Define

$$R_n^{\#}(c) \quad \equiv \quad \sum_{j=1}^{p} I(\mid Z_{jn}^{\#} \mid > c_j)$$

and let $c_n = c(Q_{0n}^{\#}, \alpha)$: that is, it satisfies $\theta\left(F_{R_n^{\#}(c)}\right) = \alpha$. Then, $MT(c_n)$ is a **bootstrap based multiple testing procedure** asymptotically controlling $\theta(F_{V_n})$ at level α.

TWO SAMPLE PROBLEM

Suppose we have n_1 observations from Population 1 with mean μ_1 and n_2 observations from Population 2 with mean μ_2.

Null Hypotheses: $H_{0,j} : \mu_j \equiv \mu_{2,j} - \mu_{1,j} = 0, j = 1, \ldots, p.$

Test Statistics:

$$\begin{aligned} D_{jn} &= \bar{X}_{2,j} - \bar{X}_{1,j}, j = 1, \ldots, p \\ \text{or } T_{jn} &= \frac{\bar{X}_{2,j} - \bar{X}_{1,j}}{\sqrt{\hat{\sigma}^2_{1,j}/n_1 + \hat{\sigma}^2_{2,j}/n_2}}, j = 1, \ldots, p. \end{aligned}$$

COMPARISON OF NULL DISTRIBUTIONS

Let $COV(X_j, X_{j'})$ be ϕ_1 in population 1 and ϕ_2 in population 2.

Distribution	$Var(D_{jn})$	$Cov(D_{jn}, D_{j'n})$
Permutations	$\frac{\sigma^2_{1,j}}{n_2} + \frac{\sigma^2_{2,j}}{n_1}$	$\frac{\phi_1}{n_2} + \frac{\phi_2}{n_1}$
Bootstrap	$\frac{\sigma^2_{1,j}}{n_1} + \frac{\sigma^2_{2,j}}{n_2}$	$\frac{\phi_1}{n_1} + \frac{\phi_2}{n_2}$

Note:

- $VAR(T_{jn}) = 1$ for both distributions.
- But the two expressions for $COV(T_{jn}, T_{j'n})$ are not equivalent unless $n_1 = n_2$.

EQUIVALENCE: MULTIPLE TESTING/CONFIDENCE REGION

Let $c_n = c(Q_n^{\#}, \alpha)$. Then, the random region $\{\mu : \sqrt{n}|\mu_n - \mu| < c_n\}$ or

$$\left\{\mu : \mu_{jn} - \frac{c_{jn}}{\sqrt{n}} < \mu_j < \mu_{jn} + \frac{c_{jn}}{\sqrt{n}}, j = 1, \ldots, p\right\} \tag{7}$$

is a θ-specific $(1-\alpha)\%$ confidence region for $\mu(P)$.

The multiple testing procedure $MT(c_n)$ equals:

$$\text{Reject } H_{0j} \text{ if } \mu_j^0 \text{ is outside the interval } \left[\mu_{jn} - \frac{c_{jn}}{\sqrt{n}}, \mu_{jn} + \frac{c_{jn}}{\sqrt{n}}\right],$$

for $j = 1, \ldots, p$.

The correlation example

Suppose we observe n i.i.d. observations $X_1, \ldots, X_n$ of a vector $X = (X(1), X(2), X(3))$. For the sake of illustrations, we will assume that the variables are standardized so that $\mathrm{VAR}(X(j)) = 1$, $j = 1, 2, 3$, and suppose that we assume the parametric model $X \sim N(0, \rho)$, where ρ is the correlation matrix of X. Let ρ_j, $j \in J \equiv \{(12), (13), (23)\}$ denote the three unknown correlations. Suppose we are concerned with testing $H_{0,j} : \rho_j = 0$, $j \in J$. Let ρ_{jn}, $j \in J$, be the empirical correlations, and suppose that we use as test-statistics $T_{nj} = \sqrt{n}\rho_{jn}$. Let $S_0 = \{j \in J : \rho_j = 0\}$ be the set of true nulls.

Let $Q_{n1} = Q_n(P_0)$ be the null distribution of T_n under the data generating distribution $P_0 = N(0, I)$, where I denotes the identity matrix. Let Q_{n2} be the distribution of $\sqrt{n}(\rho_n - \rho)$. One wants to choose a null distribution which is such that the sub-distribution corresponding with the components in S_0equals or approximates well the actual distribution of $T_{nj}, j \in S_0$. It follows immediately that the sub-distribution of Q_{n2} corresponding with the components in S_0 equals (by definition) the distribution of $T_{nj}, j \in S_0$. Consequently, an estimate of the limit distribution of Q_{n1} as one obtains with either the nonparametric bootstrap, or model based bootstrap, or influence curve approach, consistently estimates the actual distribution of T_{nj}, $j \in S_0$. We will now show that the sub-distribution of Q_{n1} fails to do this.

Firstly, if the components of X are uncorrelated, then it follows immediately that the three empirical correlations are independent. Consequently, by the CLT it follows that $Q_n(P_0)$ converges to a $N(0, I)$. However, two empirical correlations corresponding with true nulls are not necessarily (asymptotically) uncorrelated. For example, suppose that $\rho_{13} = \rho_{23} = 0$, but $\rho_{23} \neq 0$. Then it follows that

$$\sqrt{n}(\rho_{n12}, \rho_{n13}) \overset{D}{\Rightarrow} N(0, \Sigma_0),$$

where the 2 by 2 matrix Σ_0 is 1 on the diagonal and ρ_{23} off-diagonal.

DATA ANALYSIS

The publicly available data set of Alizadeh *et al.* (2000):

- Blood samples from $n = 40$ **Diffuse Large B-Cell Lymphoma (DLBCL)** patients
- Expression of $p = 13,412$ clones (relative to a pooled control) measured with cDNA arrays
- Patients belong to two molecularly distinct disease groups:
  - $n_1 = 21$ **Activated** with mean μ_1
  - $n_2 = 19$ **Germinal Center (GC)** with mean μ_2
- Survival time T measured on each patient
- Pre-processing:
  - $\log_2$
  - replace missing values with the mean for that gene
  - truncate ratios exceeding 20-fold to $\pm \log_2(20)$

DIFFERENCE IN MEANS: METHOD

- Null hypotheses: for $j = 1, \ldots, p$

$$H_{0,j} : \mu_j \equiv \mu_{2,j} - \mu_{1,j} = 0.$$

- Test statistics: $T_{jn} = \frac{\mu_{jn} - 0}{sd(\mu_{jn})} = \frac{\bar{X}_{2,j} - \bar{X}_{1,j}}{\sqrt{\hat{\sigma}^2_{1,j}/n_1 + \hat{\sigma}^2_{2,j}/n_2}}$.
- Control the usual FWER: $Pr(V \geq 1) = \alpha = 0.05$
- Estimated null distributions and thresholds:
  1. Fine-tuned common quantiles with the non-parametric bootstrap distribution,
  2. Fine-tuned common quantiles with the permutation distribution,
  3. Bonferroni common *threshold* with the tabled t-distribution.

DIFFERENCE IN MEANS: RESULTS

Null Distribution	Rejections
Non-parametric bootstrap	186
Permutations	287
T-distribution	32

Number of rejected null hypotheses (out of $p = 13,412$) for three different choices of multiple testing procedure. All 32 of the genes in the t-distribution subset are in both the permutation and the bootstrap subset, and the bootstrap and permutation subsets have 156 genes in common.

LOGISTIC REGRESSION: METHOD

- Logistic Regression Model for each gene: $j = 1, \dots, p$

$$E(Group \mid X_j) = \frac{e^{\beta_{0,j} + \beta_{1,j} * X_j}}{1 + e^{\beta_{0,j} + \beta_{1,j} * X_j}}$$

- Null hypotheses: for $j = 1, \dots, p$

$$H_{0,j} : \beta_{1,j} = 0.$$

- Test statistic: $\sqrt{n} * \beta_{1n}$.
- Control the gFWER $Pr(V \geq k) = \alpha = 0.05$ for $k = 1, \dots, 100$.
- Fine-tuned common quantiles.
- Estimated null distributions: Nonparametric bootstrap.
- RESULTS

$k =$	1	10	50	100	200
Rejections	303	303	303	471	553

Table 1: Logistic Regression Parameters. Number of rejected null hypotheses (out of $p = 13,412$) using the non-parametric bootstrap estimated null distribution and controlling the gFWER $P(V_n > k)$ for different choices of k, where V_n is the number of false positives. The test statistics used are $\sqrt{n} * (\beta_n - 0)$.

LINEAR REGRESSION: METHOD

- Accelerated failure time model for each gene: $j = 1, \ldots, p$

$$E(\log(T) \mid X_j) = \gamma_{0,j} + \gamma_{1,j} * X_j$$

- Use an Inverse Probability of Censoring Weighted (IPCW) estimator for γ since survival is right-censored for some patients.
- Null hypotheses: for $j = 1, \ldots, p$

$$H_{0,j} : \gamma_{1,j} = 0.$$

- Test statistic: $\sqrt{n} * \gamma_{1n}$.
- Control the gFWER $Pr(V \geq k) = \alpha = 0.05$ for $k = 1, \ldots, 100$.
- Fine-tuned common quantiles with the non-parametric bootstrap distribution.
- Could do for each disease group (Activated and GC) separately

and compare lists of significant genes.

- RESULTS soon...

SIMULTATIONS USING REAL DATA

- Sample from the data set derived from Alizadeh *et al.* (2000)
- $p = 100$ random genes, centered to have mean zero in both groups.

	Permutation	Non-parametric	Parametric
		Bootstrap	Bootstrap
$n_1 = 5, n_2 = 15$			
D_j	0.21	0.025	0.085
T_j	0.020	0.025	0.020
$n_1 = 9, n_2 = 11$			
D_j	0.13	0.050	0.065
T_j	0.015	0.065	0.015
$n_1 = 10, n_2 = 10$			
D_j	0.17	0.060	0.070
T_j	0.020	0.055	0.035

Estimates $\hat{\alpha}$ of the error rate $Pr(V > 10)$ over $I = 200$ independent simulated data sets for null distributions of D_j and T_j. The target error rate is $\alpha = 0.05$.

CONCLUSIONS

1. $Q_0 = N(0, \Sigma(P))$ is the asymptotically correct null distribution for the test statistics $\sqrt{n}(\mu_n - \mu^0)$ and it provides asymptotic control of type I error rates that are a function of the distribution of the number of false positives.
2. For a finite sample, Q_0 can be consistently estimated with the standard bootstrap.
3. Common practice of estimating Q_0 via a data null distribution P_0 only provides asymptotic control when $\Sigma(P_0) = \Sigma(P)$.
4. Multiple testing is equivalent with constructing an 0.95-error specific confidence region (e.g. using the bootstrap).
5. Two Sample Problem: Permutation data null distribution P_{0n} has the wrong asymptotic covariance unless $n_1 = n_2$ or $\Sigma_1 = \Sigma_2$

Resampling based Multiple Testing with Asymptotic Control of Type-I Error: General Hypotheses, Single Step and Step-Down Procedures

Mark van der Laan

Division of Biostatistics, UC Berkeley

www.stat.berkeley.edu/~laan

www.bepress.com/ucbbiostat/ 0.75cm

Multivariate Statistical Methods in Genomics

PH 243A, 2305 Tolman, MW 12-2

Fall 2003

©Copyright 2003, all rights reserved

Multiple hypothesis testing framework

Model. Let $X_1, \ldots, X_n$ be n i.i.d. copies of a random variable $X \sim P \in \mathcal{M}$, where P is known to be an element of a particular statistical model $\mathcal{M}$ (possibly nonparametric). Let $\mathcal{M}_j \subset \mathcal{M}$ be a collection of m submodels and let $H_{0j} = I(P \in \mathcal{M}_j)$ be the corresponding set of null hypotheses, $j = 1, \ldots, m$. Thus, H_{0j} is true if $P \in \mathcal{M}_j$ and false otherwise.

Let $S_0 = S_0(P) \equiv \{j : H_{0j} \text{ is true}\} = \{j : P \in \mathcal{M}_j\}$ be the set of $m_0 = |S_0|$ true null hypotheses, where we note that S_0 depends on the true data generating distribution P. Let $S_0^c = S_0^c(P) \equiv \{j : j \notin S_0\}$ be the set of $m_1 = m - m_0$ false null hypotheses, i.e., true positives.

Example: $H_{0j} : \mu(j) \leq \mu_0(j)$, where the $\mu(j) = \mu(j \mid P)$ are real-valued parameters

Type I error rates. Decision to reject or not the null hypotheses are based on test statistics $T_n(j)$, $j = 1, \ldots, m$, where we assume that large values of $T_n(j)$ provide evidence against the null hypothesis H_{0j}. Let $T_n = (T_n(j) : j = 1, \ldots, m)$ be the corresponding m-vector of test statistics, with joint distribution $Q_n = Q_n(P)$. The end-product of single-step or stepwise multiple hypothesis testing procedures is a set, $S_n = S(X_1, \ldots, X_n; Q_0, \alpha) \subseteq \{1, \ldots, m\}$, of rejected hypotheses, i.e., of null hypotheses believed to be false.

$S(X_1, \ldots, X_n; Q_0, \alpha)$, the set S_n depends on the data, $X_1, \ldots, X_n$, the choice of null distribution Q_0 for computing cut-offs for the test statistics or p-values, and on α, the nominal level of the test).

Two types of errors can be committed: a false positive, or *Type I error*, is committed by rejecting a true null hypothesis, and a false negative, or *Type II error*, is committed when the test fails to reject a false null hypothesis.

The situation can be summarized by the table below, where the number of Type I errors is $V_n = |S_n \cap S_0|$ and the number of Type II errors is $U_n = |S_n^C \cap S_0^C|$. Note that both U_n and V_n depend on the unknown data generating distribution P through $S_0 = S_0(P)$. The numbers $m_0 = |S_0|$ and $m_1 = m - m_0$ of true and false null hypotheses are unknown parameters, the number of rejected hypotheses $R_n = |S_n|$ is an observable random variable, and $m_1 - U_n$, U_n, $m_0 - V_n$, and V_n are unobservable random variables (depending on P, through $S_0(P)$).

Table 2: Type I and Type II errors in multiple hypothesis testing.

		Null hypotheses		
		not rejected	rejected	
Null hypotheses	true	$m_0 - V_n$	V_n (Type I)	m_0
	false	U_n (Type II)	$m_1 - U_n$	m_1
		$m - R_n$	R_n	m

Type-I Error Rate

In general, we make the following assumptions for the parameter $\theta(F_{V_n})$ defining a particular Type I error rate. Given the distance measure $d(F_1, F_2) = \max_{x\in\{0,\ldots,m\}} \mid F_1(x) - F_2(x) \mid$ for two cumulative distribution functions F_1 and F_2 on $\{0, \ldots, m\}$, we assume that the parameter $\theta(F)$ satisfies the following properties, where F represents a c.d.f. on $\{0, \ldots, m\}$ for V_n.

Monotonicity.

$$\text{If } F_1 \geq F_2, \text{ then } \theta(F_1) \leq \theta(F_2). \tag{8}$$

Uniform Continuity.

$$\text{If } d(F_n, G_n) \to 0, \text{ then } \theta(F_n) - \theta(G_n) \to 0, \tag{9}$$

or equivalently,

$$\sup_{\{(F,G):d(F,G)\leq\delta_n\}} \mid \theta(F) - \theta(G) \mid \to 0$$

if $\delta_n \to 0$.

Adjusted p-values

As in the single hypothesis setting, multiple testing procedures may be described in terms of p-values. Given any multiple testing procedure, the *adjusted p-value* corresponding to the test of a single hypothesis H_{0j} can be defined as the nominal level of the entire procedure at which H_{0j} would just be rejected, given the values of all test statistics involved. In terms of our previous notation, the adjusted p-value for hypothesis H_{0j}, given a multiple testing procedure $S_n = S(X_1, \ldots, X_n; Q_0, \alpha)$, is

$$\tilde{p}_n(j) = \inf \left\{ \alpha \in [0,1] : j \in S(X_1, \ldots, X_n; Q_0, \alpha) \right\}, \qquad (10)$$

where the *nominal* Type I error rate is the α-level at which the specified procedure is performed. Hypothesis H_{0j} is then rejected at nominal Type I error rate α if $\tilde{p}_n(j) \leq \alpha$. This definition of adjusted p-values applies to procedures controlling any type of error rate, e.g., gFWER, PCER, FDR.

As in the single hypothesis case, an advantage of reporting adjusted p-values, as opposed to only rejection or not of the hypotheses, is that the level of the test does not need to be determined in advance, that is, results of the multiple testing procedure are provided for all α.

Single Step Multiple Testing Procedure

A hypothesis H_{0j} is rejected if $T_n(j) > c_j$, for an m-vector of cut-offs $c = (c_j : j = 1, \ldots, m)$. Denote the number of rejected hypotheses and Type I errors by

$$\begin{aligned} R(c \mid Q) &= \sum_{j=1}^{m} I(T_n(j) > c_j) \qquad \text{and} \\ V(c \mid Q) &= \sum_{j \in S_0} I(T_n(j) > c_j), \end{aligned}$$

respectively, where the notation $R(c \mid Q)$ and $V(c \mid Q)$ acknowledges that the distribution of the above sums is defined by a distribution Q for the test statistics T_n.

Procedure 1. Single-step procedure — control of general Type I error rates $\theta(F_{V_n})$.
Given a null distribution Q_0, define a vector of cut-offs $c(Q_0, \delta) = (c_j(Q_0, \delta) : j = 1, \ldots, m)$ for the test statistics T_n, such that $c_j(Q_0, \delta)$ is the $(1-\delta)$–quantile of the marginal distribution Q_{0j}, $j = 1, \ldots, m$. Let δ be chosen as

$$\delta_0 = \delta_0(\alpha) = \max\{\delta : \theta(F_{R(c(Q_0,\delta)|Q_0)}) \leq \alpha\}. \tag{11}$$

Here $c(Q_0, \delta_0(\alpha))$ is referred to as the common-quantile cut-off rule for type-I-error θ based on the null distribution Q_0.
The single-step multiple testing procedure for controlling the Type I error rate $\theta(F_{V_n})$ at level α is defined by

$$\text{Reject } H_{0j} \text{ if } T_n(j) > c_j(Q_0, \delta_0(\alpha)), \qquad j = 1, \ldots, m.$$

Rather than simply reporting rejection or not of the hypotheses at a prespecified level α, one can report adjusted p-values for Procedure 1, computed under the null distribution Q_0. The adjusted p-value for hypothesis H_{0j} is given by

$$\tilde{P}_n(j) = \theta(F_{R(c(Q_0,\delta_{0j})|Q_0)}), \qquad \text{where} \qquad \delta_{0j} = \bar{Q}_{0j}(T_n(j)) \quad (12)$$

and $\bar{Q}_{0j}$, $j = 1, \ldots, m$, denote the marginal survival functions corresponding to the null distribution Q_0. The procedure for controlling the Type I error rate at level α can then be stated equivalently as

$$\text{Reject } H_{0j} \text{ if } \tilde{P}_n(j) \leq \alpha, \qquad j = 1, \ldots, m.$$

Theorem: Given a null distribution Q_0 and $\alpha \in (0,1)$, denote the number of Type I errors for Procedure 1 by

$$V_n = V(c(Q_0, \delta_0(\alpha)) \mid Q_n) = \sum_{j \in S_0} I(T_n(j) > c_j(Q_0, \delta_0(\alpha))),$$

where $Q_n = Q_n(P)$ is the (finite sample) joint distribution of the test statistics T_n. Assume that there exists a random m-vector $Z \sim Q_0 = Q_0(P)$ so that

$$\liminf_{n \to \infty} Pr\left(\sum_{j \in S_0} I(T_n(j) > c_j) \leq x\right) \geq Pr_{Q_0}\left(\sum_{j \in S_0} I(Z(j) > c_j) \leq x\right). \tag{13}$$

Then Procedure 1 provides asymptotic control of the Type I error rate $\theta(F_{V_n})$, that is,

$$\limsup_{n \to \infty} \theta(F_{V_n}) \leq \alpha.$$

Explicit Proposal for Null Distribution

Suppose that there exists known m-vectors $\theta_0 \in \mathbb{R}^m$ and $\tau_0 \in \mathbb{R}^m$ (null values), so that

$$\begin{aligned} \limsup_{n\to\infty} ET_n(j) &\leq \theta_0(j) \\ \limsup_{n\to\infty} Var[T_n(j)] &\leq \tau_0(j), \qquad \text{for } j \in S_0. \end{aligned} \tag{14}$$

Let

$$\nu_{0n}(j) = \min\left(1, \frac{\tau_0(j)}{Var[T_n(j)]}\right)$$

and define the m-vector $Z_n \sim Q_{0n} = Q_{0n}(P)$ by

$$Z_n(j) \equiv \sqrt{\nu_{0n}(j)}\Big(T_n(j) + \theta_0(j) - ET_n(j)\Big), \qquad j = 1, \ldots, m. \tag{15}$$

Suppose that

$$(Z_n(j) : j \in S_0) \Rightarrow_D (Z(j) : j \in S_0), \tag{16}$$

where we allow various components of $Z \sim Q_0 = Q_0(P)$ to be degenerate (e.g., at $-\infty$). Then, for this choice of null distribution Q_0 and for all $c = (c_j : j = 1, \ldots, m)$ and x

$$\liminf_{n \to \infty} Pr \left(\sum_{j \in S_0} I(T_n(j) > c_j) \leq x \right) \geq Pr_{Q_0} \left(\sum_{j \in S_0} I(Z(j) > c_j) \leq x \right),$$

so that the previous Theorem applies.

Practical remark: Given the null values θ_0, τ_0 for the mean and variance of the test-statistic distribution (when the null would be true), respectively, this explicit proposal for the null distribution corresponds with 1) simulate a large number B of vectors T_n from the actual true distribution $Q_n(P)$, 2) compute the marginal expectation ET_n and variance $\mathrm{VAR}(T_n)$, and 3) make the $m \times B$-matrix $\sqrt{\nu_{0n}}(T_n - ET_n + \theta_0$.

Step-down Procedures for FWE

We propose two step-down multiple testing procedures, based on a null distribution $Q_0 = Q_0(P)$ that provides asymptotic control of the family-wise error rate, without the requirement of subset pivotality. The first procedure involves maxima of the test statistics $T_n(j)$ (maxT, Procedure 2) and the second is based on minima of p-values $P_n(j)$, also computed under the null Q_0 (minP, Procedure 3).

Procedure 2. Step-down procedure based on maxima of test statistics (maxT) — control of FWER

Let $T_{n,(j)}$ be the ordered test statistics, $T_{n,(1)} \geq \ldots \geq T_{n,(m)}$, and $R_n(j)$ the indices for these order statistics, so that $T_{n,(j)} = T_n(R_n(j))$, $j = 1, \ldots, m$. Given a null distribution Q_0 and $\alpha \in (0,1)$, define $(1-\alpha)$–quantiles, $c(A) = c(A, Q_0, \alpha) \in \mathbb{R}$, for maxima of random variables $Z = (Z(j) : j = 1, \ldots, m) \sim Q_0$ over the complements of subsets $A \subseteq \{1, \ldots, m\}$

$$c(A) = \inf \left\{ c : Pr_{Q_0} \left(\max_{j \notin A} Z(j) \leq c \right) \geq 1 - \alpha \right\}.$$

Given the indices $R_n(j)$ for the order statistics $T_{n,(j)}$, define $(1-\alpha)$–quantiles

$$C_n(j) = c(\{R_n(1), \ldots, R_n(j-1)\}, Q_0, \alpha)$$

and test statistics

$$T^*_{n,(j)} \equiv \begin{cases} T_{n,(j)}, & \text{if } T_{n,(j-1)} > C_n(j-1) \\ -\infty, & \text{otherwise} \end{cases}, \qquad j = 1, \ldots, m.$$

The step-down maxT multiple testing procedure for controlling the FWER at level α is defined by

$$\text{Reject } H_{0,R_n(j)} \text{ if } T^*_{n,(j)} > C_n(j), \qquad j = 1, \ldots, m.$$

Adjusted P-values

Note that the definition $T^*_{n,(j)} = -\infty$, if $T_{n,(j-1)} \leq C_n(j-1)$, ensures that the procedure is indeed step-down, that is, one can only reject a particular hypothesis provided all hypotheses with larger test statistics were rejected beforehand. Rather than simply reporting rejection or not of the hypotheses at a prespecified level α, one can report adjusted p-values for Procedure 2, computed under the null distribution Q_0. The adjusted p-value for hypothesis $H_{0,R_n(j)}$ is given by

$$\tilde{P}_n(R_n(j)) = \max_{k=1,\ldots,j} \left\{ Pr_{Q_0}\left(\max_{l \in \{R_n(k),\ldots,R_n(m)\}} Z(l) > T_n(R_n(k)) \right) \right\}. \tag{17}$$

Here the adjusted p-values are conditional on the observed test statistics and their ranks. The procedure for controlling the FWER at level α can then be stated equivalently as

$$\text{Reject } H_{0,R_{n(j)}} \text{ if } \tilde{P}_n(R_n(j)) \leq \alpha, \qquad j = 1, \ldots, m.$$

Assumptions

In order to prove asymptotic control of the FWER by Procedure 2, we make the following two assumptions.

Assumption A1T. There exists an m-dimensional random vector $Z \sim Q_0(P)$ so that

$$\limsup_{n\to\infty} Pr\left(\max_{j\in S_0} T_n(j) > x\right) \leq Pr_{Q_0}\left(\max_{j\in S_0} Z(j) > x\right) \text{ for all } x. \tag{18}$$

We also assume that for $\alpha \in (0,1)$

$$\max_{A\subseteq\{1,\ldots,m\}} c(A, Q_0, \alpha) < \infty. \tag{19}$$

Assumption A2T. There exists a degenerate maximal value M_1

(e.g., $+\infty$) so that for all $M < M_1$

$$Pr\left(\min_{j\in S_0^c} T_n(j) \geq M\right) \rightarrow 1 \qquad \text{as } n \rightarrow \infty \tag{20}$$

and

$$\lim_{M\uparrow M_1} \lim_{n\rightarrow\infty} Pr\left(\max_{j\in S_0} T_n(j) \geq M\right) = 0. \tag{21}$$

Note that these assumptions only require that T_n represents a sensible set of test statistics.

Theorem: Given a null distribution Q_0 and $\alpha \in (0, 1)$, denote the number of Type I errors for Procedure 2 by

$$V_n \equiv \sum_{j=1}^{m} I(T^*_{n,(j)} > C_n(j), R_n(j) \in S_0).$$

Suppose Assumptions A1T and A2T on the test statistics $T_n(j)$ and null distribution Q_0 hold. Then, Procedure 2 provides asymptotic control of the family-wise error rate at level α, that is,

$$\limsup_{n \to \infty} Pr(V_n \geq 1) \leq \alpha.$$

If (18) in Assumption A1T holds with equality, then asymptotic control is exact:

$$\lim_{n \to \infty} Pr(V_n \geq 1) = \alpha.$$

Outline Proof.

Note that, with probability one in the limit, the first $m_1 = |S_0^c|$ rejected hypotheses correspond to the m_1 false null hypotheses. Thus, no Type I errors are committed for these first m_1 rejections and one can focus on the m_0 least significant statistics, $T_{n,(j)}$, $j = m_1 + 1, \ldots, m$, which now correspond to the test statistics for the true nulls, $T_n(j)$, $j \in S_0$. By definition of the step-down procedure, a Type I error is committed iff $\max_{j \in S_0} T_n(j) > C_n(S_0^c)$, which is controlled at level α. Thus, conditional on having rejected the first m_1 correct rejections, with probability $1 - \alpha$ the procedure will not reject at step $m_1 + 1$ and thus result in zero false rejections.

Remark: Local alternatives cause non-control of FWE for step-down procedures.

Step-down procedure (FWE) based on minima of p-values

Procedure 2 above is a step-down analogue of the single-step *common-cut-off* procedure. One can also prove asymptotic control of the FWER for an analogue of Procedure 2, where maxima of test statistics $T_n(j)$ are replaced by minima of unadjusted p-values $P_n(j)$, also computed under the proposed null distribution Q_0: $P_n(j) = \bar{Q}_{0j}(T_n(j))$, where $\bar{Q}_{0j}$, $j = 1, \ldots, m$, denote the marginal survival functions corresponding to the null distribution Q_0. Such a procedure corresponds to (2.10) in Section 2.6 of Westfall and Young (1993), (with the important distinction in the choice of null distribution Q_0) and is a step-down version of the *common-quantile* procedure in Pollard, van der Laan (2003).

Note that procedures based on maxima of test statistics (maxT) and minima of p-values (minP) are equivalent, when the test statistics $T_n(j)$ are identically distributed, $j = 1, \ldots, m$. In this case, the marginal survival functions $\bar{Q}_{0j}$ are the same for each j, and thus the significance rankings based on $T_n(j)$ and $P_n(j)$ coincide. In general, however, the two procedures produce different results, and considerations of balance, power, and computational feasibility should dictate the choice between the two approaches. In the case of non-identically distributed test statistics $T_n(j)$, not all tests are weighted equally in the maxT procedure and this can lead to unbalanced adjustments. When the null distribution Q_0 is replaced by a resampling-based estimator $\hat{Q}_{0n}$ (Section **??**), procedures based on minima of p-values tend to be more sensitive to the number of resampling steps and more conservative than those based on maxima of test statistics, due to discreteness when estimating quantiles. Also, minP procedures require more computations than maxT procedures, because the unadjusted p-values $P_n(j)$ must be estimated before considering the distribution of their successive minima.

Finally, note that while nominal p-values computed from a standard normal or other type of distribution may not be correct, a step-down procedure based on minima of such transformed test statistics nonetheless provides asymptotic control of the FWER (e.g., $P_n(j) = \bar{\Phi}(T_n(j))$, where $\bar{\Phi}$ is the standard normal survival function). That is, these p-values can be viewed as just another type of test statistic $T_n(j)$ and one can appeal to previous theorems.

Here, however, we propose a step-down multiple testing procedure where p-values are also defined in terms of the null distribution Q_0, that is, $P_n(j) = \bar{Q}_{0j}(T_n(j))$. We therefore have a more specific procedure and assumptions for proving asymptotic control of the family-wise error rate. Type I error control by the minP procedure relies on Assumptions A1P and A2P, below.

Procedure 3. Step-down procedure based on minima of p-values (minP) — control of the FWER.

Given a null distribution Q_0, define marginal or unadjusted p-values as

$$P_n(j) = Pr_{Q_0}(Z(j) \geq T_n(j)) = \bar{Q}_{0j}(T_n(j)), \tag{22}$$

where Z is an m-dimensional random vector $Z \sim Q_0$ and $\bar{Q}_{0j}$, $j = 1, \ldots, m$, denote the marginal survival functions corresponding to the null distribution Q_0. Let $P_{n,(j)}$ be the ordered p-values, $P_{n,(1)} \leq \ldots \leq P_{n,(m)}$, and $R_n(j)$ the indices for these order statistics, so that $P_{n,(j)} = P_n(R_n(j))$, $j = 1, \ldots, m$. Define α–quantiles, $c(A) = c(A, Q_0, \alpha) \in \mathbb{R}$, $\alpha \in (0,1)$, for minima of p-values $(\bar{Q}_{0j}(Z(j)) : j = 1, \ldots m)$ over the complements of subsets $A \subseteq \{1, \ldots, m\}$

$$c(A) = \inf\left\{c : Pr_{Q_0}\left(\min_{j \notin A} \bar{Q}_{0j}(Z(j)) \leq c\right) \geq \alpha\right\}.$$

Given the indices $R_n(j)$ for the ordered p-values $P_{n,(j)}$, define α–quantiles

$$C_n(j) = c(\{R_n(1), \ldots, R_n(j-1)\}, Q_0, \alpha)$$

and statistics

$$P^*_{n,(j)} \equiv \begin{cases} P_{n,(j)}, & \text{if } P_{n,(j-1)} < C_n(j-1) \\ 1, & \text{otherwise} \end{cases}, \qquad j = 1, \ldots, m.$$

The step-down minP multiple testing procedure for controlling the FWER at level α is defined by

$$\text{Reject } H_{0,R_n(j)} \text{ if } P^*_{n,(j)} < C_n(j),\ j = 1, \ldots, m.$$

Adjusted P-values

Note that the definition $P^*_{n,(j)} = 1$, if $P_{n,(j-1)} \geq C_n(j-1)$, ensures that the procedure is indeed step-down, that is, one can only reject a particular hypothesis provided all hypotheses with smaller unadjusted p-values were rejected beforehand. Adjusted p-values are defined similarly as for Procedure 2. The adjusted p-value for hypothesis $H_{0,R_n(j)}$ is given by

$$\tilde{P}_n(R_n(j)) = \max_{k=1,\ldots,j} \left\{ Pr_{Q_0} \left(\min_{l \in \{R_n(k),\ldots,R_n(m)\}} \bar{Q}_{0j}(Z(j)) < P_n(R_n(k)) \right) \right\}. \tag{23}$$

Assumptions

Theorem below, proves asymptotic control of the FWER by Procedure 3 under the following two assumptions, which are the p-value analogues of Assumptions A1T and A2T, respectively.

Assumption A1P. There exists an m-dimensional random vector $Z \sim Q_0(P)$ so that

$$\limsup_{n\to\infty} Pr\left(\min_{j\in S_0} P_n(j) < x\right) \leq Pr_{Q_0}\left(\min_{j\in S_0} \bar{Q}_{0j}(Z(j)) < x\right) \text{ for all } x, \tag{24}$$

where $\bar{Q}_{0j}$, $j = 1, \ldots, m$, denote the marginal survival functions corresponding to the null distribution $Q_0 = Q_0(P)$. We also assume that for $\alpha \in (0, 1)$

$$\min_{A\subseteq\{1,\ldots,m\}} c(A, Q_0, \alpha) > 0. \tag{25}$$

Assumption A2P. For each $\epsilon > 0$,

$$Pr\left(\max_{j\in S_0^c} P_n(j) \leq \epsilon\right) \to 1 \qquad \text{as } n \to \infty \tag{26}$$

and

$$\lim_{\epsilon\downarrow 0}\lim_{n\to\infty} Pr\left(\min_{j\in S_0} P_n(j) \leq \epsilon\right) = 0. \tag{27}$$

Theorem: Given a null distribution Q_0 and $\alpha \in (0,1)$, denote the number of Type I errors for Procedure 3 by

$$V_n \equiv \sum_{j=1}^{m} I(P^*_{n,(j)} < C_n(j), R_n(j) \in S_0).$$

Suppose Assumptions A1P and A2P hold, specifically, conditions (24), (25), (26), and (27) are satisfied by the p-values $P_n(j)$, i.e., by the test statistics $T_n(j)$ and null distribution Q_0. Then, Procedure 3 provides asymptotic control of the family-wise error rate at level α, that is,

$$\limsup_{n\to\infty} Pr(V_n \geq 1) \leq \alpha.$$

If (24) in Assumption A1P holds with equality, then asymptotic control is exact

$$\lim_{n\to\infty} Pr(V_n \geq 1) = \alpha.$$

Procedure 4. Step-down procedure based on maxima of test statistics (maxT) — control of GFWER

Let $T_{n,(j)}$ be the ordered test statistics, $T_{n,(1)} \geq \ldots \geq T_{n,(m)}$, and $R_n(j)$ the indices for these order statistics, so that $T_{n,(j)} = T_n(R_n(j))$, $j = 1, \ldots, m$. Given a null distribution Q_0 and $\alpha \in (0,1)$, define $(1-\alpha)$–quantiles, $c(A) = c(A, Q_0, \alpha) \in \mathbb{R}$, for maxima of random variables $Z = (Z(j) : j = 1, \ldots, m) \sim Q_0$ over the complements of subsets $A \subseteq \{1, \ldots, m\}$

$$c(A) = \inf\left\{c : Pr_{Q_0}\left(\max_{j \notin A} Z(j) \leq c\right) \geq 1-\alpha\right\}.$$

Given the indices $R_n(j)$ for the order statistics $T_{n,(j)}$, define $(1-\alpha)$–quantiles

$$C_n(j) = c(\{R_n(1), \ldots, R_n(j-1)\}, Q_0, \alpha)$$

and test statistics

$$T^*_{n,(j)} \equiv \begin{cases} T_{n,(j)}, & \text{if } T_{n,(j-1)} > C_n(j-1) \\ -\infty, & \text{otherwise} \end{cases}, \qquad j = 1, \ldots, m.$$

Let

$$l^* \equiv \min(j : T^*_{n,(j)} = \infty)$$

be the number of test-statistics which are not set to $-\infty$. The step-down maxT multiple testing procedure for controlling the GFWER= $P(V_n > k)$ at level α is defined by

$$\text{Reject } H_{0,R_n(j)} \text{ if } T^*_{n,(j)} > C_n(j), \qquad j = 1, \ldots, m$$

and also reject $\{H_{0,R_n(l^*)}, H_{0,R_n(l^*+1)}, \ldots, H_{0,R_n(l^*+k-1)}\}$.

Remark. Note that this procedure is nothing else than first carrying out the Step-down procedure for controlling FWE and subsequently rejecting the next k in the ordered list of test-statistics. Let $V_n(1)$ be the number of False-positives for the procedure 2 controlling FWE, and $V_n(k)$ be the number of false positives for Procedure 4 controlling GFWE.. Since the set of rejections for the above procedure equals the union of the set of rejections for Procedure 2 controlling FWE and another k rejections, we have that $V_n(k) \leq V_n(1) + k$. Since $\limsup_{n\rightarrow\infty} P(V_n(1) > 0) \leq \alpha$, it follows that the above procedure satisfies

$$\limsup_{n\rightarrow\infty} P(V_n(k) > k) \leq \alpha.$$

In addition, if $\limsup_{n\rightarrow\infty} P(V_n(1) > 0) = \alpha$, then we have

$$\limsup_{n\rightarrow\infty} P(V_n(k) = k) = 1 - \alpha.$$

That is, with probabilty tending to 1-α, this procedure will select precisely k false positives. This gives us the following theorem. This procedure and theorem immedidately generalizes to a step-down procedure based on minima of p-values controlling GFWER.

Theorem for Procedure 4 controlling GFWER: Given a null distribution Q_0 and $\alpha \in (0,1)$, denote the number of Type I errors for Procedure 4 by

$$V_n(k) \equiv \sum_{j=1}^{l^*+k-1} I(R_n(j) \in S_0).$$

Suppose Assumptions A1T and A2T on the test statistics $T_n(j)$ and null distribution Q_0 hold. Then, Procedure 4 provides asymptotic control of the generalized family-wise error rate at level α, that is,

$$\limsup_{n\to\infty} Pr(V_n(k) > k) \leq \alpha.$$

If (18) in Assumption A1T holds with equality, then asymptotic control is exact:

$$\lim_{n\to\infty} Pr(V_n(k) > k) = \alpha,$$

and, in fact, in that case, we have

$$\lim_{n\to\infty} Pr(V_n(k) = k) = 1 - \alpha.$$

Asymptotic Control for Consistent Estimator Null Distribution

If $\hat{Q}_{0n}$ is a consistent estimator of Q_0, then the corollary below shows that procedures based on estimated cut-offs $\hat{C}_n(j)$ also provide asymptotic control of the Type I error rate. We state the Corollary for the step-down procedure based on maxima of test statistics (Procedure 2), but the same result applies to the general single-step procedure (Procedure 1) and the step-down procedure based on minima of p-values (Procedure 3).

Corollary: Let $\hat{Q}_{0n}$ be such that, given the empirical distribution P_n of $X_1, \ldots, X_n$, it converges pointwise (i.e., converges weakly) to a limit distribution Q_0 with continuous and strictly increasing marginal cumulative distribution functions Q_{0j}, $j = 1, \ldots, m$. This implies that, conditional on P_n,

$$\max_{A \subseteq \{1,\ldots,m\}} \mid c(A, \hat{Q}_{0n}, \alpha) - c(A, Q_0, \alpha) \mid \rightarrow 0. \tag{28}$$

Denote the number of Type I errors for Procedure 2 based on the estimator $\hat{Q}_{0n}$ by

$$\hat{V}_n \equiv \sum_{j=1}^{m} I(T^*_{n,(j)} > \hat{C}_n(j), R_n(j) \in S_0).$$

Then, the family-wise error rate is controlled asymptotically at level α

$$\limsup_{n \rightarrow \infty} Pr(\hat{V}_n \geq 1) \leq \alpha.$$

If (18) in Assumption A1 holds with equality, asymptotic control is exact

$$\lim_{n\to\infty} Pr(\hat{V}_n \geq 1) = \alpha.$$

Bootstrap Estimator of the Null Distribution

The asymptotic null distribution $Q_0 = Q_0(P)$ can be estimated with the non-parametric or model-based bootstrap. Let $\tilde{P}_n$ denote an estimator of the true data generating distribution P. For the non-parametric bootstrap, $\tilde{P}_n$ is simply the empirical distribution P_n, that is, samples of size n are drawn at random with replacement from the observed $X_1, \ldots, X_n$. For the model-based bootstrap, $\tilde{P}_n$ is based on a model $\mathcal{M}$, such as the m-variate normal distribution.

Each bootstrap sample consists of n i.i.d. realizations $X_1^{\#}, \ldots, X_n^{\#}$ of a random variable $X^{\#} \sim \tilde{P}_n$. Denote test statistics computed from bootstrap samples by $T_n^{\#}$. The proposed null distribution Q_0 from Theorems can be estimated by the distribution $\hat{Q}_{0n}$ of

$$Z_n^{\#}(j) \equiv \hat{\nu}_{0n}(j)\Big(T_n^{\#}(j) + \theta_0(j) - E_{\tilde{P}_n} {T_n}^{\#}(j)\Big), \qquad j = 1, \ldots, m, \tag{29}$$

where

$$\hat{\nu}_{0n}(j) = \min\left(1, \frac{\tau_0(j)}{Var_{\tilde{P}_n}[T_n{}^{\#}(j)]}\right).$$

Under regularity conditions, the bootstrap is known to be consistent, in the sense that $Z_n^{\#} \Rightarrow_D Z \sim Q_0$ conditional on $\tilde{P}_n$

In practice, one can only approximate the distribution of $Z_n^{\#}$ by an empirical distribution over B bootstrap samples drawn from $\tilde{P}_n$. That is, the estimator $\hat{Q}_{0n}$ is the empirical distribution of Z_n^b, where Z_n^b corresponds to the test statistics for the bth bootstrap sample, $b = 1, \ldots, B$.

For procedures based on maxima of the test statistics T_n (Procedure 2), the quantiles $c(A, \hat{Q}_{0n}, \alpha)$ are simply the quantiles of $\max_{j \notin A} Z_n^b(j)$ over the B bootstrap samples, that is,

$c(A, \hat{Q}_{0n}, \alpha)$ is such that

$$c(A, \hat{Q}_{0n}, \alpha) = \inf \left\{ c : \frac{1}{B} \sum_{b=1}^{B} I(\max_{j \notin A} Z_n^b(j) \leq c) \geq 1 - \alpha \right\}.$$

Resampling-based procedures for minima of p-values (Procedure 3) are more complex, as one must first estimate p-values $P_n(j) = Pr_{Q_0}(Z(j) \geq T_n(j))$ using $\hat{Q}_{0n}$, before considering the distribution of their successive minima. Unadjusted p-values $P_n(j)$ are estimated by

$$\hat{P}_n(j) = \frac{1}{B} \sum_{b=1}^{B} I(Z_n^b(j) \geq T_n(j)).$$

The reader is referred to ? for a fast algorithm for resampling estimation of adjusted p-values for step-down procedures based on minima of p-values.

Example: t-statistics for single parameter hypotheses

Consider testing m single parameter null hypotheses of the form $H_{0j} : \mu(j) \leq \mu_0(j)$ against alternative hypotheses $H_{1j} : \mu(j) > \mu_0(j)$, where $\mu(j) = \mu(j \mid P)$ is a real-valued parameter, $j = 1, \ldots, m$. Then, the set of true null hypotheses can be represented as $S_0 = \{j : \mu(j) \leq \mu_0(j)\}$.

Let $\mu_n(j)$ be an asymptotically linear estimator of $\mu(j)$, with influence curve $IC_j(X \mid P)$, that is,

$$\mu_n(j) - \mu(j) = \frac{1}{n} \sum_{i=1}^{n} IC_j(X_i \mid P) + o_P(1/\sqrt{n}), \tag{30}$$

where $E[IC_j(X \mid P)] = 0$ and $IC(X \mid P) = (IC_j(X \mid P) : j = 1, \ldots, m)$ denotes the m-dimensional vector influence curve. Let

$$T_n(j) \equiv \sqrt{n} \frac{\mu_n(j) - \mu_0(j)}{\sigma_n(j)} \tag{31}$$

be the standardized test statistic, or t-statistics, for the null hypothesis H_{0j}, where $\sigma_n(j)$ is a consistent estimator of $\sigma(j) \equiv E[IC_j(X \mid P)^2]$, $j = 1, \ldots, m$. Large values of $T_n(j)$ provide evidence against $H_{0j} : \mu(j) \leq \mu_0(j)$. Let $T_n = (T_n(j) : j = 1, \ldots, m)$ be the corresponding m-vector of test statistics, with joint distribution $Q_n = Q_n(P)$.

Choice of null distribution Q_0

The test statistics $T_n = (T_n(j) : j = 1, \ldots, m)$ satisfy Assumptions A1T and A2T and Assumptions A1P and A2P, where the null distribution $Q_0 = Q_0(P)$ is the m-variate normal distribution with mean zero and covariance matrix $\rho(P)$, the correlation matrix of the vector influence curve $IC(X \mid P)$. Thus, step-down Procedures 2 and 3, based on T_n and the null distribution Q_0, provide asymptotic control of the FWER for the test of single-parameter null hypotheses of the form $H_{0j} : \mu(j) \leq \mu_0(j)$ against alternative hypotheses $H_{1j} : \mu(j) > \mu_0(j)$, $j = 1, \ldots, m$.

Correspondence with explicit construction

The above theorems involve a null distribution Q_0 that was derived specifically in terms of the statistics μ_n in equation (31). It turns out that, under mild regularity conditions, this null distribution Q_0 corresponds to the general proposal Q_0^* defined as the asymptotic distribution of m-vectors Z_n^*, where

$$Z_n^*(j) \equiv \nu_{0n}(j)\Big(T_n(j) + \theta_0(j) - ET_n(j)\Big), \qquad j = 1, \ldots, m.$$

The proposal above for the null distribution is defined simply as the asymptotic distribution of m-vectors Z_n, where

$$Z_n(j) \equiv \sqrt{n}\frac{\mu_n(j) - \mu(j)}{\sigma_n(j)}, \qquad j = 1, \ldots, m.$$

With $\theta_0(j) \equiv 0$ and $\tau_0(j) \equiv 1$, one can show that Z_n^* and Z_n have the same asymptotic joint distribution, that is Q_0 and Q_0^* coincide.

Example: Tests of means

A familiar testing problem that falls within this framework is that where $X_1, \ldots, X_n$ are n i.i.d. random m-vectors, $X \sim P$, and the parameter of interest is the mean vector $\mu = \mu(P) = (\mu_j = \mu(j \mid P) : j = 1, \ldots, m) = EX$. Null hypotheses $H_{0j} : \mu(j) \leq \mu_0(j)$ then refer to individual components of the mean vector μ and the test statistics $T_n(j)$ are the usual one sample t-statistics, where $\mu_n(j) = \bar{X}_n(j) = \frac{1}{n} \sum_i X_i(j)$ and $\sigma_n^2(j) = \frac{1}{n} \sum_i (X_i(j) - \bar{X}_n(j))^2$ are empirical means and variances for the m components, respectively.

Example: Tests of correlations

Another common testing problem covered by this framework is that where the parameter of interest in the correlation matrix $\rho = \rho(P) = (\rho_{jk}(P))$ for the random vectors in the previous example, $\rho_{jk} = \rho_{jk}(P) = Cor(X_j, X_k)$, $j, k = 1, \ldots, m$. Suppose we are interested in testing the $m(m-1)/2$ null hypotheses that the m components of X are uncorrelated, $H_{jk} : \rho_{jk} = 0$, $j = 1, \ldots, m$, $k = j + 1, \ldots, m$. Common test statistics for this problem are $T_n(jk) = \sqrt{n} r_{jk}$, where r_{jk} are the sample correlations.

As discussed in Westfall and Young (1993), Example 2.2, p. 43, subset pivotality fails for this testing problem. To see this, consider the simple case $m = 3$ and assume H_{12} and H_{13} are true, so that $\rho_{12} = \rho_{13} = 0$. Then the joint distribution of (T_{12}, T_{23}) is asymptotically normal with mean vector zero, variance 1, and correlation ρ_{23}, and thus depends on the truth or falsity of the third hypothesis H_{23}. In other words, the asymptotic covariance of the vector influence curve for the sample correlations is not the same under the true P as it is under a null distribution P_0 for which $\rho_{jk} \equiv 0 \; \forall j \neq k$.

However, our proposed null distribution Q_0 (and bootstrap estimator thereof) for the test statistics T_n does control the Type I error rate when used in Procedures 1, 2, and 3. Tests of correlations thus provide an example where standard procedures based on subset pivotality fail, while procedures based on our general null distribution Q_0 achieve the desired control.

F-statistics for multiple parameter hypotheses

Consider random m-vectors $X_k \sim P_k$, $k = 1, \ldots, K$, from K different populations with data generating distributions P_k ***. Denote the mean vector and covariance matrix in population k by $\mu_k = EX_k$ and Σ_k, respectively. We are interested in testing the m null hypotheses $H_{0j} : \mu_k(j) \equiv \mu(j) \; \forall k$, that, for each population, the jth components $\mu_k(j)$ of the mean vectors are equal to a common value $\mu(j)$, $j = 1, \ldots, m$. As before, let S_0 denote the set of true null hypotheses. Suppose, we observe i.i.d. samples $X_{k,1}, \ldots, X_{k,n_k}$, of size n_k from population k, $k = 1, \ldots, K$. Let $n = \sum_k n_k$ denote the total sample size and $\delta_{k,n} = n_k/n$ the proportion of observations from population k in the sample, where it is assumed that, $\forall k$, $\delta_{k,n} \rightarrow \delta_k > 0$ as $n \rightarrow \infty$.

As test statistics we can use the well-known F-statistics

$$T_n(j) = \frac{1/(K-1)\sum_k n_k(\bar{X}_k(j) - \bar{X}(j))^2}{1/(n-K)\sum_{i,k}(X_{k,i}(j) - \bar{X}_k(j))^2}, \qquad j = 1, \ldots, m, \tag{32}$$

where $\bar{X}_k$ denotes the sample mean vector for population k and $\bar{X} = \sum_k \delta_{k,n}\bar{X}_k$ denotes the overall mean vector.

Choice of null distribution Q_0

Theorem: The F-statistics $T_n = (T_n(j) : j = 1, \ldots, m)$ satisfy Assumptions A1T and A2T of Theorem **??** and Assumptions A1P and A2P of Theorem **??**, where the null distribution $Q_0 = Q_0(P)$ is the joint distribution of the random m-vector $Z = f(Z_1, \ldots, Z_K)$, defined in terms of independent Gaussian m-vectors $Z_k \sim N(0, \Sigma_k)$ and a quadratic function f specified below. Thus, step-down Procedures 2 and 3, based on T_n and the null distribution Q_0, provide asymptotic control of the FWER for the test of multiple parameter null hypotheses $H_{0j} : \mu_k(j) \equiv \mu(j)\ \forall k,\ j = 1, \ldots, m$.

Proof of Theorem

Firstly, note that the denominators of the F-statistics can be written as

$$D_n(j) = \frac{n}{n-K} \sum_k \delta_{k,n} \hat{\sigma}^2_{k,n}(j), \qquad j = 1, \ldots, m,$$

where $\hat{\sigma}^2_{k,n}(j)$ are consistent estimators of the population variances $\sigma^2_k(j)$, i.e., of the diagonal elements of covariance matrices Σ_k, $k = 1, \ldots, K$. Thus, as $n \to \infty$,

$$D_n(j) \Rightarrow_P D(j) = \sum_k \delta_k \sigma^2_k(j), \qquad j = 1, \ldots, m.$$

The numerator of the F-statistics $T_n(j)$ can be rewritten as

$$N_n(j) = \frac{1}{K-1} \sum_k \left((1 - \delta_{k,n}) Z_{k,n}(j) - \sum_{l \neq k} \sqrt{\delta_{k,n}\delta_{l,n}} Z_{l,n}(j) \right)^2,$$

where $Z_{k,n} = \sqrt{n_k}(\bar{X}_k - \mu)$, $k = 1, \ldots, K$. Thus, the m-vector $T_n = (T_n(j) : j = 1, \ldots, m)$ of F-statistics can be approximated by a random m-vector Z_n that is a simple quadratic function $f(Z_{1,n}, \ldots, Z_{K,n}) = (f_j(Z_{1,n}, \ldots, Z_{K,n}) : j = 1, \ldots, m)$ of the K independent m-vectors $Z_{k,n}$, $k = 1, \ldots, K$,

$$T_n(j) \approx \frac{N_n(j)}{D(j)} = Z_n(j) = f_j(Z_{1,n}, \ldots, Z_{K,n}), \qquad j = 1, \ldots, m. \tag{33}$$

By the Central Limit Theorem, $(Z_{k,n}(j) : j \in S_0) \Rightarrow_D (Z_k(j) : j \in S_0)$, where $Z_k \sim N(0, \Sigma_k)$, $k = 1, \ldots, K$. For $j \notin S_0$, $Z_{k,n}(j) = \sqrt{n_k}(\bar{X}_k(j) - \mu_k(j)) + \sqrt{n_k}(\mu_k(j) - \mu(j))$ converge to either $+\infty$ or $-\infty$ for some k. Applying the Continuous Mapping Theorem to the function $(f_j(Z_{1,n}, \ldots, Z_{K,n}) : j \in S_0)$ proves that $(T_n(j) : j \in S_0)$ converges in distribution to $(Z(j) : j \in S_0)$, where $Z = f(Z_1, \ldots, Z_K)$ and the Z_k are independent m-vectors with

$Z_k \sim N(0, \Sigma_k)$, $k = 1, \ldots, K$. That is, the limit distribution of $(T_n(j) : j \in S_0)$ is directly implied by the multivariate normal distributions $N(0, \Sigma_k)$, where Σ_k denotes the $m \times m$ covariance matrix of $X_k \sim P_k$, $k = 1, \ldots, K$. For $j \notin S_0$, $T_n(j) \rightarrow \infty$.

Therefore, the F-statistics T_n satisfy Assumptions A1T and A2T, where the null distribution $Q_0 = Q_0(P)$ is the joint distribution of the random m-vector $Z = f(Z_1, \ldots, Z_K)$, for independent m-vectors $Z_k \sim N(0, \Sigma_k)$ and the quadratic function f defined in equation (33). In Assumption A2T, $M_1 = \infty$, and condition (19) in Assumption A1T follows immediately by continuity of Q_0. For this definition of Q_0, Assumptions A1P and A2P are also satisfied.

Explicit Null Distribution for F-test Statistics

Multiple Testing with Asymptotic Control of False Discovery Rate

Mark van der Laan

Division of Biostatistics, UC Berkeley

www.stat.berkeley.edu/~laan

www.bepress.com/ucbbiostat/

©Copyright 2003, all rights reserved

The nonparametric mixture model for test-statistics

Let $T_1^n, \ldots, T_m^n$ be m independent and identically distributed test statistics for null hypotheses $H_{0,j}$, $j = 1, \ldots, m$, with density being a mixture of a known null density $f_{0,n}$ and unknown density $f_{1,n}$ with unknown mixing proportion p_0:

$$T_j^n \sim f_n \equiv p_0 f_{0,n} + (1 - p_0) f_{1,n}.$$

Let $F_n, F_{0,n}, F_{1,n}$ be the corresponding cdf's.

Remark: A more common situation is that the finite sample null distribution $F_{0,n}$ can be consistently estimated with an estimator $\hat{F}_{0,n}$ in the sense that

$$\begin{aligned} F_{0,n} - F_0 &\rightarrow 0 \\ \hat{F}_{0,n} - F_0 &\rightarrow 0, \end{aligned}$$

where F_0 denotes a limit null distribution. In this case, one replaces the null distribution $F_{0,n}$ by its estimate $\hat{F}_{0,n}$.

Equivalently, let $B_j \sim \text{Bernoulli}(1 - p_0)$ be the hidden label, T_j^n, given B_j, has density $f_{B_j,n}$, the full data are m i.i.d copies (B_j, T_j^n), $j = 1, \dots, p$, but we only observe the test-statistics T_j^n. Here $B_j = 1 - I(H_{0,j} \text{ is true})$ indicates if the null-hypothesis $H_{0,j}$ is true.

Let (B, T^n) denote the random variables described by:
$B \sim \text{Bernoulli}(p_0)$, and T^n, given B, has density $f_{B,n}$.

Approximate correspondence between frequentist independence and nonparametric mixture model

Frequentist independence model for test-statistics: Suppose that it is known that the test-statistics are independent, that all the marginal distributions of test statistics corresponding with a true null hypothesis $H_{0,j}$ equal a common known distribution $F_{0,n}$, but that distributions $F_{j,n}$ of T_j^n, are unknown otherwise.

Let $S_0 = \{j : F_{j,n} = F_{0,n}\}$ be the set of true nulls. Let $p_0 \equiv \mid S_0 \mid /m$. Let $F_{1,n}$ be the distribution of the mixture of $F_{j,n}$, $j \in S_0^c$, with uniform mixing distribution. For a dominating measure μ, let $f_{0,n} \equiv dF_{0,n}/d\mu$, and $f_{1,n} \equiv dF_{1,n}/d\mu$.
Approximate correspondence: Consider a parameter (such as FDR of the set $\{j : T_j^n > t\}$) of the distribution of $\vec{T}^n$ under the independence model which only depends on the m-marginal distributions $F_{j,n}$ through $p_0, F_{0,n}, F_{1,n}$. Then this parameter has

approximately the same value under this independence distribution as under the corresponding mixture model distribution defined by $B_j \sim$ Bernoulli(p_0), $T_j^n \sim f_{B_j,n}$, $j = 1, \ldots, p$. Therefore, the mixture model provides a convenient working model to find the right cut-off $t(\alpha)$ so that the FDR of the set $\{j : T_j^n > t\}$ equals α.

For example, it can be verified that

$$\frac{E_{IND} \sum_{j=1}^{m} I(T_j^n > t, j \in S_0)}{E_{IND} \sum_{j=1}^{m} I(T_j^n > t)} = \frac{E_{MIXT} \sum_{j=1}^{m} I(T_j^n > t, B_j = 0)}{E_{MIXT} \sum_{j=1}^{m} I(T_j^n > t)},$$

where E_{IND} denotes the expectation under the distribution of $\vec{T}_n$ in the frequentist independence model identified by S_0, F_{0n}, F_{jn}, $j \notin S_0$, and E_{MIXT} denotes the expectation under the distribution of $(\vec{T}_n, \vec{B}_n)$ in the i.i.d. mixture model identified by $p_0 =\mid S_0 \mid /m$, $F_{1n} = sum_{j \notin S_0} F_{jn} / \mid S_0^c \mid$, and F_{0n}.

It remains to be seen till what degree we have:

$$E_{IND}\left(\frac{\sum_{j=1}^m I(T_j^n > t, j \in S_0)}{\sum_{j=1}^m I(T_j^n > t)}\right) \approx E_{MIXT}\left(\frac{\sum_{j=1}^m I(T_j^n > t, B_j = 0)}{\sum_{j=1}^m I(T_j^n > t)}\right).$$

One fundamental difference between the two (with each other) corresponding distributions is that under the frequentist model $\mid S_0 \mid$ is fixed, while $\sum_j I(B_j = 0)$ is random with mean $\mid S_0 \mid$. To obtain a stronger similarity one could enforce in the mixture model the constraint that $\sum_j I(B_j = 0) =\mid S_0 \mid$. The effect of this additional constraint on our calculations in the mixture model might need to be investigated.

Finite Sample Identifiability in nonparametric mixture model

Given the actual density f_n of the test-statistics and the null density $f_{0,n}$, the proportion of true nulls p_0 and the alternative density $f_{1,n}$ are identified up till:

$$\begin{aligned} 0 &\leq p_0 \leq \min\left(\min_t \frac{f_n(t)}{f_{0,n}(t)}, \min_t \frac{F_n(t)}{F_{0,n}(t)}\right) \\ f_{1,n} &= \frac{f_n - p_0 f_{0,n}}{1 - p_0}. \end{aligned}$$

Parameter of interest in nonparametric mixture model.

$$\begin{aligned} \theta_n(t) &\equiv P(B = 0 \mid T^n = t) = p_0 \frac{f_{0,n}(t)}{f_n(t)} \\ \Phi_n(t) &\equiv P(B = 0 \mid T^n > t) = p_0 \frac{\bar{F}_{0,n}(t)}{\bar{F}_n(t)} \end{aligned}$$

John Storey refers to $\Phi_n(T_j^n)$ as q-values in his work on FDR, and therefore we will refer to $\theta_n(T_j^n)$ as local q-values.

Unknown Multiple Testing Procedures Controlling FDR

In the following lemmas we adopt the common convention to define the proportion of false discoveries as zero when the set of rejections is empty.

Lemma Let $S_n^* \equiv \{j : T_j^n \leq t_n^*(\alpha)\}$, where $t_n^*(\alpha) \equiv \min\{t : \theta_n(t) \leq \alpha\}$. Let $\alpha' = \theta(t_n^*(\alpha))$. Then

$$E\frac{\mid S_n^* \cap S_0 \mid}{\mid S_n^* \mid} = \alpha' P(\mid S_n^* \mid > 0).$$

If $S_n^* \equiv \{j : \theta_n(T_j^n) \leq \alpha\}$, then

$$E\frac{\mid S_n^* \cap S_0 \mid}{\mid S_n^* \mid} \leq \alpha Pr(\mid S_n^* \mid > 0).$$

Proof: We prove the last statement. The proof of the first statement is similar. Note

$$I(\mid S_n^* \mid > 0)\frac{\mid S_n^* \cap S_0 \mid}{\mid S_n^* \mid} = \frac{\sum_j I(\theta(T_j^n) \leq \alpha, B_j = 0)}{\sum_j I(\theta(T_j^n) \leq \alpha)} I(\mid S_n^* \mid > 0).$$

The conditional expecation of this quantity, given $T_1^n, \ldots, T_p^n$, equals

$$I(\mid S_n^* \mid > 0) \frac{\sum_j I(\theta(T_j^n) \leq \alpha)\theta(T_j^n)}{\sum_j I(\theta_n(T_j^n) \leq \alpha)} \leq \alpha I(\mid S_n^* \mid > 0)$$

Lemma: Let $S_n \equiv \{j : T_j^n > t_n(\alpha)\}$, where

$$t_n(\alpha) \equiv \min\{t : \Phi_n(t) \leq \alpha\}.$$

Let $\alpha^* = \Phi_n(t_n(\alpha))$. Then

$$E\frac{\mid S_n \cap S_0 \mid}{\mid S_n \mid} = P(\mid S_n \mid > 0)\alpha^*.$$

Similarly, if $S_n \equiv \{j : \Phi_n(T_j^n) \leq \alpha)\}$, then

$$E\frac{\mid S_n \cap S_0 \mid}{\mid S_n \mid} \leq P(\mid S_n \mid > 0)\alpha.$$

Proof: We will now only proof the first statement, since the last statement is proved similarly. Note

$$\frac{\mid S_n \cap S_0 \mid}{\mid S_n \mid} = \frac{\sum_j I(T_j^n > t_n(\alpha), B_j = 0)}{\sum_j I(T_j^n > t_n(\alpha))}.$$

The conditional expectation of this quantity, given the Bernoulli indicators $I(T_1^n > t_n(\alpha)), \ldots, I(T_p^n > t_n(\alpha))$, equals

$$I(\mid S_n \mid > 0) \frac{\sum_j I(T_j^n > t_n(\alpha)) \Phi_n(t_n(\alpha))}{\sum_j I(T_j^n > t_n(\alpha))} = I(\mid S_n \mid > 0) \alpha^*.$$

Remark: Thus, if $Pr(| S_n |> 0) \approx 1$, and $\alpha^* = \alpha$ (as is the case for continuous distributions), we have that S_n and S_n^* control the FDR exactly at level α. Since Φ_n is easier to estimate from the data than θ_n, we prefer the multiple testing procedure based on estimating S_n (i.e., Φ_n) in comparison with a multiple testing procedure estimating S_n^* (i.e. θ_n).

Equivalence with Benjamini-Hochberg Method

Under the assumption that $\Phi(t)$ is monotone in t, without any loss we can use

$$S_n = \{j : \Phi_n(T_j^n) \leq \alpha\}.$$

We need an estimate of Φ_n, and thus of F_n and p_0. Let $\hat{p}_0$ be an estimate or upper bound of p_0: e.g., $\hat{p}_0 = 1$. A possible estimate of $F_n(t)$ is given by:

$$\hat{F}_n(t) \quad = \quad \frac{1}{p} \sum_{j=1}^{p} I(T_j^n \leq t).$$

Let $\hat{\Phi}_n = \hat{p}_0 \frac{\hat{\bar{F}}_{0,n}(t)}{\bar{F}_n(t)}$. Let

$$p_j \equiv \bar{F}_{0,n}(T_j^n), \; j = 1, \ldots, p$$

denote the p-values, as calculated under the null distribution $F_{0,n}$.

Then

$$\hat{S}_n = \left\{ j : p_j \leq \frac{\alpha}{\hat{p}_0} \hat{F}_n(T_j^n) \right\}.$$

This equals the Benjamini-Hochberg method for controlling the False Discovery Rate. An equivalent (common) representation of this BH-procedure is obtained by ordering the p-values as $p_{(1)} \leq \ldots \leq p_{(m)}$ and denoting the ranks by $r(1), \ldots, r(m)$ so that:

$$\hat{S}_n = \left\{ r(j) : p_{(j)} \leq \frac{\alpha}{p * \hat{p}_0} j \right\}.$$

Asymptotic control of FDR

Under the assumption that the test-statistics $T_{j,n}$, $j \in S_0^c$, converge to infinity for $n \to \infty$, we have that $\bar{F}_{1,n}(t) \to 1$ for all t. Assume that the null distribution converges as well to a limit distribution: $F_{0,n} \to F_0$ for $n \to \infty$. Then, $F_n \to p_0 F_0$ (or equivalently, $\bar{F}_n \to 1 - p_0 F_0$), and

$$\Phi_n(t) \to \Phi(t) \equiv p_0 \frac{1 - F_0(t)}{1 - p_0 F_0(t)}.$$

Let

$$\begin{aligned} S_n &= \{j : T_{j,n} > t_n(\alpha)\},\ t_n(\alpha) = \min\{t : \Phi_n(t) \leq \alpha\} \\ S'_n &= \{j : T_{j,n} > t(\alpha)\}, t(\alpha) = \min\{t : \Phi(t) \leq \alpha\}. \end{aligned}$$

Under the above assumptions and that Φ is differentiable at $t(\alpha)$ with non-zero derivative, we have

$$Pr(S_n = S'_n) \to 1.$$

We also have:

$$E\frac{| S_n \cap S_0 |}{| S_n |} - E\frac{| S_n' \cap S_0 |}{| S_n' |} \to 0.$$

Since the left-hand side equals $\alpha * Pr(| S_n |> 0)$, this shows that in order to obtain asymptotic control of the FDR, it suffices to obtain a consistent estimator of Φ and thereby of $t(\alpha)$.

Let $\hat{p_0}$, $\hat{F}_n$ and $\hat{F}_{0,n}$ be asymptotically consistent estimators of p_0, F_n (that is, $\hat{F}_n - F_n \to 0$), and F_0, and let

$$\hat{\Phi}_n(t) = \hat{p}_0 \frac{1 - \hat{F}_{0,n}(t)}{1 - \hat{F}_n(t)}$$

be the corresponding consistent estimator of $\Phi(t)$. Let

$$\hat{t}_n(\alpha) \equiv \min\{t : \hat{\Phi}_n(t) \leq \alpha\}$$

be the corresponding consistent estimator of $t(\alpha)$.

The proposed multiple testing procedure is now given by:

$$\hat{S}_n \equiv \{j : T_{j,n} > \hat{t}_n(\alpha)\},$$

and it asymptotically controls the FDR:

$$\limsup_{n\to\infty} E\frac{\mid \hat{S}_n \cap S_0 \mid}{\mid \hat{S}_n \mid} = \limsup_{n\to\infty} \alpha P(\mid \hat{S}_n \mid > 0).$$

Consistent Conservative Estimation of p_0.

Suppose we have consistent estimators $\hat{F}_n$, $\hat{F}_{0,n}$ of F, F_0 available. We note that

$$p_0 = \lim_{n\to\infty} \frac{F_n(t)}{F_{0,n}(t)}$$

for all $t < \infty$. Thus, instead of using the upper bound $p_0 = 1$, one can also consistently estimate p_0 with

$$\hat{p}_0(t) = \frac{\hat{F}_n(t)}{\hat{F}_{0,n}(t)},$$

where t is user supplied. Note, that for finite n, the estimate $\hat{p}_0(t)$ of p_0 will be (typically) conservative:

$$\hat{p}_0(t) \approx p_0 + (1 - p_0)\frac{\hat{F}_{1,n}(t)}{\hat{F}_{0,n}(t)},$$

with the bias being $(1 - p_0)\frac{\hat{F}_{1,n}(t)}{\hat{F}_{0,n}(t)}$.

One might improve on this estimate by averaging a range of these estimates over an interval $[a, b]$:

$$\hat{p}_0 \equiv \frac{1}{b-a} \int_a^b \hat{p}_0(t)dt.$$

Estimation of Null and True Distribution of Test-Statistics

Bootstrap: Suppose $T_{j,n} = T_j(X_1, \ldots, X_n)$, $j = 1, \ldots, p$, where X_i, $i = 1, \ldots, n$, are i.i.d. observations from a data generating distribution P. Let P_n be the empirical distribution. We can estimate the distribution $F_{j,n}$ of $T_{j,n}$ with the distribution of $T_j(X_1^{\#}, \ldots, X_n^{\#})$, where $X_i^{\#}$ are i.i.d. observations from the empirical distribution P_n. The uniform mixture of these p distributions $\hat{F}_{j,n}$ represents now an estimate of F_n.

Similarly, we can use as an estimate of the asymptotic null distribution F_0 the bootstrap distribution of null-value centered (and scaled) test-statistics: e.g., if $T_{j,n} = \sqrt{n}(\mu_{j,n} - \mu_{j0})$, then we estimate F_0 with the distribution of $\sqrt{n}(\mu_{j,n}^{\#} - \mu_{j,n})$ (Pollard, van der Laan, 2003).

Nonparametric control of rate of expected false discoveries.

Dependent mixture model

Let $\vec{B} = (B_1, \ldots, B_m)$ be a joint vector of bernoulli indicators $B_j = 1 - I(H_{0j}$ True), and assume that the marginal distributions of B_j is Bernoulli$(1 - p_0)$. Let $Q_{n|\vec{B}}$ denote the joint conditional distribution of the vector of test-statistics, given $\vec{B}$. Assume that the marginal distributions of $Q_{n|\vec{B}}$ corresponding with the true null hypotheses H_{0j} (i.e. $B_j = 0$) equal a common known distribution $F_{0,n}$, and that the other marginal distributions equal a common unknown distribution $F_{1,n}$. For a dominating measure μ, let $f_0^n \equiv dF_{0,n}/d\mu$, and $f_{1,n} \equiv dF_{1,n}/d\mu$. Let $F_n = p_0 F_{0,n} + 1 - p_0 F_{1,n}$, and $\Phi_n = p_0 \frac{1-F_{0,n}}{1-F_n}$. Under the above dependent mixture model, we have the following result.

Lemma: Let $S_n \equiv \{j : T_j^n > t_n(\alpha)\}$, where

$$t_n(\alpha) \equiv \min\{t : \Phi_n(t) \leq \alpha\}.$$

Let $\alpha^* = \Phi_n(t_n(\alpha))$. Then (ratio of expected number of false discoveries, REFD)

$$REFD = \frac{E_{Q_n} \mid S_n \cap S_0 \mid}{E_{Q_n} \mid S_n \mid} = \alpha^*.$$

Proof: Since $REFD$ only depends on the joint distribution of $(\vec{B}_n, \vec{T}_n)$ through its marginal distributions we have that we can replace its distribution Q_n by a distribution Q_n^* with the same but independent marginals. Under this independence distribution Q_n^* we have

$$\mid S_n \cap S_0 \mid = \sum_j I(T_j^n > t_n(\alpha), V_j = 0).$$

The conditional expectation of this quantaty, given the Bernoulli indicators $I(T_1^n > t_n(\alpha)), \ldots, I(T_p^n > t_n(\alpha))$, equals

$$\sum_j I(T_j^n > t_n(\alpha))\Phi_n(t_n(\alpha)) = \alpha^* \sum_j I(T_j^n > t_n(\alpha)).$$

This proves the statement.

Large p case

Suppose that the number of tests $p = p(n)$ converges to infinity if n converges to infinity. In this case, it is reasonable to assume that

$$\frac{E_{Q_n} \mid S_n \cap S_0 \mid}{E_{Q_n} \mid S_n \mid} - E_{Q_n} \frac{\mid S_n \cap S_0 \mid}{\mid S_n \mid} \rightarrow 0.$$

Consequently, under this assumption the multiple testing procedure $\hat{S}_n$ which controls asymptotically the REFD at level α will also control asymptotically the FDR at level α. This teaches us that the multiple testing procedure $\hat{S}_n$ can be applied in many genomics applications in which the independence assumption is invalid, and still control the FDR well.

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: Multivariate Statistical Methods in Genomics

PH 243A, 2305 Tolman, MW 12-2

Fall 2003

©Copyright 2003, all rights reserved

Acknowledgments

Joint work with

Sündüz Keleş, Division of Biostatistics, UC Berkeley.

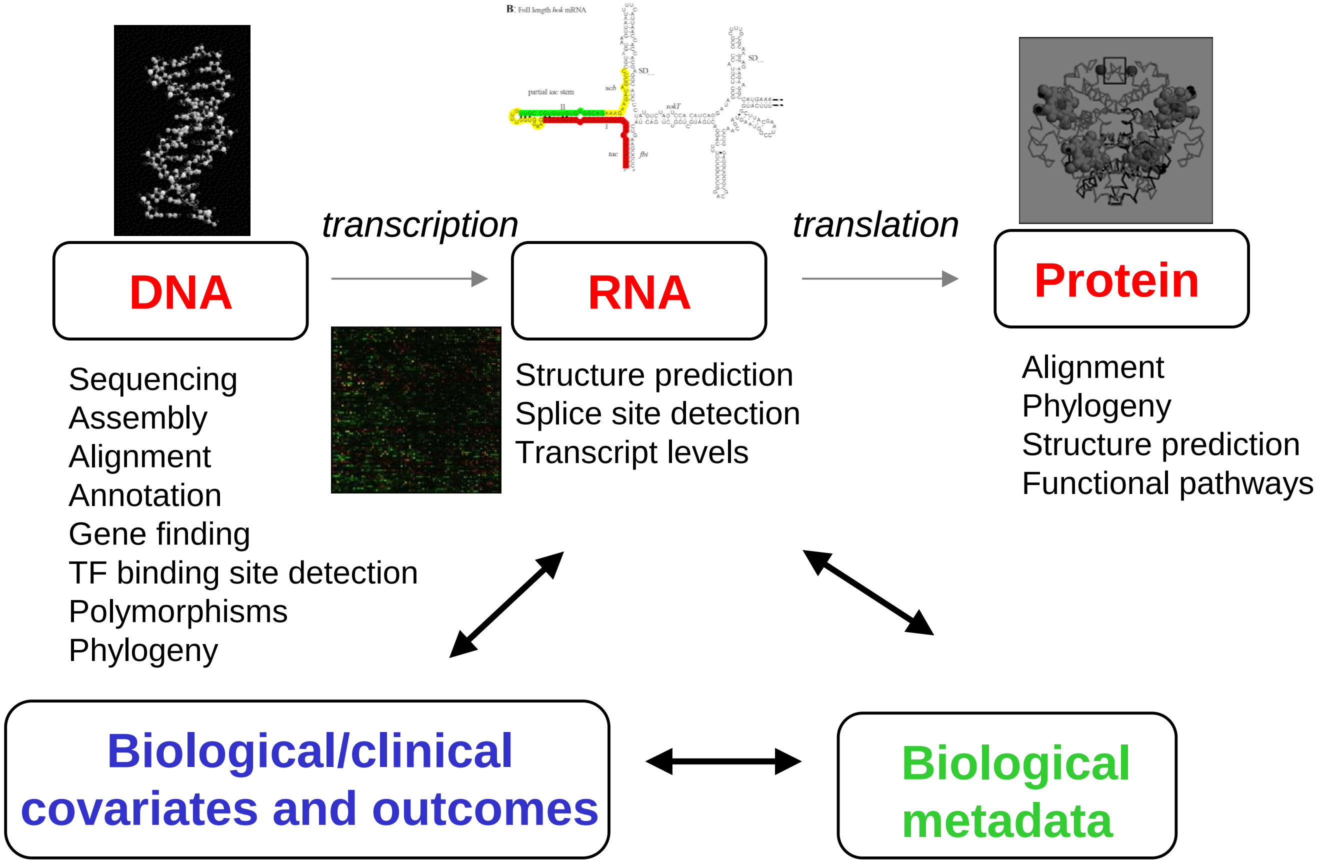
Annette Molinaro, Division of Biostatistics, UC Berkeley.

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

- Motivation: estimator construction, selection, and performance assessment in genomics.
- 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.



DNA —*transcription*→ **RNA** —*translation*→ **Protein**

DNA:
Sequencing
Assembly
Alignment
Annotation
Gene finding
TF binding site detection
Polymorphisms
Phylogeny

RNA:
Structure prediction
Splice site detection
Transcript levels

Protein:
Alignment
Phylogeny
Structure prediction
Functional pathways

Biological/clinical covariates and outcomes ↔ **Biological metadata**

Motivation: Microarray experiments

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/unaffectedness
  — 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).

Motivation: Sequence analysis

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

```
>YBR019C TCGGCGATACCTTCACCG
>YBR020W CGGGCGACGATTACCCG
>YLR081W TATCGGAGCGTAGGCGGCCGAAC
>YML051W CGGCATCCTACATGCCG
>YOR120W TCGGTTCAGACAGGTCCGG
```

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.
  E.g. Bailey & Elkan (1994), Kechris et al. (2002), Keleş et al. (2003b), Lawrence & Reilly (1990).

- Prediction of gene expression levels based on sequence features.
  E.g. Keleş 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.

General framework: van der Laan & Dudoit (2003).
Special cases and applications: Dudoit & van der Laan (2003), Keleş 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 3: Examples of loss functions for different estimation problems.

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

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

$$X \equiv \bar{X}(T) = \{X(t) : 0 \leq t \leq T\},$$

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.

Observed data structure

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

$$Pr_0(C = t \mid C \geq t, \bar{X}(T)) = Pr_0(C = t \mid C \geq t, \bar{X}(t)), \qquad \text{for } t < T.$$

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, ψ_0, is a mapping, $\psi : \mathcal{S} \rightarrow \Re$, from a covariate space $\mathcal{S}$ into the real line $\Re$. Denote the parameter space by Ψ.

The parameter ψ_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 ψ_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 $L(O, \psi \mid \eta_0)$ with the same risk

$$\int \underbrace{L(o, \psi \mid \eta_0)}_{\text{Observed data}} \, dP_0(o) = \int \underbrace{L(x, \psi)}_{\text{Full data}} \, dF_{X,0}(x).$$

Here, η_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

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

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:

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

DR-IPCW loss function

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

$$
\begin{aligned}
& L(O, \psi \mid Q, G) \\
\equiv \; & \frac{L(X,\psi)\Delta}{\bar{G}(T|X)} + \int E_{G,Q}\left(\frac{L(X,\psi)\Delta}{\bar{G}(T|X)} \mid \bar{X}(u), \tilde{T} \geq u \right) dM_G(u),
\end{aligned}
$$

where

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

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,

$$\begin{aligned} d_n(\hat{\psi}_{\hat{k}}, \psi_0) &\equiv \int \Big\{ L(o, \hat{\psi}_{\hat{k}} \mid \eta_0) - L(o, \psi_0 \mid \eta_0) \Big\} \, dP_0(o) \\ &\qquad \text{(observed data loss function)} \\ &= \int \Big\{ L(x, \hat{\psi}_{\hat{k}}) - L(x, \psi_0) \Big\} \, dF_{X,0}(x) \\ &\qquad \text{(full data loss function)} \\ &\longrightarrow 0 \qquad \text{at asymptotically optimal rate.} \end{aligned}$$

The estimator selection problem

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

$$\begin{aligned} d_n(\hat{\psi}_k,\psi_0) &= \int \left\{ L_2(x,\hat{\psi}_k) - L_2(x,\psi_0) \right\} dF_{X,0}(x) \\ &= \int \left(\hat{\psi}_k(w) - \psi_0(w) \right)^2 dF_{W,0}(w). \end{aligned}$$

- For the negative log-likelihood loss function, the risk difference is the Kullback-Leibler divergence between $\hat{\psi}_k$ and ψ_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

$$\tilde{k}_n \quad \equiv \quad \text{argmin}_k \; d_n(\hat{\psi}_k, \psi_0)$$

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

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

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

$$\tilde{\theta}_n(k) \equiv \int L(o, \psi_k(\cdot \mid P_n) \mid \eta_0) dP_0(o)$$

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

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

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

The particular distribution of S_n defines the type of cross-validation procedure.

General framework for cross-validation

Let P^0_{n,S_n} and P^1_{n,S_n} 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

$$\begin{aligned}\hat{k} &\equiv \text{argmin}_k\, E_{S_n} \int L(o, \underbrace{\psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n})}_{\text{Training}} \underbrace{dP^1_{n,S_n}(o)}_{\text{Validation}} \\ &= \text{argmin}_k\, E_{S_n} \sum_{\{i:S_{n,i}=1\}} L(O_i, \psi_k(\cdot \mid P^0_{n,S_n}) \mid \eta^0_{n,S_n}).\end{aligned}$$

Here, $\psi_k(\cdot \mid P^0_{n,S_n})$ and η^0_{n,S_n} 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

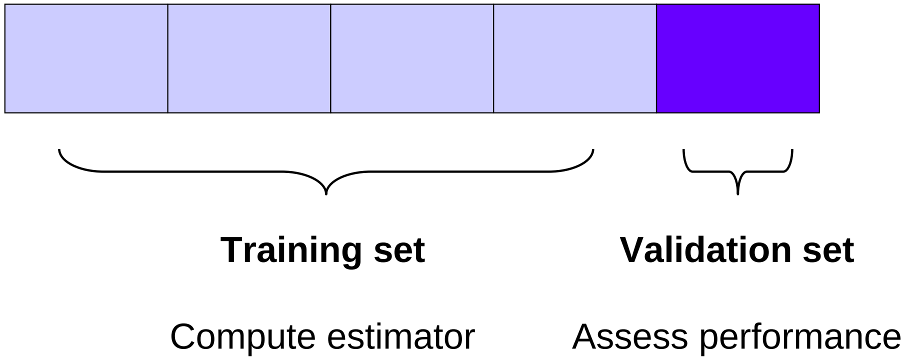


Figure 1: *Five-fold cross-validation.* S_n has 5 realizations.

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 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."*

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

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

$\Longrightarrow$ Predictors are not compared on an equal footing.

Honest cross-validation

Prediction error rates (risk) can 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.

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

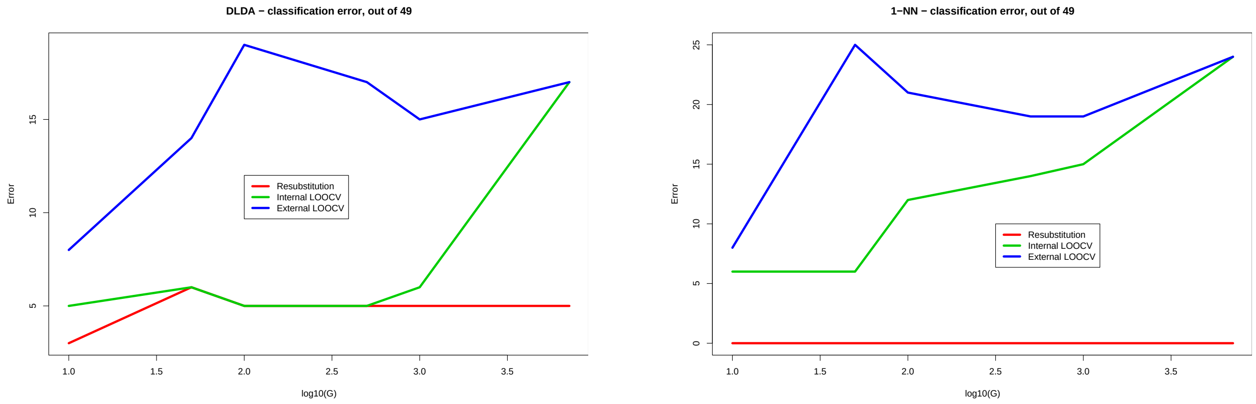
Honest cross-validation

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.

Honest cross-validation



(a) DLDA

(b) 1–NN

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).

Estimator selection using cross-validation

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

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

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

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

Estimator selection using cross-validation

Theorem 1. (Stated in special case of known η_0, $L(O,\psi \mid \eta_0) = L(O,\psi)$).
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

$$\begin{aligned}\int \left\{L(o,\psi_k(\cdot \mid P^0_{n,S_n})) - L(o,\psi_0(\cdot))\right\}^2 dP_0(o) \\ \leq M_2 \int \left\{L(o,\psi_k(\cdot \mid P^0_{n,S_n})) - L(o,\psi_0(\cdot))\right\} dP_0(o) \text{ a.s.}\end{aligned}$$

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

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

Estimator selection using cross-validation

Asymptotic optimality. If

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

then

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

Estimator selection using cross-validation

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

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

and

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

Then,

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

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 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)

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 $np_n \rightarrow \infty$; 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 $\hat{\psi}(\cdot) = \psi(\cdot \mid P_n)$ and loss function $L(O, \psi \mid \eta_0) = L(O, \psi)$ with known η_0.

Cross-validation risk estimator (observable random variable)

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

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

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

Conditional risk, n observations (unknown random variable)

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

Asymptotic risk (unknown parameter)

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

Asymptotic linearity of CV risk estimator

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

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

Then

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

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

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

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

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

Simulation study: Consistency and asymptotic linearity

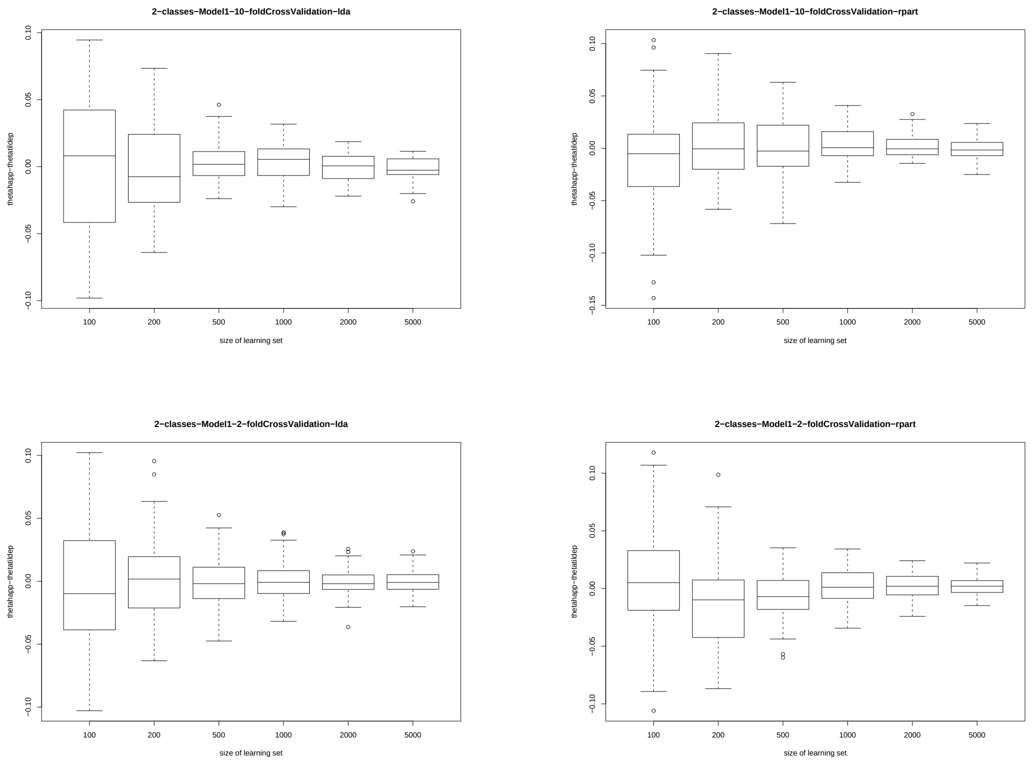


Figure 3: *Convergence to zero of* $\hat{\theta}_{n(1-p)} - \tilde{\theta}_{n(1-p)}$. $X|Y \sim \mathrm{N}(Y1_2, I_2)$, $Y \sim \mathrm{B}(1/2)$, LDA, rpart, two- and ten-fold CV, 200 simulations.

Simulation study: Risk confidence intervals

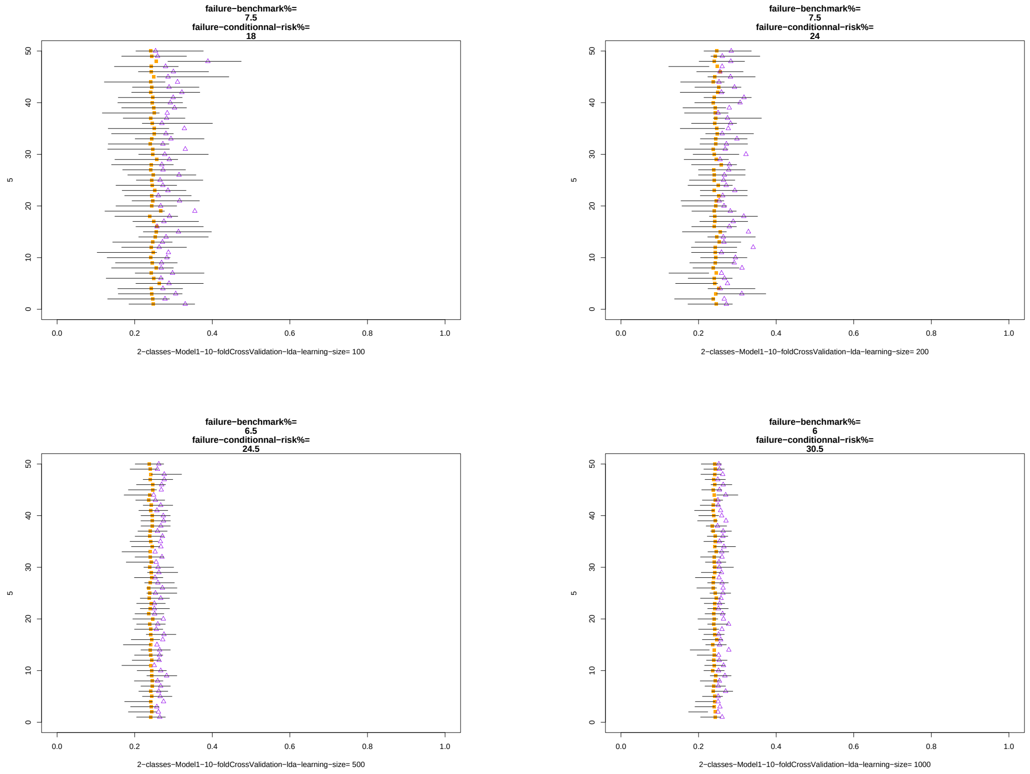


Figure 4: *Risk confidence intervals.* $X|Y \sim \mathrm{N}(Y1_2, I_2)$, $Y \sim \mathrm{B}(1/2)$, $n = 100, 200, 500, 1000$, LDA, ten-fold CV, $\tilde{\theta}_{n(1-p)}$ and $\tilde{\theta}_n$.

Applications of estimation road map

The above results hold for general cross-validation procedures and apply to general distributions P_0, general loss functions $L(O, \psi \mid \eta_0)$, and general estimators $\psi(\cdot \mid P_n)$. 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.
6. Counterfactual prediction in causal inference.

van der Laan & Dudoit (2003)

Example 1: Predictor selection

Suppose we have a learning set of n i.i.d. observations $O = (W, Z) \sim P_0$, 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

$$\begin{aligned} \hat{k} &= \operatorname{argmin}_k \; E_{S_n} \int (z - \psi_k(w \mid P^0_{n,S_n}))^2 dP^1_{n,S_n}(o) \\ &= \operatorname{argmin}_k \; E_{S_n} \sum_{\{i: S_{n,i}=1\}} (Z_i - \psi_k(W_i \mid P^0_{n,S_n}))^2. \end{aligned}$$

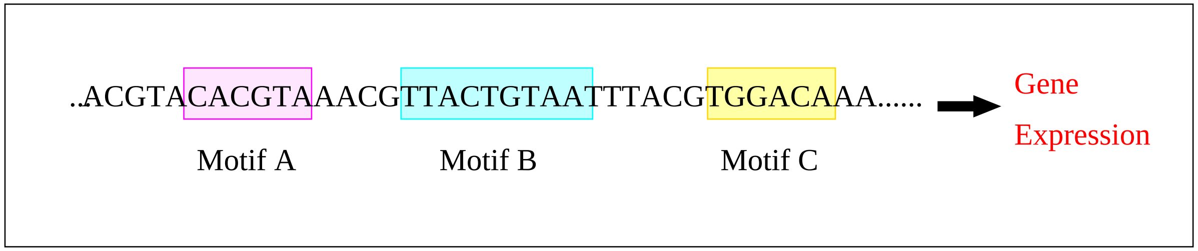
Example 1: Predictor selection

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.

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, $\hat{\psi}_k = f_k(\cdot \mid P_n)$, 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

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

Example 2: Density estimator selection

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

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

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

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

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, $\hat{f}_k(\cdot) = f_k(\cdot \mid P_n)$, for a learning set $x_1, \cdots, x_n$ 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

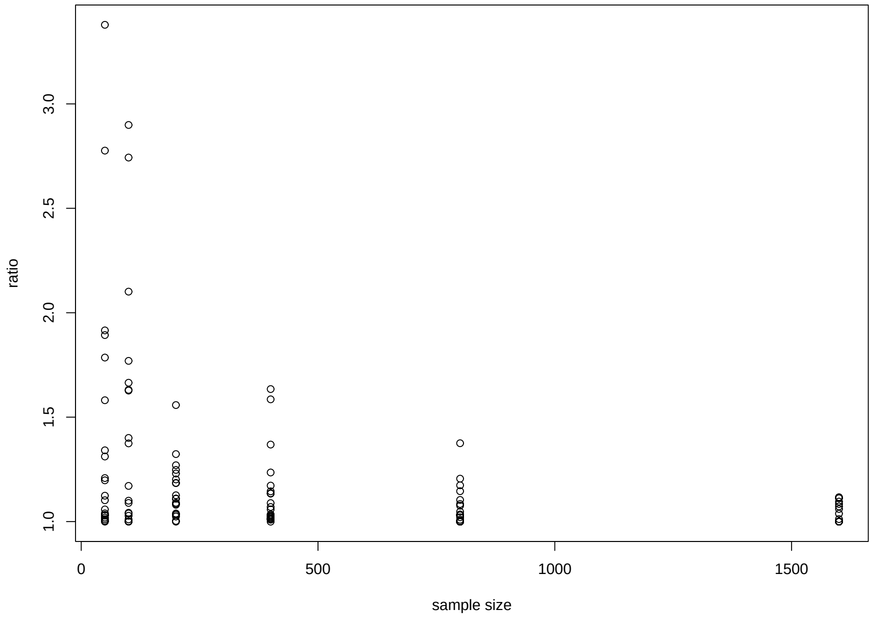


Figure 5: $\frac{d_{n(1-p)}(\hat{\psi}_{\hat{k}},\psi_0)}{d_{n(1-p)}(\hat{\psi}_{\tilde{k}_{n(1-p)}},\psi_0)}$ *vs.* n, *for* $p = 1/10$. The bandwidth $\hat{k}$ was selected using ten-fold CV ($p = 1/10$), for 20 replicate datasets at each of six sample sizes, n.

Example 2: Density estimator selection

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

Table 4: $\frac{\hat{E} d_{n(1-p)}(\hat{\psi}_{\hat{k}}, \psi_0)}{\hat{E} d_{n(1-p)}(\hat{\psi}_{\tilde{k}_{n(1-p)}}, \psi_0)}$ *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

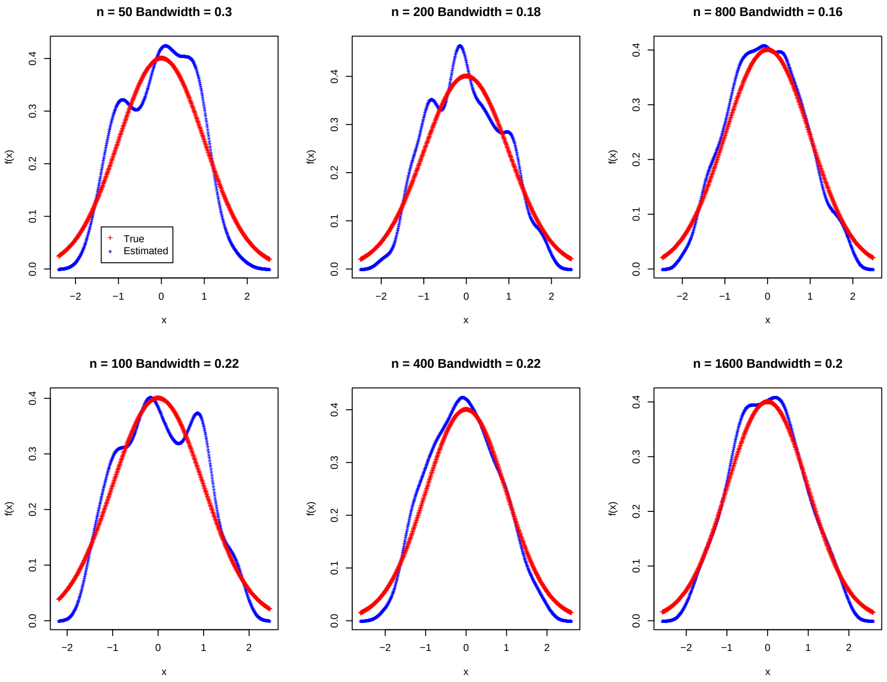


Figure 6: *Cross-validation density estimates* $\hat{f}_{\hat{k}}$ *and true density* f_0. The cross-validation kernel density estimate $f_{\hat{k}}(\cdot \mid P_n)$ is shown for six sample sizes, $n = 50, 100, 200, 400, 800, 1600$, for one simulated dataset. The bandwidth $\hat{k}$ was selected using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

Example 2: Density estimator selection

		n					
		50	100	200	400	800	1600
	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
p	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 5: *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 $p = 1/V$.

Example 2: Density estimator selection

		n					
		50	100	200	400	800	1600
p	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
	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 6: *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.

Example 2: Density estimator selection

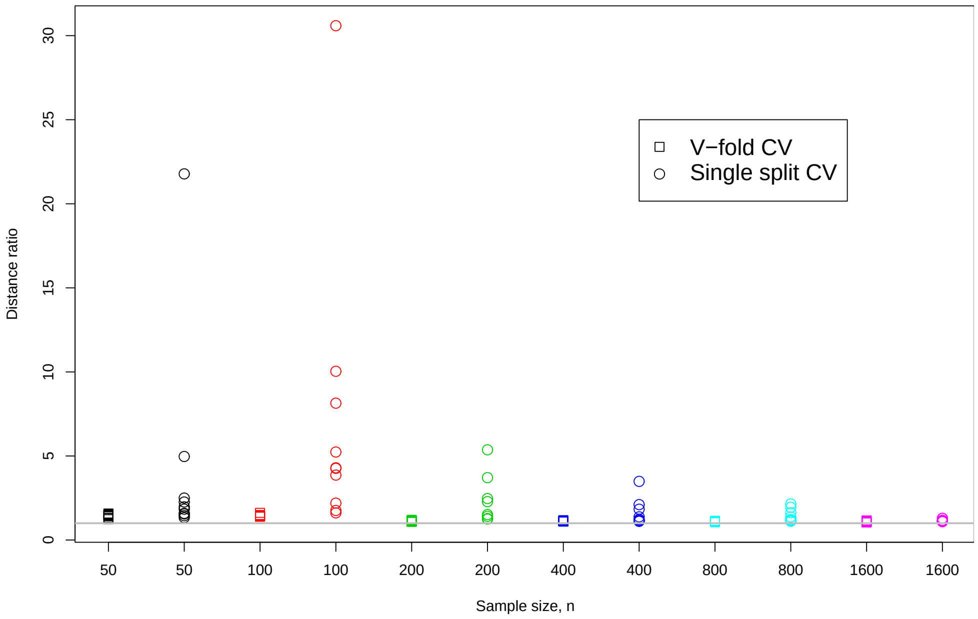


Figure 7: *V-fold vs. single split CV:* $\frac{\hat{E}d_n(\hat{\psi}_{\hat{k}(p)},\psi_0)}{\hat{E}d_n(\hat{\psi}_{\tilde{k}_n},\psi_0)}$ *vs.* n. Estimated distance ratios are based on 20 replicate datasets at six different sample sizes n and for ten different validation set proportions p.

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).

Application 1‘: Identification of regulatory motifs

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

```
>YBR019C TCGGCGATACCTTCACCG
>YBR020W CGGGCGACGATTACCCG
>YLR081W TATCGGAGCGTAGGCGGCCGAAC
>YML051W CGGCATCCTACATGCCG
>YOR120W TCGGTTCAGACAGGTCCGG
```

Application 1: Identification of regulatory motifs

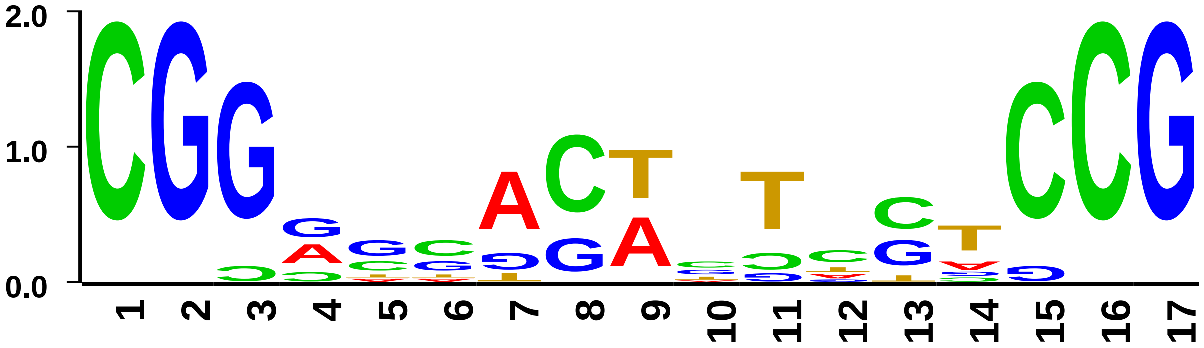


Figure 8: *GAL4 binding sites sequence logo.* www-lmmb.ncifcrf.gov/~toms/sequencelogo.html.

Application 1: Identification of regulatory motifs

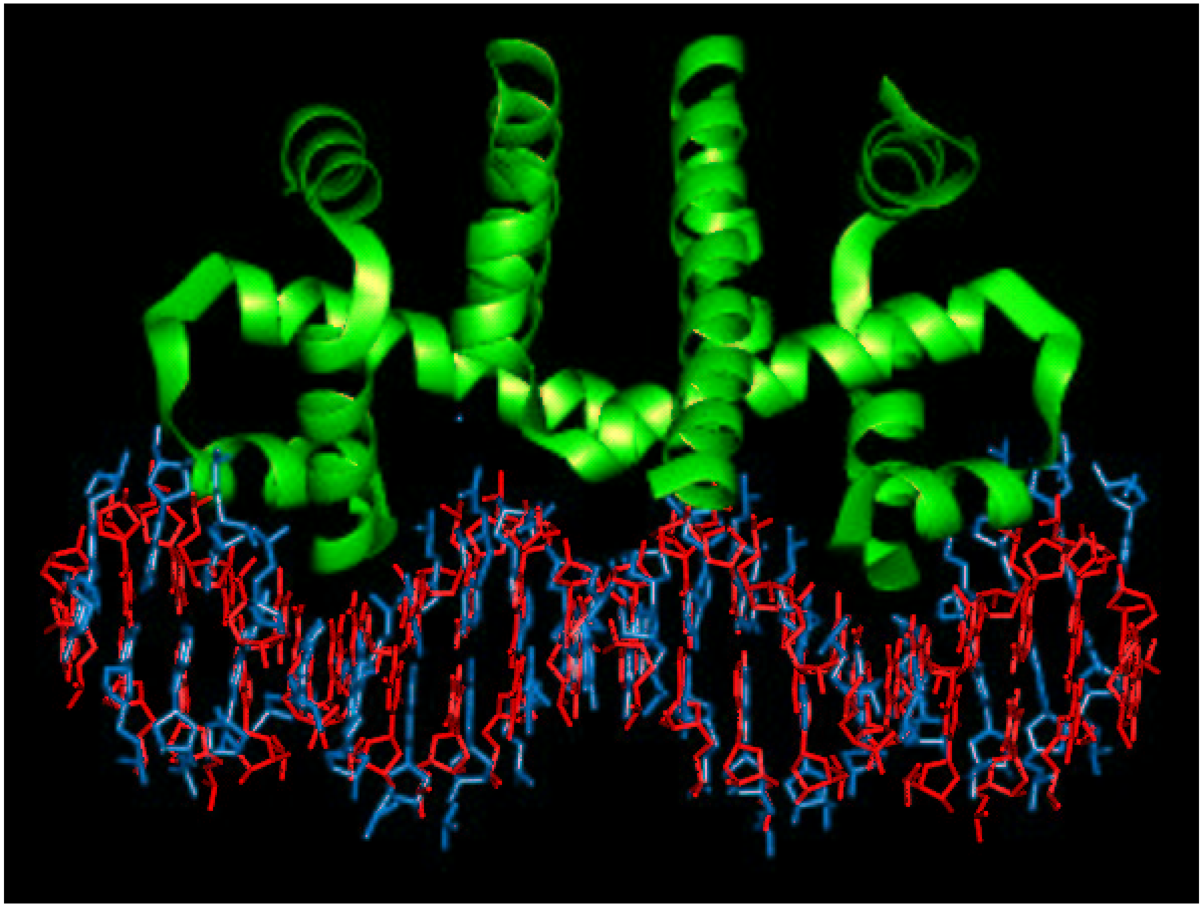


Figure 9: *GAL4 binding.* From www.cryst.bbk.ac.uk/PPS2/.

Application 1: Identification of regulatory motifs

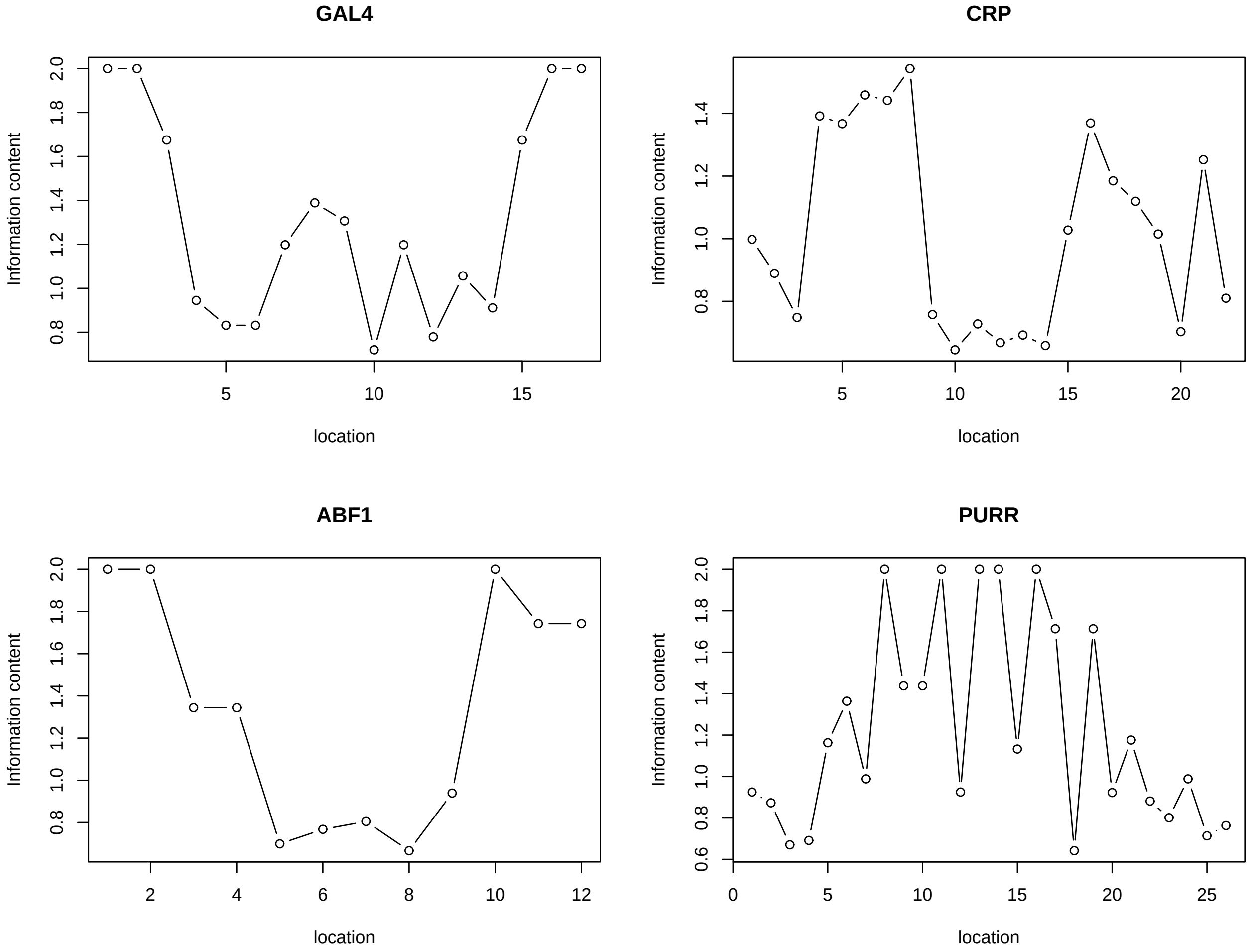


Figure 10: *Information content profiles.* GAL4, CRP, ABF1, PURR.

Application 1: COMODE

Keleş et al. (2003b). COnstrained MOtif DEtection – COMODE. Likelihood-based method for detecting structured regulatory motifs in biological sequences.

- 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.

  $\Longrightarrow$ likelihood-based cross-validation.

Application 1: COMODE

Examples of constraints on PWM.

- Constraints on the information content profile.
  E.g. parametric model such as

  $$IC(w;\phi_1,\phi_2,w^*) = \phi_1 - |w - w^*| \tan\phi_2, \qquad w = 1,\ldots,W.$$

  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. $p_{w1} > 0.8 \Longrightarrow$ preference for nucleotide A at position w.

Application 1: COMODE

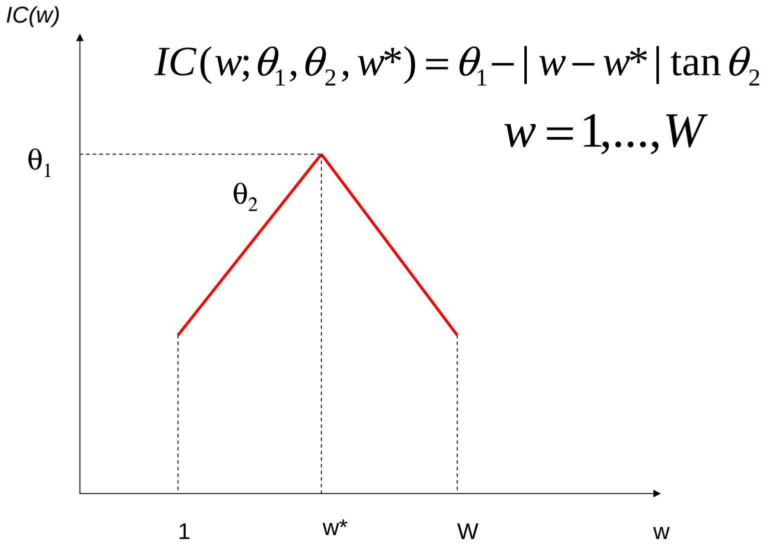


Figure 11: *Example of parameterization for the IC profile of a motif PWM.*

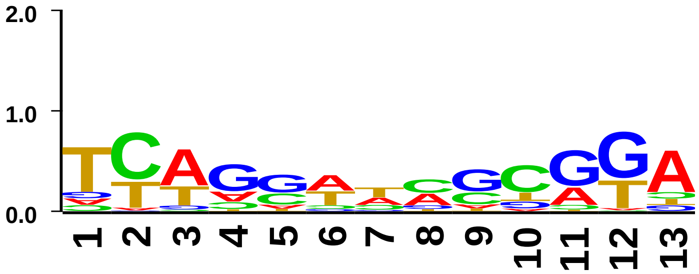
Application 1: COMODE

Input	**Output**
• n unaligned sequences	• Estimated PWM
• Motif length W	• Predicted start site
• PWM constraint functions	for each input sequence

Available from Sündüz Keleş, `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 θ_1 and θ_2).
- c.zoops-II: ordered IC profile, first and last three positions have equal high IC, middle positions have equal low IC, HHHLLLLLLHHH.
- c.zoops-III: piecewise linear IC profile, hat-shaped, mirror image of c.zoops-I (two additional parameters θ_1 and θ_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

$$\widehat{sens}_b = \frac{|K_b \cap \hat{K}_b|}{|K_b|},$$

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: COMODE

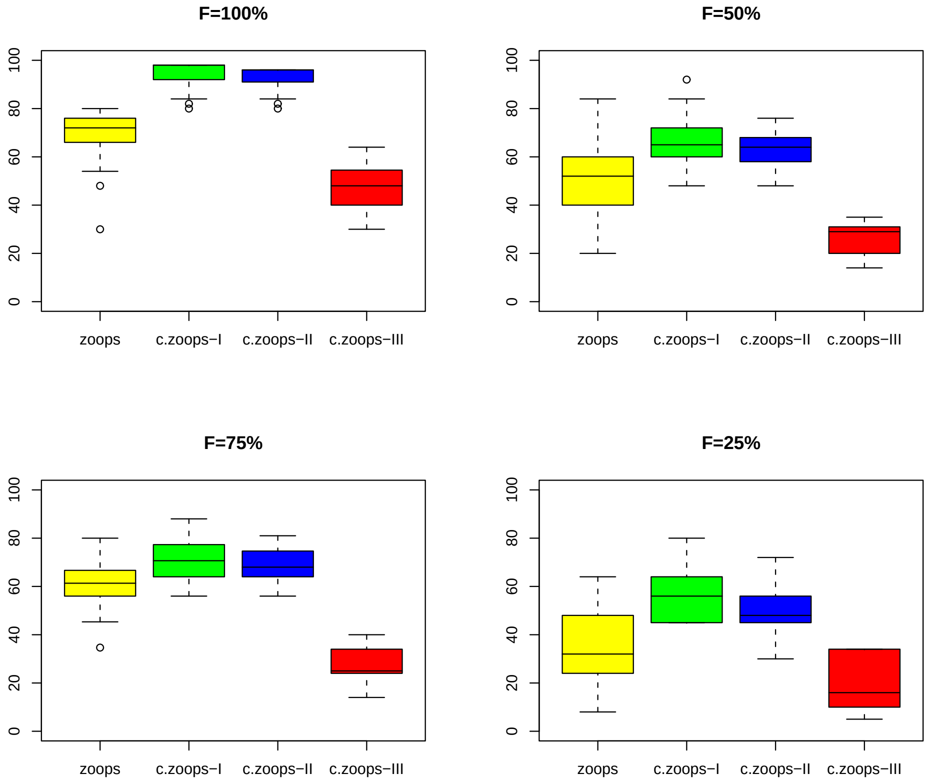


Figure 12: *COMODE.* Boxplot of sensitivity measures for ZOOPS, C.ZOOPS-I, C.ZOOPS-II, and C.ZOOPS-III.

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.

Application 1: Likelihood CV for motif width selection

		$n = 20$		$n = 100$	
		2-fold	5-fold	2-fold	5-fold
	6	0	20	0	22
	7	24	42	0	10
	8	40	10	15	14
	9	11	17	3	3
w	**10**	**121**	**98**	**147**	**142**
	11	0	10	35	9
	12	0	3	0	0
	13	0	0	0	0
	14	0	0	0	0
	15	1	0	0	0

Table 7: *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).

Application 2: Tree-based estimation with censored data

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 ⟶ squared error loss function;
- conditional median of (log) survival time given covariates ⟶ absolute error loss function;
- conditional density (survival or hazard function) of survival time give covariates ⟶ negative log-likelihood loss function.

Application 2: Tree-based estimation with censored data

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.

- Breiman (2003). Partition of time-covariate space using negative log-likelihood loss for a constant hazards model within nodes.
- Davis & Anderson (1989). Negative log-likelihood loss for an exponential model within nodes.
- Gordon & Olshen (1985). L_p, L_p Wasserstein, and Hellinger distances for within-node Kaplan-Meir estimates of survival distribution.
- LeBlanc & Crowley (1992). Negative log-likelihood loss based on first step of a full likelihood estimation procedure for a Cox proportional hazards model within nodes.
  Default method in R `rpart` function (Therneau & Atkinson, 1997).
- 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.

Application 2: Tree-based estimation with censored data

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.

Application 2: Tree-based estimation with censored data

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

Note that in the absence of censoring, i.e., when $\Delta = 1$, then $L(O, \psi \mid \eta_0) = L(X, \psi)$ 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 $L(O, \psi \mid \eta_0)$.
- **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.

Application 2: Tree-based estimation with censored data

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

- *NLL_PH*: negative log-likelihood loss function for Cox proportional hazards model (LeBlanc & Crowley, 1992), `rpart` default for survival data, `method=''exp''`;
- *square_IPCW*: IPCW squared error loss function, `rpart` with `method=''anova''`, `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, and 6000.

  Risk estimates are based on test samples of size $N = 5000$ generated from the full data distribution.

Application 2: Simulation study

Table 8: 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.

Sample size, n	Survival estimation method KM median	 IPCW mean
250	0.9422	0.8838
600	0.9524	0.9062
1250	0.9629	0.9244
6000	0.9767	0.9533

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).

Application 2: Breast cancer survival and CGH copy number

- 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 $\bar{G}_0$ 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 (`coxph` function).
- Overall performance was assessed on the test sample.

Application 2: Breast cancer survival and CGH copy number

Chromosomal Location	
BAC 294	3q26
BAC 1226	10q22
BAC 529	5q11
BAC 542	5q21

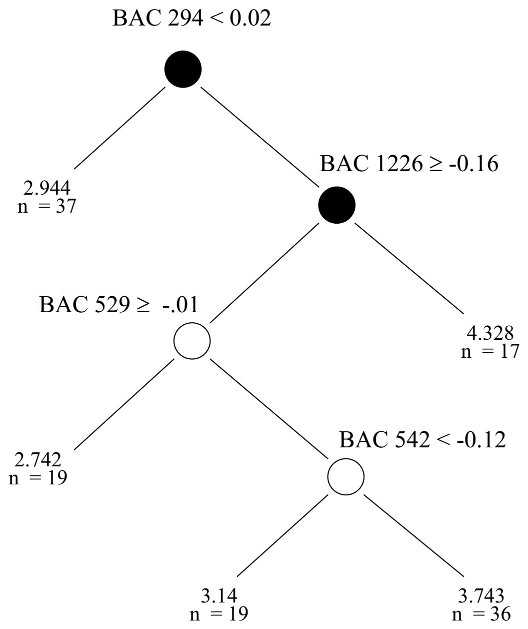


Figure 13: *Breast cancer survival and CGH copy number dataset.* Learning set survival tree, IPCW mean log survival time (in months).

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.

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).
- More aggressive strategies for generating candidate estimators (Step 2): addition/deletion/substitution algorithm (van der Laan & Dudoit, 2003; Sinisi & van der Laan, 2003).
- Loss-based variable importance statistics.
- R package.

References

www.bepress.com/ucbbiostat/

www.bepress.com/sagmb/

- 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. General framework

- 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. CV in prediction

- S. Keleş, M. J. van der Laan, and S. Dudoit (2003a). Asymptotically Optimal Model Selection Method for Regression on Censored Outcomes. Division of Biostatistics, UC Berkeley, Technical Report #124. CV in prediction with censored data

- 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. Likelihood CV

- A. Molinaro, S. Dudoit, and M. J. van der Laan (2003). Tree-based Multivariate Regression and Density Estimation based on Right-censored Data. Division of Biostatistics, UC Berkeley, Technical Report #135. Tree-based estimation with censored data

- S. Keleş, M. J. van der Laan, S. Dudoit, B. Xing, and M. B. Eisen (2003b). Supervised detection of regulatory motifs in DNA sequences. Statistical Applications in Genetics and Molecular Biology, Vol. 2, No. 1, Article 5. Motif finding

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 $G_0(\cdot \mid X)$. Assume $C \perp Y$, given W.

Suppose we have a learning set of n i.i.d. observations of the right-censored data structure
$O = \Big(\min(Y, C),\, \Delta = I(Y \leq C),\, W\Big) \sim P_0 = P_{F_{X,0},G_0}$.

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 $\psi_0(W) = E_0[Y \mid W]$.

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

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 $IC(O \mid Q(F_X), G, D)$, indexed by nuisance parameters G and (possibly) $Q(F_X)$. The estimating functions satisfy

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

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

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

Example 3: Predictor selection for right-censored outcomes

Inverse probability of censoring weighted (IPCW) estimating function

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

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

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

Example 3: Predictor selection for right-censored outcomes

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

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

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.

Example 3: Predictor selection for right-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 $Y = \log T = W^2 + \epsilon$, $\epsilon \sim \mathrm{N}(0, \sigma^2)$, $\sigma^2 = 2$, 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 $k = 1, \ldots, K_n$, the unit interval is divided into k bins with width $1/k$ each.

Example 3: Predictor selection for right-censored outcomes

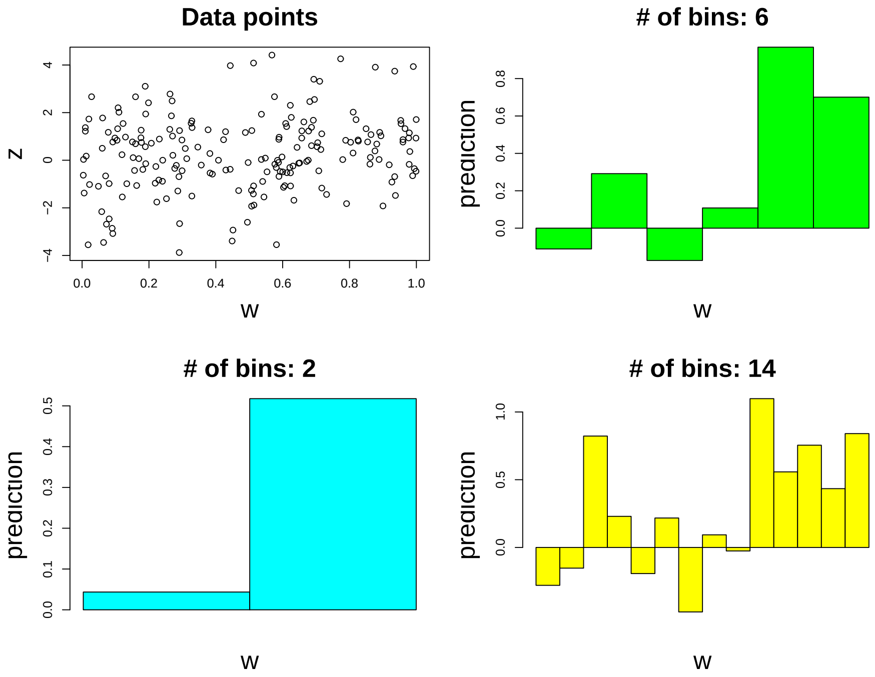


Figure 14: *Histogram regression.* Predictors are indexed by the number of bins and the prediction for a given bin is the mean outcome for observations in that bin.

Example 3: Predictor selection for right-censored outcomes

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

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

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

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

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

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

Example 3: Predictor selection for right-censored outcomes

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

Table 9: *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).

CROSS-VALIDATED DELETION/SUBSTITUTION/ADDITION ALGORITHM FOR PREDICTION: APPLICATIONS IN GENOMICS

Mark van der Laan

www.stat.berkeley.edu/ laan

Joint work with Sandrine Dudoit, Sandra Sinisi, Annette Molinaro, Mike Eisen.

Division of Biostatistics, University of California, Berkeley.

www.bepress.com/ucbbiostat/

October 8, 2003

The Constance van Eeden Lecture

University of Britisch Columbia, Vancouver

MOTIVATION

1. Prediction of clinical outcomes based on epidemiological and genomic data such as gene expression/ single nucleotide polymorphism (SNP)/ comparative genomic hybridization (CGH).
2. Prediction of gene-expression from regularitory-DNA-sequence.
3. And so on!

Interesting features: High dimensional covariates, censored clinical outcomes such as survival.

OVERVIEW

- The **optimal predictor** in terms of loss function.
- **Selection** among candidate data-based predictors (estimators): Cross-validation selector, theory, take home lesson.
- **Parametrizing** predictors as linear combinations of basis functions (i.e., choose a Sieve).
- **Construction** of a candidate data dependent predictor for each subset of basis functions.
- **Minimizing** criteria (cross-validated risk/empirical risk) for subset-specific predictor over all possible subsets of basis functions: Deletion/Substitution/Addition algorithm.
- **Generalizing** to censored data.

OPTIMAL PREDICTOR IN TERMS OF LOSS FUNCTION

Let $O_1 = (Y_1, W_1), \ldots, O_n = (Y_n, W_n)$ be n i.i.d. observations of $O = (Y, W) \sim P_0$, where Y denotes an outcome of interest and W is a d-dimensional vector of covariates. Let $\mathcal{M}$ be a model for P_0: that is, it is given that $P_0 \in \mathcal{M}$.

Predictor: A function $W \rightarrow \psi(W)$ from W to an outcome.

Loss function: Let $L(O, \psi) = (Y - \psi(W))^2$ be the squared error loss function for a candidate predictor ψ.

Risk of predictor: The risk of a predictor $W \rightarrow \psi(W)$ equals the expected loss (w.r.t. the true distribution P_0).

Optimal predictor: $\psi_0(W) = E_{P_0}(Y \mid W)$ is the optimal (minimal risk) predictor over a set $\mathbf{\Psi}$ (e.g., parameter space implied by model $\mathcal{M}$) of allowed predictors:

$$\begin{aligned} \psi_0 &= \text{argmin}_{\psi \in \mathbf{\Psi}} E_0 L(O, \psi) \\ &= \text{argmin}_{\psi \in \mathbf{\Psi}} \int (Y - \psi(W))^2 dP_0(Y, W). \end{aligned}$$

Risk “distance” : For a given predictor $W \rightarrow \psi(W)$, we have that its risk minus the optimal risk equals the expected squared deviation $\psi(W) - \psi_0(W)$:

$$\begin{aligned} d(\psi, \psi_0) &\equiv \int \{L(O, \psi) - L(O, \psi_0)\}\, dP_0(O) \\ &= \int (\psi(W) - \psi_0(W))^2 dP_0(W). \end{aligned}$$

SELECTION

Let P_n be the empirical distribution of the observed sample $O_1, \ldots, O_n$.

Estimator: An estimator of the optimal predictor $\psi_0(W)$ is a mapping (i.e., an algorithm) from P_n into a particular predictor in $\mathbf{\Psi}$. Notation: $P_n \rightarrow \psi(P_n) \in \mathbf{\Psi}$.

Candidate estimators: Let $P_n \rightarrow \psi_k(P_n) \in \mathbf{\Psi}$, $k = 1, \ldots, K(n)$, be a collection of estimators of ψ_0 (i.e., algorithms which map data into a predictor).

THE ORACLE SELECTOR

The oracle selector $\tilde{k}_n$ chooses the estimator with minimal (true) risk. Equivalently

$$\tilde{k}_n = \text{argmin}_k \int (\psi_k(P_n)(W) - \psi_0(W))^2 dP_0(W).$$

Since risk depends on the true distribution P_0 this selector is not available in practice.

Asymptotic equivalence with oracle selector: Given the $K(n)$ candidate estimators, a selector $\hat{k} = \hat{k}(P_n) \in \{1, \ldots, K(n)\}$ is asymptotically equivalent with the oracle selector if the risk of the estimator chosen by the selector approaches (when sample size converges to infinity) as fast to the optimal risk of ψ_0 as the risk of the estimator chosen by the oracle selector: that is,

$$\frac{d(\psi_{\hat{k}}(P_n), \psi_0)}{d(\psi_{\tilde{k}_n}(P_n), \psi_0)} \rightarrow 1 \text{ in probability.}$$

THE CROSS-VALIDATION SELECTOR

Empirical risk estimate: Given an estimator $\psi(P_n)$, the empirical risk estimate is simply the empirical mean of the squared error loss $L(O, \psi(P_n)) = (Y - \psi(P_n)(W))^2$:

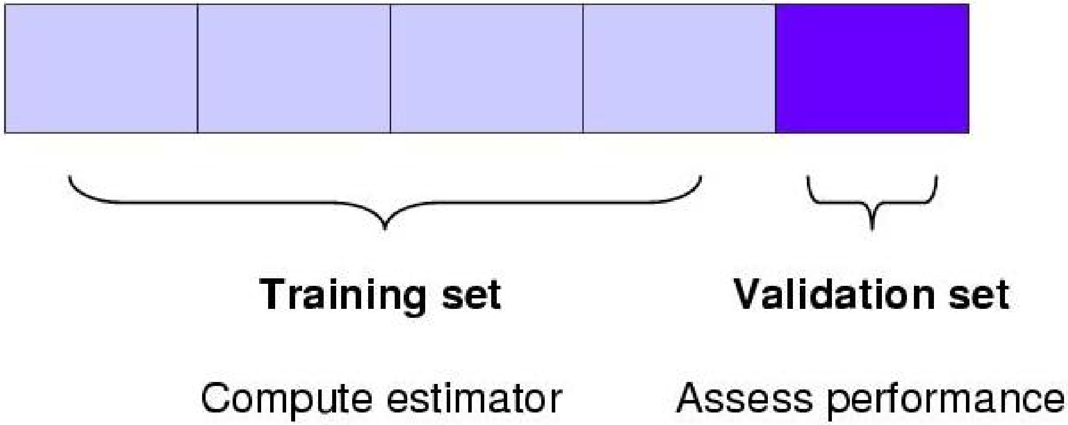
$$\frac{1}{n}\sum_{i=1}^{n} L(O_i, \psi(P_n)).$$

Cross-validated risk estimate: In this case, one applies the estimator to a part of the sample (training sample) and one computes the average loss of the obtained estimator over the remaining sample (validation sample). One averages this risk estimate over a particular number of splits of the sample.

Formally, define a random vector $S_n \in \{0,1\}^n$ for splitting the sample into a validation and a training sample.

$$S_{n,i} = \begin{cases} 0 & \text{if} \quad \text{i-th observation is in the training sample} \\ 1 & \text{if} \quad \text{i-th observation is in the validation sample} \end{cases}$$

Different choices of S_n cover all types of cross-validation schemes including $V-$ fold cross-validation, monte carlo cross validation, and bootstrap cross-validation. For example, in 5-fold cross-validation S_n has 5 possible outcomes.



Training set

Compute estimator

Validation set

Assess performance

Let $p = n_1/n$ be the proportion constituting the validation sample.

Let P^0_{n,S_n}, P^1_{n,S_n} be the empirical distributions of the training and validation sample, respectively.

The cross-validated risk estimate of a candidate estimator $P_n \rightarrow \psi_k(P_n)$ is defined by:

$$E_{S_n} \frac{1}{np} \sum_{i:S_n(i)=1} (Y_i - \psi_k(P^0_{n,S_n})(W_i))^2.$$

Cross-validation selector: The cross-validation selector chooses the estimator minimizing the cross-validated risk estimate of the risk of $\psi_k(P_n)$, $k = 1, \ldots, K(n)$.

EQUIVALENCE WITH ORACLE SELECTOR

If 1) the proportion $p = p(n)$ constituting the validation sample converges to zero with sample size n, and 2) the **logarithm** of the number of estimators, $K(n)$, **divided by** the validation sample size, np, converges faster to zero than the risk distance of the oracle choice estimator and ψ_0, then the cross-validation selector $\hat{k}$ is asymptotically equivalent (and thus optimal) with the oracle selector.

Sensitivity to proportion p: Simulations (and theoretical argument) show that, in practice, the sensitivity of the performance of the cross-validation selector to the choice of p is remarkably low: e.g. 2-fold performs well!

DESCRIBING/PARAMETRIZING PREDICTORS

We describe any of the allowed predictors in $\boldsymbol{\Psi}$ with linear combinations $W \to \sum_{j \in I} \beta_j \Phi_j(W)$ of basis functions $W \to \Phi_j(W)$ indexed by an index set I.

Tensor products of univariate basis functions: For example, if we use a polynomial basis, then for each $\vec{p} = (p_1, \ldots, p_d)$, we have a basis function $\phi_{\vec{p}}(X) = X_1^{p_1} \cdots X_d^{p_d}$.

Each index set $I = \{\vec{p}_1, \ldots, \vec{p}_k\}$, corresponds now with a linear regression model in variables being tensor products of polynomial powers.

Indicators of sets of a partition: Let $\mathcal{W}$ be the covariate space. Given a region R in $\mathcal{W}$, let $\Phi_R(\cdot) = I(\cdot \in R)$ be the indicator of this region. Each partition $I = \{R_1, \ldots, R_k\}$ of $\mathcal{W}$ corresponds with a linear regression model in variables being indicators of sets R_j: thus, a histogram regression model.

SUBSET SPECIFIC LEAST-SQUARES ESTIMATOR

For each index set I (indicating tensor products of univariate basis functions, or indicators of sets corresponding with a partition), let

$$\Psi_I(P_n)$$

be the minimizer of residual sum of squared errors (i.e., empirical mean of squared error loss function) over the linear regression model $\{\psi_{I,\beta} : \beta\}$ corresponding with the subset of basis functions identified by I.

ESTIMATING THE SUBSET OF BASIS FUNCTIONS

The optimal subset would be the minimizer over all subsets I of the true risk (say) $f_0(I)$ of the corresponding estimator $\psi_I(P_n)$. So we need to estimate this true risk function.

Subset Estimator 1: Minimize the empirical risk estimate over all subsets of basis functions of size k, but choose k by minimizing the cross-validated risk estimate.

Subset Estimator 2: Minimize the cross-validated risk estimate over all subsets of basis functions. That is, minimize

$$I \rightarrow E_{S_n} \frac{1}{np} \sum_{i:S_n(i)=1} \{Y_i - \Psi_I(P^0_{n,S_n})(W_i)\}^2.$$

Below, we specify a DELETION/SUBSTITUTION/ADDITION (D/S/A) algorithm for minimizing over I the empirical risk estimate $f_{RSS}(I)$ or cross-validated risk estimate $f_{CV.RSS}(I)$ of $\psi_I(P_n)$.

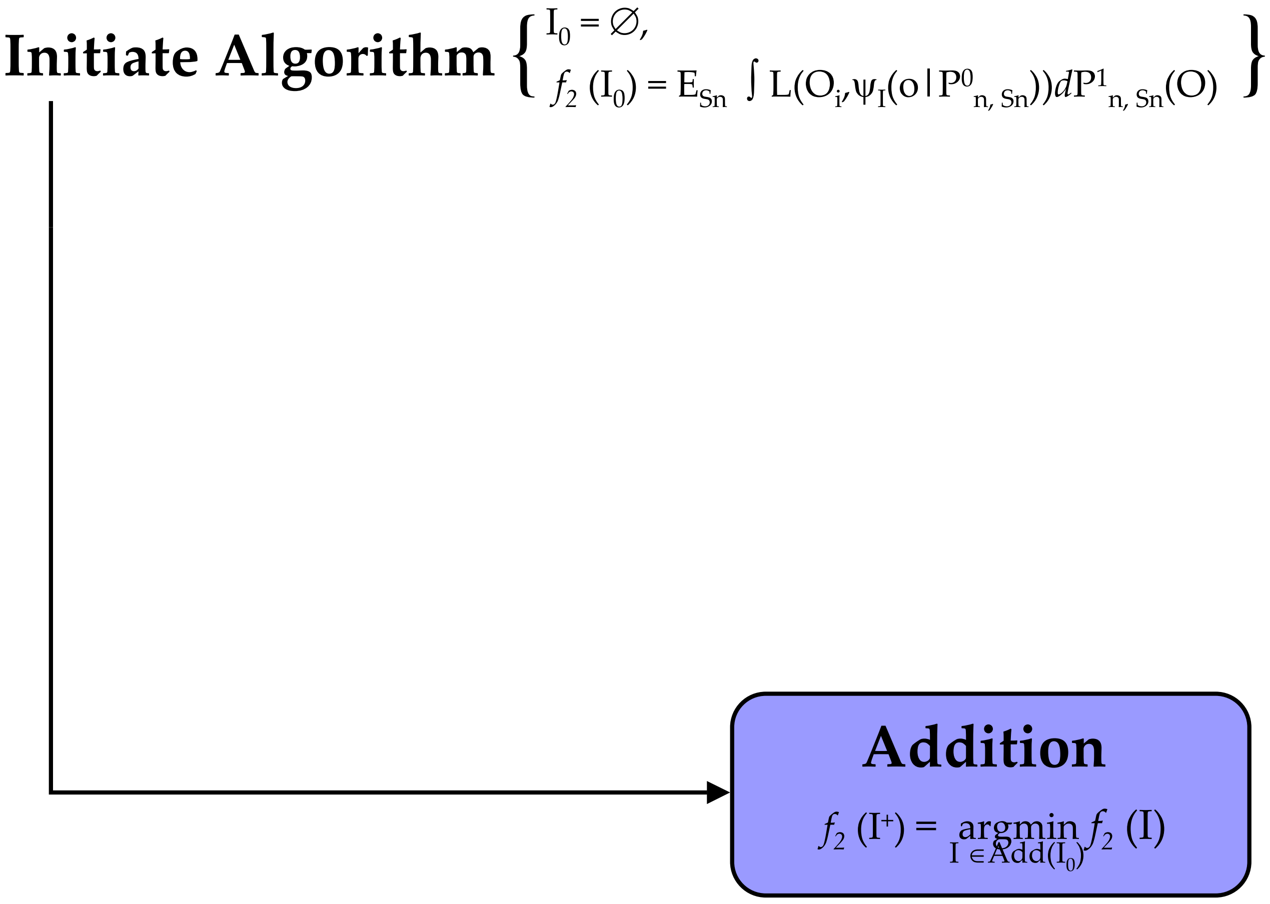
DELETION/SUBSTITUTION/ADDITION ALGORITHM

The D/S/A algorithm aims to minimize a function $I \rightarrow f(I)$ (e.g., f_{RSS}, $f_{CV.RSS}$) over subsets of basis functions, and is defined by three set functions $DEL(I_0)$, $SUB(I_0)$, and $ADD(I_0)$, which maps a current subset I_0 into a collection of subsets of size $\mid I_0 \mid -1$ (deletion moves), $\mid I_0 \mid$ (substitution moves), and $\mid I_0 \mid +1$ (addition moves), respectively.

ALGORITHM

$$\textbf{Initiate Algorithm} \left\{ \begin{array}{l} \mathrm{I}_0 = \varnothing, \\ f_2\,(\mathrm{I}_0) = \mathrm{E}_{\mathrm{Sn}} \int \mathrm{L}(\mathrm{O}_i, \psi_{\mathrm{I}}(\mathrm{o} \mid \mathrm{P}^0{}_{\mathrm{n,\,Sn}}))d\mathrm{P}^1{}_{\mathrm{n,\,Sn}}(\mathrm{O}) \end{array} \right\}$$

ALGORITHM

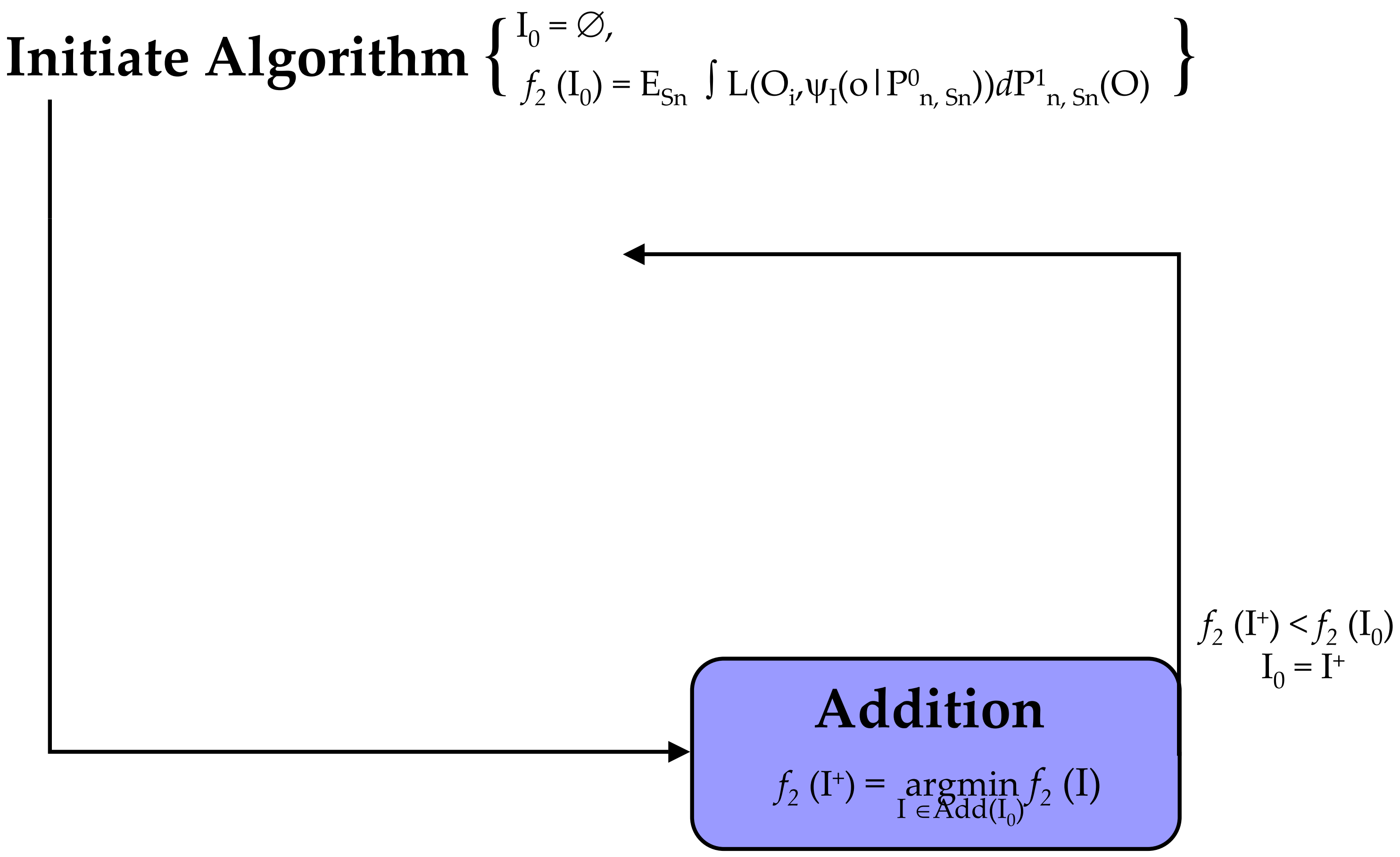


Initiate Algorithm $\left\{ \begin{array}{l} I_0 = \emptyset, \\ f_2\,(I_0) = E_{Sn} \int L(O_i, \psi_I(o \,|\, P^0{}_{n,\,Sn}))\,dP^1{}_{n,\,Sn}(O) \end{array} \right\}$

Addition

$$f_2\,(I^+) = \underset{I \in Add(I_0)}{\operatorname{argmin}} f_2\,(I)$$

ALGORITHM



Initiate Algorithm $\left\{ \begin{array}{l} I_0 = \emptyset, \\ f_2\,(I_0) = E_{Sn} \int L(O_i, \psi_I(o \mid P^0_{n,\,Sn}))\,dP^1_{n,\,Sn}(O) \end{array} \right\}$

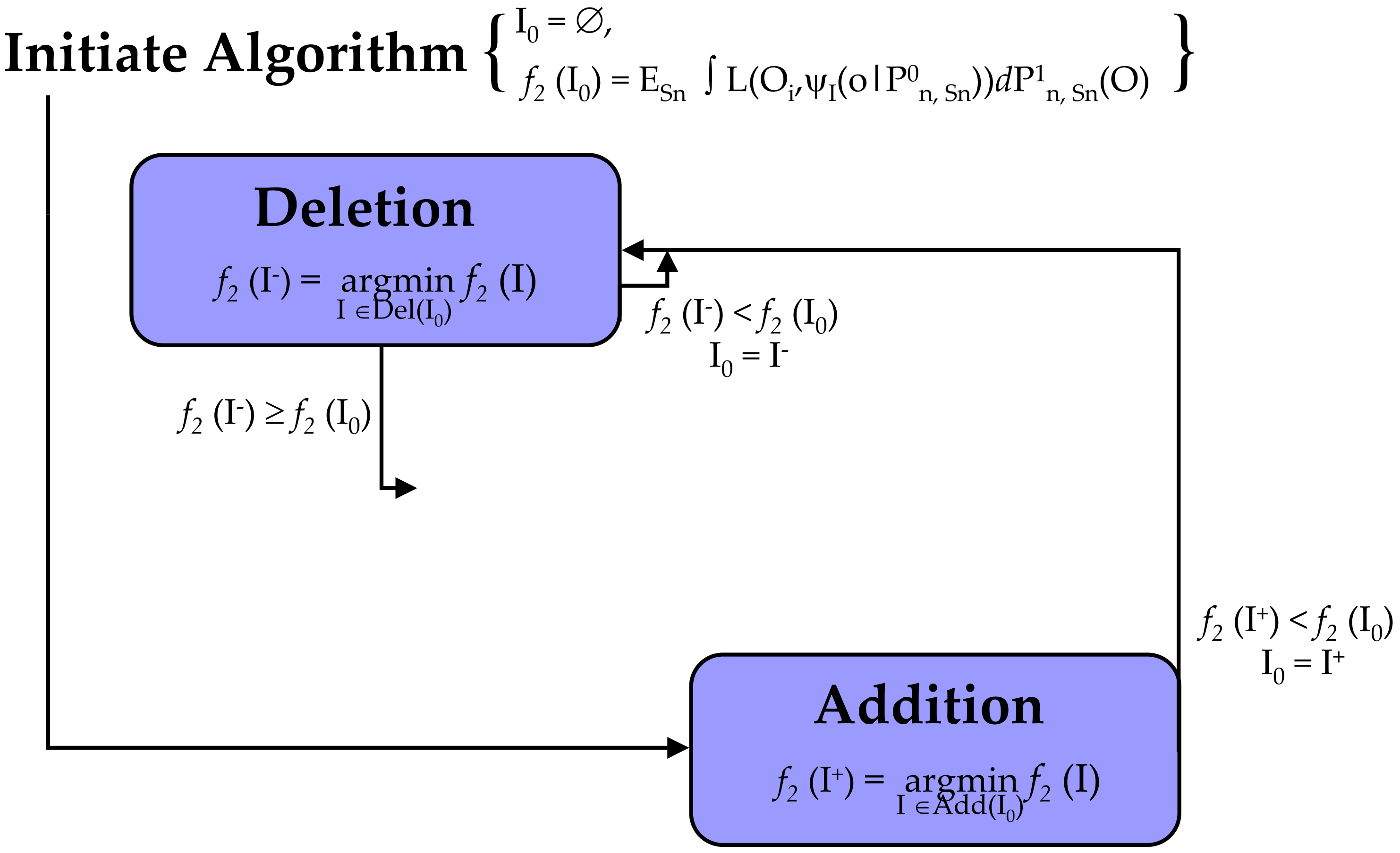
$f_2\,(I^+) < f_2\,(I_0)$

$I_0 = I^+$

Addition

$f_2\,(I^+) = \underset{I \in Add(I_0)}{\text{argmin}}\, f_2\,(I)$

ALGORITHM



Initiate Algorithm $\left\{ \begin{array}{l} I_0 = \emptyset, \\ f_2\,(I_0) = E_{Sn} \int L(O_i, \psi_I(o \mid P^0_{n,\,Sn}))dP^1_{n,\,Sn}(O) \end{array} \right\}$

Deletion

$f_2\,(I^-) = \underset{I \in Del(I_0)}{\text{argmin}}\, f_2\,(I)$

$f_2\,(I^-) < f_2\,(I_0)$

$I_0 = I^-$

$f_2\,(I^-) \geq f_2\,(I_0)$

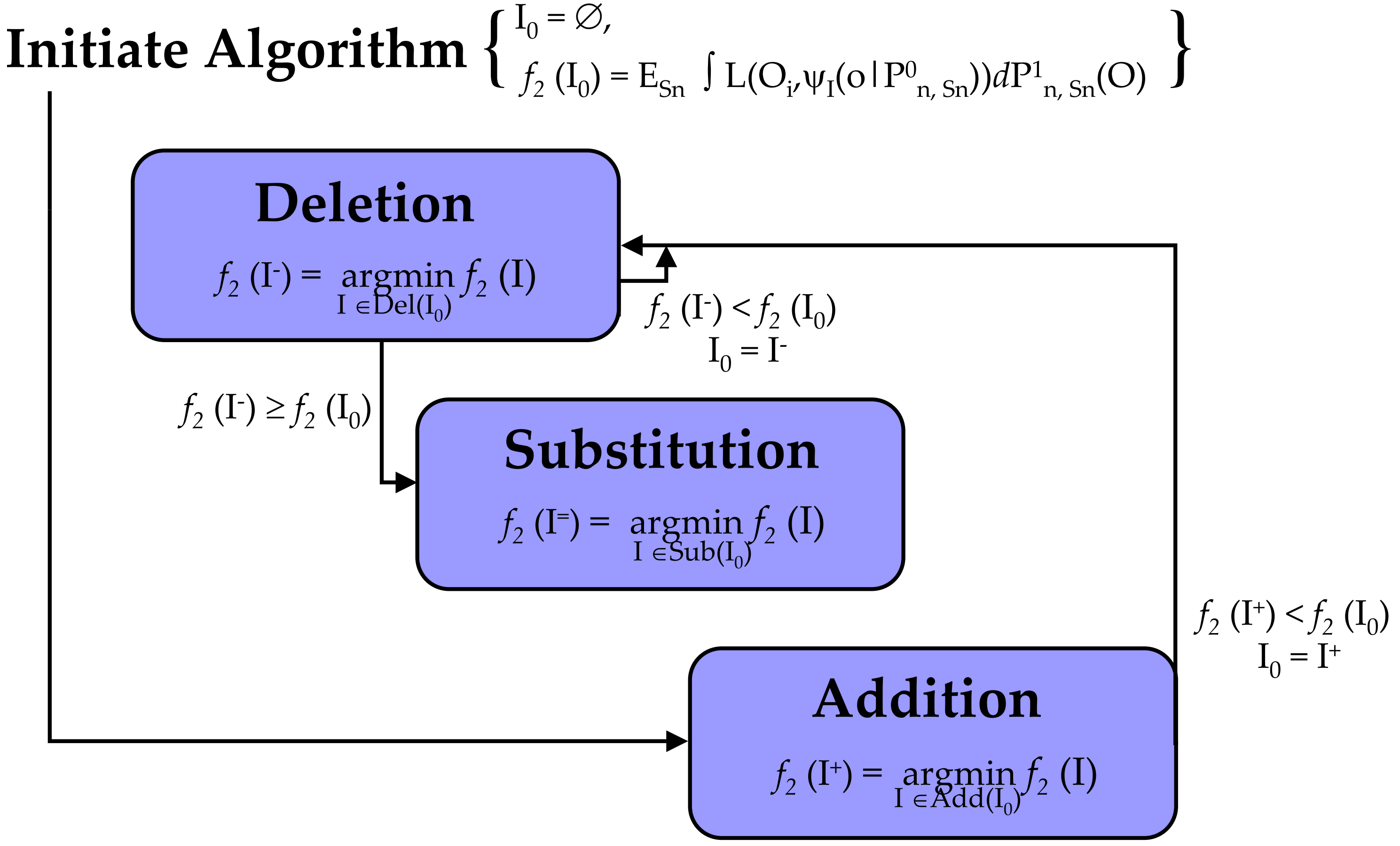
$f_2\,(I^+) < f_2\,(I_0)$

$I_0 = I^+$

Addition

$f_2\,(I^+) = \underset{I \in Add(I_0)}{\text{argmin}}\, f_2\,(I)$

ALGORITHM



Initiate Algorithm $\left\{ \begin{array}{l} I_0 = \emptyset, \\ f_2\,(I_0) = E_{Sn} \int L(O_i, \psi_I(o \mid P^0_{n,\,Sn}))\,dP^1_{n,\,Sn}(O) \end{array} \right\}$

Deletion

$f_2\,(I^-) = \underset{I \in Del(I_0)}{\operatorname{argmin}} f_2\,(I)$

$f_2\,(I^-) < f_2\,(I_0)$

$I_0 = I^-$

$f_2\,(I^-) \geq f_2\,(I_0)$

Substitution

$f_2\,(I^=) = \underset{I \in Sub(I_0)}{\operatorname{argmin}} f_2\,(I)$

$f_2\,(I^+) < f_2\,(I_0)$

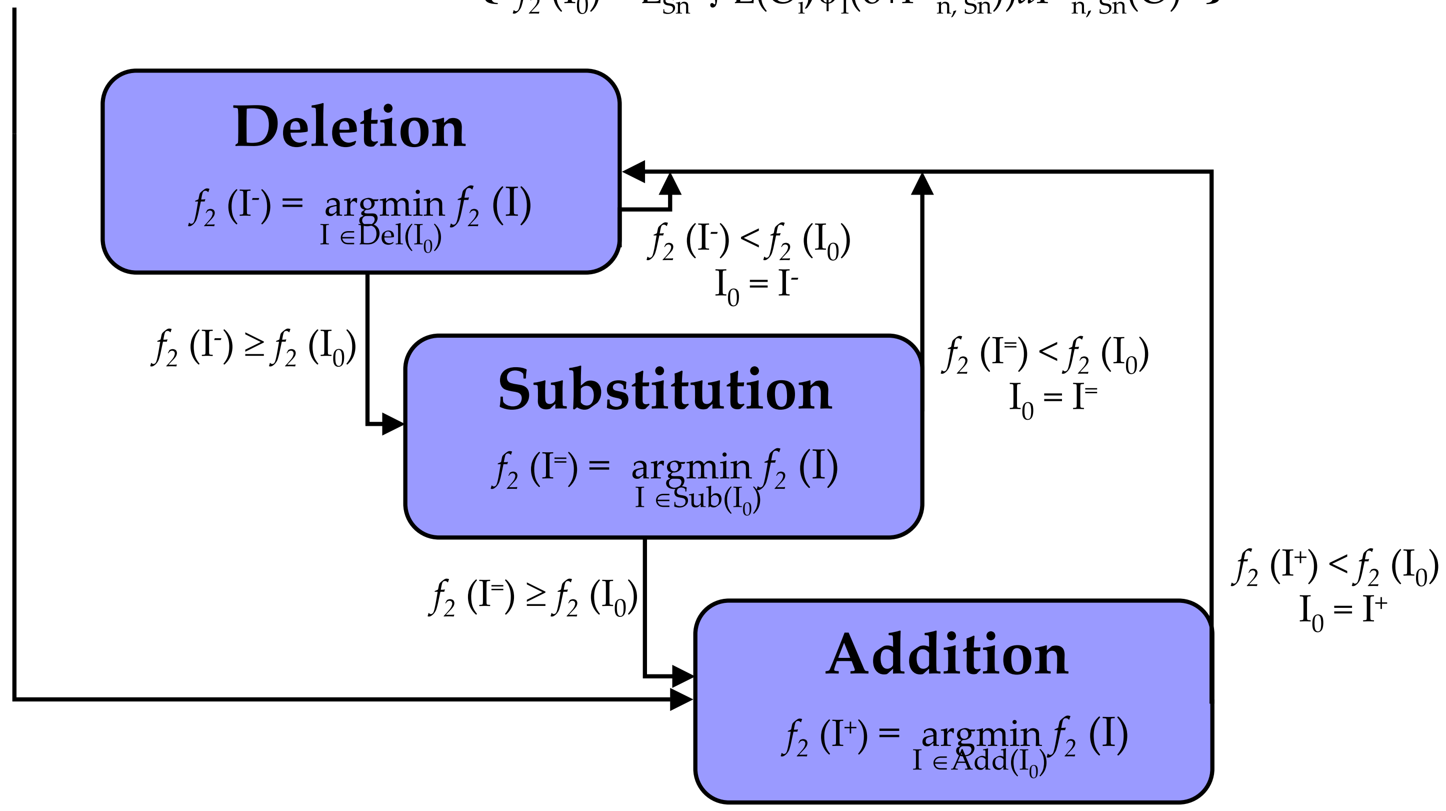
$I_0 = I^+$

Addition

$f_2\,(I^+) = \underset{I \in Add(I_0)}{\operatorname{argmin}} f_2\,(I)$

ALGORITHM

Initiate Algorithm $\left\{ \begin{array}{l} I_0 = \varnothing, \\ f_2\,(I_0) = E_{Sn} \int L(O_i, \psi_I(o \mid P^0{}_{n,\,Sn}))dP^1{}_{n,\,Sn}(O) \end{array} \right\}$



Deletion

$f_2\,(I^-) = \underset{I \in Del(I_0)}{\operatorname{argmin}} f_2\,(I)$

$f_2\,(I^-) < f_2\,(I_0)$
$I_0 = I^-$

$f_2\,(I^-) \geq f_2\,(I_0)$

Substitution

$f_2\,(I^=) = \underset{I \in Sub(I_0)}{\operatorname{argmin}} f_2\,(I)$

$f_2\,(I^=) < f_2\,(I_0)$
$I_0 = I^=$

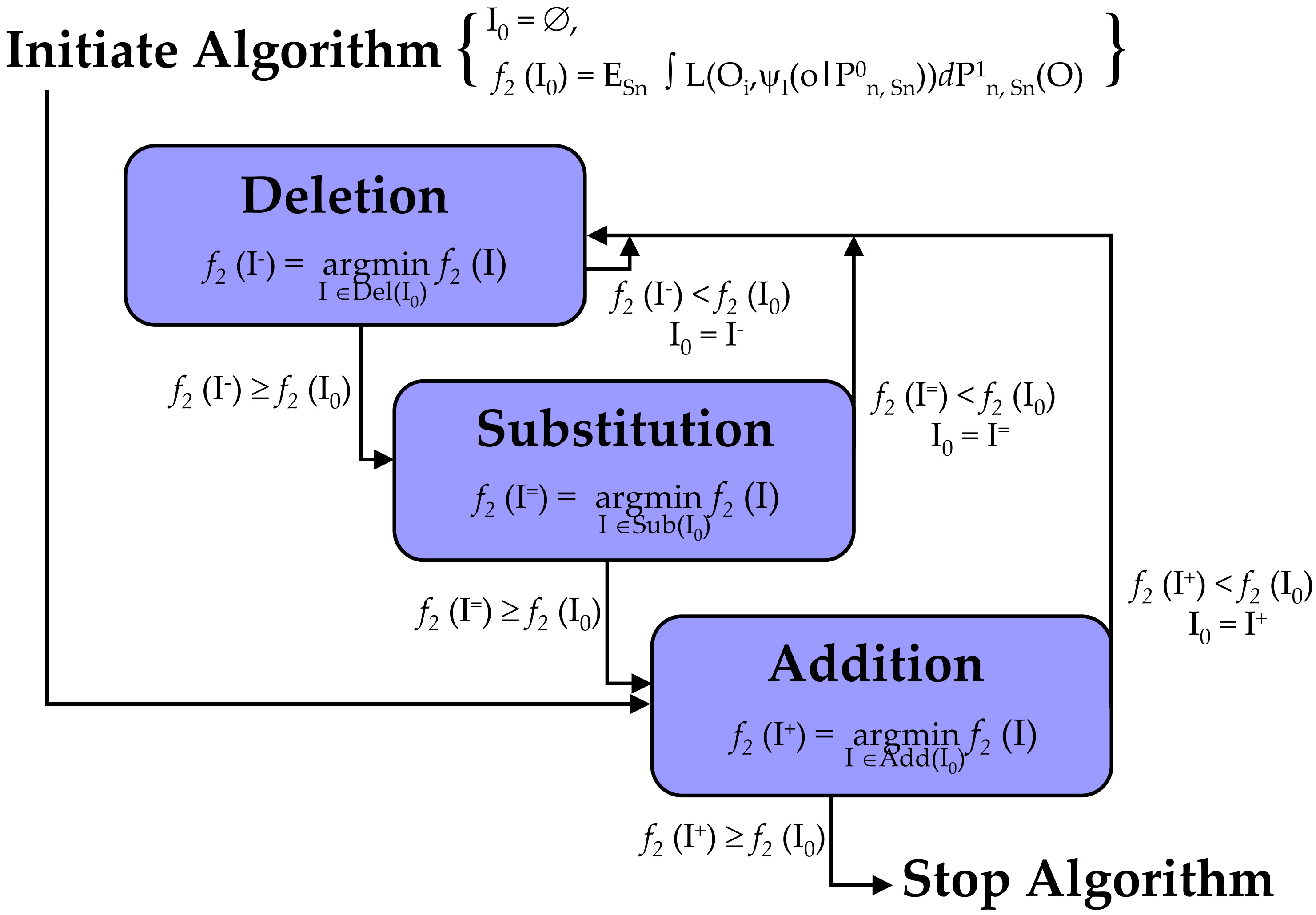
$f_2\,(I^=) \geq f_2\,(I_0)$

Addition

$f_2\,(I^+) = \underset{I \in Add(I_0)}{\operatorname{argmin}} f_2\,(I)$

$f_2\,(I^+) < f_2\,(I_0)$
$I_0 = I^+$

ALGORITHM



Initiate Algorithm $\left\{ \begin{array}{l} I_0 = \varnothing, \\ f_2\,(I_0) = E_{Sn} \int L(O_i, \psi_I(o \mid P^0{}_{n,\,Sn}))\,dP^1{}_{n,\,Sn}(O) \end{array} \right\}$

Deletion

$f_2\,(I^-) = \underset{I\,\in Del(I_0)}{\operatorname{argmin}} f_2\,(I)$

$f_2\,(I^-) < f_2\,(I_0)$

$I_0 = I^-$

$f_2\,(I^-) \geq f_2\,(I_0)$

Substitution

$f_2\,(I^=) = \underset{I\,\in Sub(I_0)}{\operatorname{argmin}} f_2\,(I)$

$f_2\,(I^=) < f_2\,(I_0)$

$I_0 = I^=$

$f_2\,(I^=) \geq f_2\,(I_0)$

$f_2\,(I^+) < f_2\,(I_0)$

$I_0 = I^+$

Addition

$f_2\,(I^+) = \underset{I\,\in Add(I_0)}{\operatorname{argmin}} f_2\,(I)$

$f_2\,(I^+) \geq f_2\,(I_0)$

Stop Algorithm

PROPOSAL FOR TENSOR PRODUCT MOVES

Deletion moves. $DEL(I_0)$ maps into the k subsets of size $k-1$ corresponding with deleting one of the k basis functions in I_0.

Substitution moves. Given a basis function indexed by $\vec{p} \in I_0$, replace it by the basis function indexed by $\vec{p} \pm \vec{e}_j$, where $\vec{e}_j$ denotes the j-th unit vector, $j = 1, \ldots, d$. Apply this to each basis function in I_0, which gives a total of $2d \times k$ substitution moves.

Illustration:

$$\vec{p} \to \begin{cases} (p_1 + 1, p_2, p_3, \ldots, p_d) \\ (p_1, p_2 + 1, p_3, \ldots, p_d) \\ \vdots \\ (p_1, p_2, p_3, \ldots, p_d + 1) \\ (p_1 - 1, p_2, p_3, \ldots, p_d) \\ (p_1, p_2 - 1, p_3, \ldots, p_d) \\ \vdots \\ (p_1, p_2, p_3, \ldots, p_d - 1) \end{cases}$$

for each $\vec{p} \in I_0$.

Addition moves. Given the current index set I_0, the addition moves are obtained by adding to I_0 the basis functions indexed by one of the unit vectors or by one of the basisi functions in $SUB(I_0)$. This gives a total of $3d$ addition moves.

Illustration:

$$\vec{p}_{k+1} = \begin{cases} (1, 0, \ldots, 0) \\ \vdots \\ (0, \ldots, 0, 1) \\ (p_1 + 1, p_2, p_3, \ldots, p_d) \\ \vdots \\ (p_1, p_2, p_3, \ldots, p_d + 1) \\ (p_1 - 1, p_2, p_3, \ldots, p_d) \\ \vdots \\ (p_1, p_2, p_3, \ldots, p_d - 1) \end{cases}$$

By replacing the jumps of size 1 in the definition of $SUB(I_0)$ and $ADD(I_0)$ by jumps of size in $\{1, \ldots, S\}$, this algorithm can be made increasingly agressive.

SIMPLE EXAMPLE FOR POLYNOMIAL BASIS

Let $d = 4$ and $Y = X_1X_2X_3 + X_2X_4^5 + \varepsilon$. Then $k = 2$, $\vec{p}_1 = (1, 1, 1, 0)$, $\vec{p}_2 = (0, 1, 0, 5)$.

A deletion move simply means removing one of the terms of the current model and fitting a model of size $k - 1$.

The substitution moves involve replacing the s^{th} term for $s = 1, \ldots, k$ with a new term, keeping the size of the model fixed at k.

The possible substitution moves are given by:

$$Y = \begin{cases} X_1^2X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_1 = (2,1,1,0) \\ X_1X_2^2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_1 = (1,2,1,0) \\ X_1X_2X_3^2 + X_2X_4^5 + \varepsilon & \vec{p}_1 = (1,1,2,0) \\ X_1X_2X_3X_4 + X_2X_4^5 + \varepsilon & \vec{p}_1 = (1,1,1,1) \\ X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_1 = (0,1,1,0) \\ X_1X_3 + X_2X_4^5 + \varepsilon & \vec{p}_1 = (1,0,1,0) \\ X_1X_2 + X_2X_4^5 + \varepsilon & \vec{p}_1 = (1,1,0,0) \\ X_1X_2X_4^5 + X_1X_2X_3 + \varepsilon & \vec{p}_2 = (1,1,0,5) \\ X_2^2X_4^5 + X_1X_2X_3 + \varepsilon & \vec{p}_2 = (0,2,0,5) \\ X_2X_3X_4^5 + X_1X_2X_3 + \varepsilon & \vec{p}_2 = (0,1,1,5) \\ X_2X_4^6 + X_1X_2X_3 + \varepsilon & \vec{p}_2 = (0,1,0,6) \\ X_4^5 + X_1X_2X_3 + \varepsilon & \vec{p}_2 = (0,0,0,5) \\ X_2X_4^4 + X_1X_2X_3 + \varepsilon & \vec{p}_2 = (0,1,0,4) \end{cases}$$

If none improve RSS, then find the best fit among the following *addition* moves:

$$Y = \begin{cases}
X_1 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (1, 0, 0, 0) \\
X_2 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (0, 1, 0, 0) \\
X_3 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (0, 0, 1, 0) \\
X_4 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (0, 0, 0, 1) \\
X_1^2X_2X_3 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (2, 1, 1, 0) \\
X_1X_2^2X_3 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (1, 2, 1, 0) \\
X_1X_2X_3^2 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (1, 1, 2, 0) \\
X_1X_2X_3X_4 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (1, 1, 1, 1) \\
X_2X_3 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (0, 1, 1, 0) \\
X_1X_3 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (1, 0, 1, 0) \\
X_1X_2 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (1, 1, 0, 0) \\
X_1X_2X_4^5 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (1, 1, 0, 5) \\
X_2^2X_4^5 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (0, 2, 0, 5) \\
X_2X_3X_4^5 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (0, 1, 1, 5) \\
X_2X_4^6 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (0, 1, 0, 6) \\
X_4^5 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (0, 0, 0, 5) \\
X_2X_4^4 + X_1X_2X_3 + X_2X_4^5 + \varepsilon & \vec{p}_3 = (0, 1, 0, 4)
\end{cases}$$

DOES THE DSA ALGORITHM DO THE JOB?

Is the algorithm capable to find the global minimum (i.e., the optimal predictor $W \to \psi_0(W) = E_0(Y \mid W)$) when n is large enough?

We generated $n = 1000$ observations from the following three true regression models with zero error, $d = 100$, $X_j \sim U(0,1)$, and we check if the D/S/A algorithm finds the truth.

$E_1[Y|X] = X_1 X_{12} X_{13}^2 X_{22} X_{24} X_{54} X_{79} X_{83} X_{95} + X_{15} X_{18} X_{37} X_{42} X_{68} + X_6 X_{22} X_{33}^3 X_{40} X_{58} X_{75} X_{82} X_{87} + X_{15} X_{31}$

$E_2[Y|X] =$
$X_7 X_{25} X_{31} X_{59} X_{63} X_{68} X_{70} X_{83} X_{88} X_{98} + X_0 X_{32} X_{47} X_{54} X_{66} X_{72} X_{73} X_{77} + X_{82} + X_7 X_{49} X_{55} X_{73} X_{80} + X_{33} X_{40} + X_{18} X_{21} X_{40} X_{56} X_{59} X_{71} X_{91} + X_9 X_{13} X_{18} X_{20} X_{41} X_{53} X_{69} X_{95} + X_3 X_{38} X_{78} X_{96} + X_0 X_{20} X_{64} X_{88} X_{91} X_{96} + X_2 X_6 X_{16} X_{37} X_{45} X_{46} X_{61} X_{68} X_{91} X_{95}$

$$\begin{gathered}E_3[Y|X] = X_0 X_1^2 X_4^4 X_{99}^{10} + X_{45} + \\ X_2^2 X_8 X_{14} X_{20} X_{22} X_{29} X_{36} X_{39} X_{41} X_{44} X_{48} X_{56} X_{62} X_{63} X_{65} X_{87} + \\ X_{27} X_{48} X_{63} X_{77} X_{78} X_{93} X_{94} + X_{71} + \\ X_{12} X_{18} X_{22} X_{44} X_{50} X_{55} X_{57} X_{64} X_{73}^2 X_{80} X_{83} X_{93} X_{94} X_{96} + X_{69} X_{91} + \\ X_2 X_4 X_{22} X_{23} X_{28} X_{36} X_{53} X_{79} X_{88} + X_{48} X_{70} X_{82} X_{97} + \\ X_3 X_{24} X_{29} X_{54} X_{64} X_{80}\end{gathered}$$

Simulation Results for Three Models

Zero error

$E[Y\|X]$	X	n	d	p	RSS
$E_1[Y\|X]$	$\mathcal{U}(0,1)$	1000	100	1.0	0.000000
$E_2[Y\|X]$	$\mathcal{U}(0,1)$	1000	100	1.0	0.000000
$E_3[Y\|X]$	$\mathcal{U}(0,1)$	1000	100	0.8	0.000001

SIMULATIONS

Consider the nonparametric polynomial regression (NPR) model for $E[Y|X]$, defined by the collection of sums of tensor-product polynomial basis functions:

$$Y = \sum_{s=1}^{size} \beta_s \prod_{j=1}^{d} X_j^{p_s(j)} + \varepsilon, \;\; E(\varepsilon|\vec{X}) = 0.$$

To assess the DSA algorithm's ability to minimize residual sum of squares over this NPR-model for large sample size, we randomly generated true regressions in this NPR-model and set $\epsilon = 0$, and verified if the algorithm found the truth.

The true regression model is randomly generated as follows:

$$\begin{aligned}
&size \sim \mathcal{U}\{1,\ldots,5\}\\
&\sum_{j=1}^{d} p_s(j) \sim \mathcal{U}\{1,\ldots,5\}\\
&\vec{p}_s \sim \text{Multinomial}(\textstyle\sum_{j=1}^{d} p_s(j), d, (\frac{1}{d},\ldots,\frac{1}{d}))
\end{aligned}$$

After randomly choosing $size$ and $\vec{p}_s$, each formed $\prod_j X_j^{p_s(j)}$ tensor-product was ensured to be unique. The sum of these randomly generated unique terms and ε yielded the true response variable Y.

REPORTED QUANTATIES

The following quantities are represented in the tables summarizing simulation results:

- p: proportion of correctly fitted terms given the true model
- $\bar{p}$: average proportion of correctly fitted terms across the number of repetitions
- RSS: residual sum of squares of fitted model
- $\overline{RSS}$: average RSS across the number of repetitions

Results for Randomly Generated Polynomial Regressions

Zero error

X	n	d	nsims	$\bar{p}$	$\overline{RSS}$
$\mathcal{U}(0,1)$	1000	5	1000	1.000	0.0000
$\mathcal{U}(0,1)$	1000	100	500	1.000	0.0000
Bernoulli(p)	1000	5	100	0.996	0.0000
Bernoulli(p)	2000	10	100	0.921	0.0000
Bernoulli(p)	1000	25	100	0.884	0.0000
Bernoulli(0.6)	500	5	100	1.000	0.0000
Bernoulli(0.6)	500	25	100	1.000	0.0000

Comparing $\varepsilon = 0$ to $\varepsilon \sim \mathcal{N}(0,1)$

The following two models were generated, first with $\varepsilon = 0$ and then with $\varepsilon \sim \mathcal{N}(0,1)$.

$E_3[Y|X] = X_0X_1^2X_2^2 + X_0X_1X_2^2X_3 + X_2^3 + X_4^4$

$E_4[Y|X] = X_7X_{25}X_{31}X_{59}X_{63}X_{68}X_{70}X_{83}X_{88}X_{98} +$
$X_0X_{32}X_{47}X_{54}X_{66}X_{72}X_{73}X_{77} + X_{82} + X_7X_{49}X_{55}X_{73}X_{80} +$
$X_{33}X_{40} + X_{18}X_{21}X_{40}X_{56}X_{59}X_{71}X_{91} +$
$X_9X_{13}X_{18}X_{20}X_{41}X_{53}X_{69}X_{95} + X_3X_{38}X_{78}X_{96} +$
$X_0X_{20}X_{64}X_{88}X_{91}X_{96} + X_2X_6X_{16}X_{37}X_{45}X_{46}X_{61}X_{68}X_{91}X_{95}$

The following quantities are used in the next table:

- RSS_n: $RSS/(n-k)$ represents the estimate of the variance of the error where k is the number of independent variables in fitted model
- RSS_0: the RSS of the true model
- *: indicates the model for which $\varepsilon \sim \mathcal{N}(0,1)$

Comparing $\varepsilon = 0$ to $\varepsilon \sim \mathcal{N}(0,1)^*$

$E[Y\|X]$	X	n	d	p	RSS_n	RSS_0
$E_3[Y\|X]$	$\mathcal{N}(5, 0.25)$	1000	5	1.0	0.0000	0.0000
$E_3[Y\|X]^*$	$\mathcal{N}(5, 0.25)$	10000	5	1.0	0.9886	0.9890
$E_4[Y\|X]$	$\mathcal{N}(5, 0.25)$	1000	100	1.0	0.0000	0.0000
$E_4[Y\|X]^*$	$\mathcal{N}(5, 0.25)$	10000	100	1.0	0.9955	0.9961

DSA ALGORITHM VERSUS stepAIC() R-FUNCTION

The DSA algorithm creates variables data-adaptively and therefore does not require enumeration of all potential variables.
The stepAIC() for linear regression does require enumeration of all variables. To compare the two black-box algorithms (data → predictor), we enumerated all main terms and two way interactions.

The following three true regression models were generated where $X_j \sim \mathcal{U}(1, 10)$, $j = 1, \ldots, d$, and $\varepsilon \sim \mathcal{N}(0, 1)$.

$$E_1[Y|X] = X_1 + X_2^2$$

$$E_2[Y|X] = X_1 X_3$$

$$E_3[Y|X] = X_1 X_3 + X_5^2 + X_7 X_{10}$$

The following quantities are represented in the table:

- $\hat{k}$: size of the final fitted model for each method
- RISK: estimate of the true risk, based on 20,000 independent observations, of the final model given by both methods

COMPARING STEP-AIC and CV-DSA

$E[Y\|X]$	n	d	$\hat{k}_{AIC}$	$\hat{k}_{CV}$	STEP-AIC	DSA-CV
$E_1[Y\|X]$	5000	3	2	2	0.9963	0.9963
$E_2[Y\|X]$	5000	10	19	1	0.9995	0.9932
$E_3[Y\|X]$	5000	10	22	3	1.0174	1.0106

PROPOSAL FOR PARTITION-MOVES

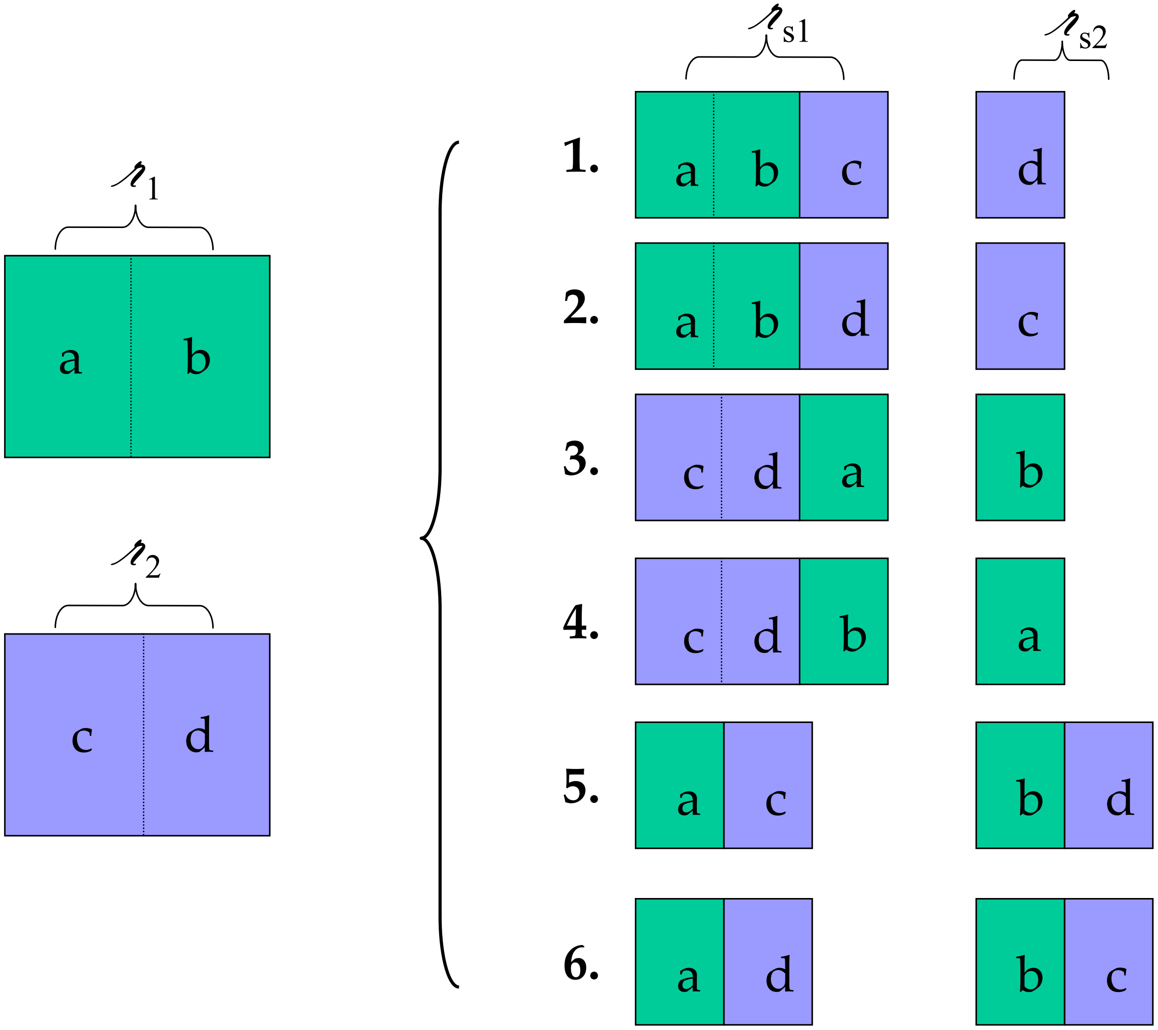
Given a partition $I_0 = \{R_1, \ldots, R_k\}$ of the covariate space $\mathcal{W}$, we propose the following moves.

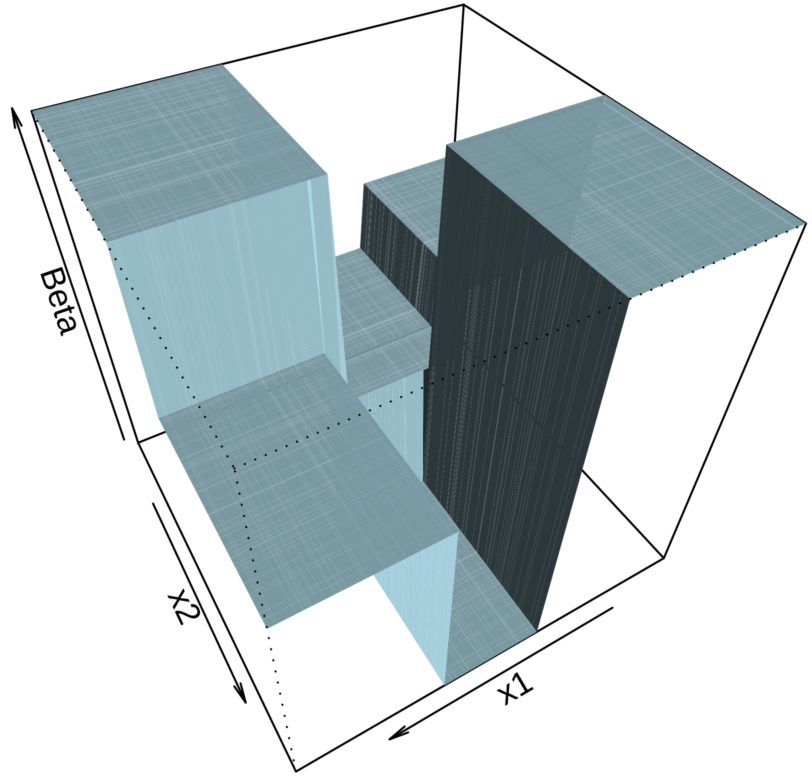
deletion moves: Replace two indicators I_{R_i}, I_{R_j} of sets R_i, R_j by the indicator $I_{R_i \cup R_j}$ of the union $R_i \cup R_j$.

substitution moves: See below.

addition moves: Replace an indicator I_{R_i} of a set R_i by two indicators of $R_i \cap \{W_l < c\}$ and $R_i \cap \{W_l > c\}$.

POSSIBLE SUBSTITUTIONS




Beta
x2
x1

INDICATOR OF SET REPRESENTATION

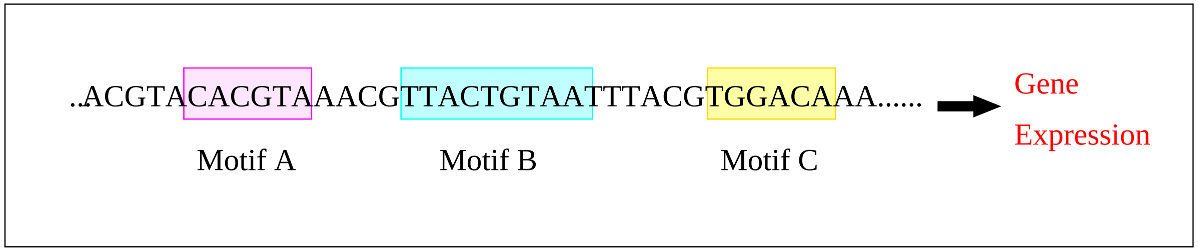
$$\left(\left\{ \begin{array}{c} (a_{111},b_{111}] \\ \textit{and} \\ (a_{211},b_{211}] \\ \textit{and} \\ \dots \\ \dots \\ \dots \\ \textit{and} \\ (a_{p11},b_{p11}] \end{array} \right\} \textit{or} \left\{ \begin{array}{c} (a_{121},b_{121}] \\ \textit{and} \\ (a_{221},b_{221}] \\ \textit{and} \\ \dots \\ \dots \\ \dots \\ \textit{and} \\ (a_{p21},b_{p21}] \end{array} \right\} \textit{or} \dots \textit{or} \left\{ \begin{array}{c} (a_{11m},b_{11m}] \\ \textit{and} \\ (a_{21m},b_{21m}] \\ \textit{and} \\ \dots \\ \dots \\ \dots \\ \textit{and} \\ (a_{p1m},b_{p1m}] \end{array} \right\} \right)$$

DSA vs. RECURSIVE PARTITIONING

Table 10: 100 repetitions of full data simulated from $y = x^2 + er$, where $x \sim N(0,1)$ and $er \sim N(0,.25)$. Conditional risk of our method (*ours*), rpart (R-implementation of CART) with 1-SE (*rpart*) and rpart by minimizing CV-error (*rpart0*).

n	Method	Mean	Std.Dev.	Avg. size	Ratio
250	ours	0.26125	0.09384	7.44	1
	rpart	0.45305	0.14195	5.69	.577
	rpart0	0.35172	0.09927	14.45	.743
500	ours	0.18935	0.07318	9.95	1
	rpart	0.27216	0.08574	9.55	.696
	rpart0	0.22187	0.07544	21.26	.853
1000	ours	0.14080	0.04016	12.06	1
	rpart	0.18489	0.05206	13.02	.762
	rpart0	0.15403	0.04916	28.44	.914

PREDICTING GENE EXPRESSION FROM SEQUENCE



..ACGTACACGTAAACGTTACTGTAATTTACGTGGACAAA......

Motif A

Motif B

Motif C

Gene Expression

Goal: To identify binding sites (regulatory motifs).

n

Data:

- Gene Expression Data: $P \times N$ matrix with entries $Y_{ij}, i = 1, \cdots, P, j = 1, \cdots, N$. Y_{ij} is the logarithm of the relative gene expression for gene i in experiment j.
- Upstream Control Region (UCR): Roughly 600 to 1000 base pairs of the gene start site.

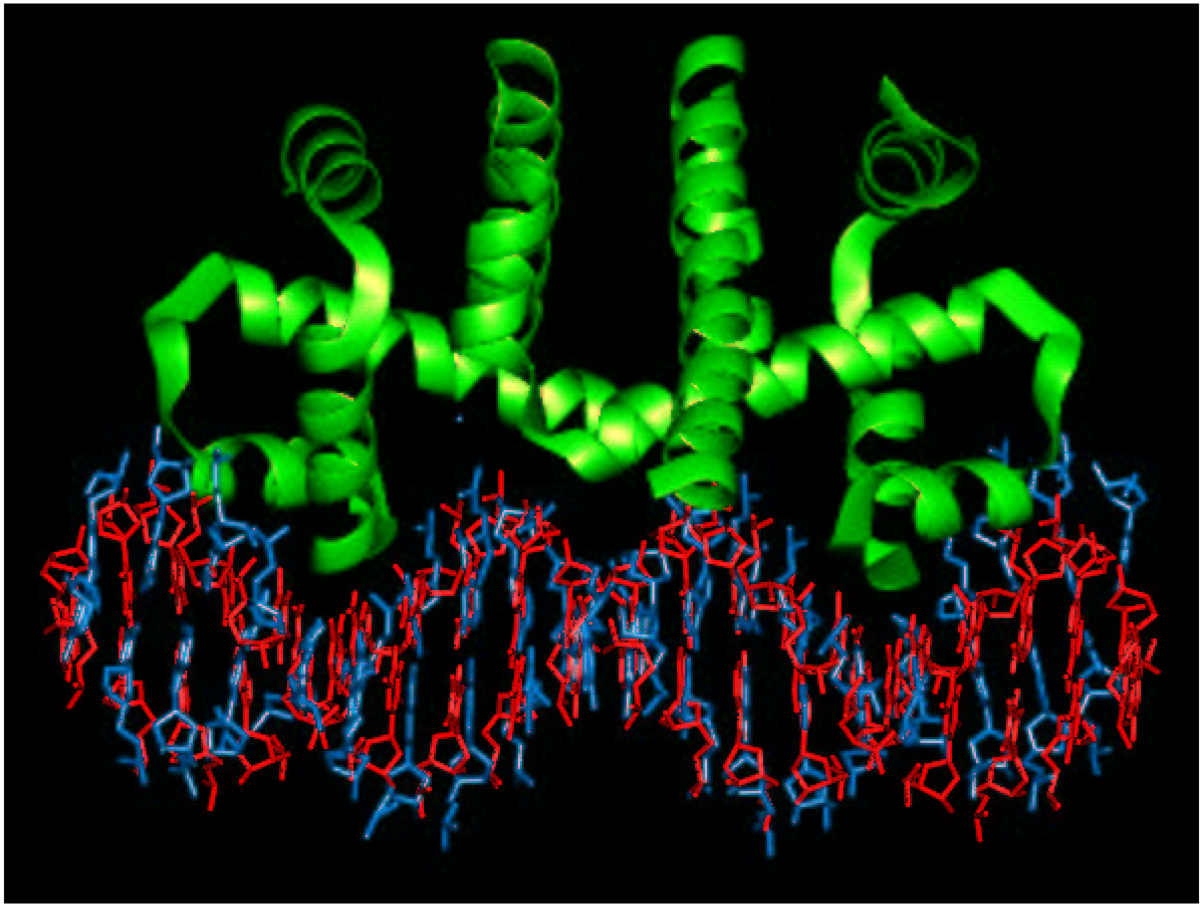
WHAT ARE BINDING SITES AND WHY ARE THEY IMPORTANT?

DNA binding proteins (transcription factors) bind to DNA in a sequence specific manner. These short DNA sequences (5-25 base pairs) are called binding sites or regulatory motifs.

All cells from bacteria to mammals respond to various treatments by activating or repressing the expression of particular genes.

Gene expression is regulated by transcription factors binding selectively to their specific binding sites.

GAL4 BINDING



From http://www.cryst.bbk.ac.uk/PPS2/.

CELL CYLE IN YEAST

We used the DSA algorithm with polynomial basis to regress the $512 = 1024/2$ indicators of "Presence of length 5 motiff" on gene expression at each time-point in the 16 time point cell cycle experiment in yeast (Cho et al., 1998).

In the DSA algorithm we use 2-fold cross-validation, maximal size of model $K = 5$, and Subset Estimator 1 for the subset I of basis functions.

T=30 min				
$\hat{k}_{CV}$	Run time	$\hat{\theta}_{full}$	$\hat{\theta}_{main}$	$\hat{\theta}_{full}/\hat{\theta}_{main}$
3	6.2 hrs	1176.144	1176.746	0.999

Selected pentamers	$\sum_{i=1}^{n} X_i, \quad X_i = \{0, 1\}$
● ACGCG [MCB]	774
● 10-way interaction:	60
AAATC	2116
AACTA	2023
AATAT	2492
ACAAA	2410
ACGCG [MCB]	774
AGCCG	937
ATGAA	2207
CAAGA	2051
CCACC	960
GAAAC	1967
● 10-way interaction:	26
AACTT	2150
ACGCG [MCB]	774
AGATA	2127
AGCAA	2009
ATAAC	2027
ATATG	2122
CTGCC	1078
CTGTC	1188
GGCCC	615
TGACA	1603

T=50 min

$\hat{k}_{CV}$	Run time	$\hat{\theta}_{full}$	$\hat{\theta}_{main}$	$\hat{\theta}_{full}/\hat{\theta}_{main}$
1	10.9 hrs	1149.242	1138.828*	1.009
Selected pentamers		$\sum_{i=1}^{n} X_i, \quad X_i = \{0, 1\}$		
● 15-way interaction:		62		
AAGAG		2120		
AAGCA		1994		
AAGGA		2118		
AATCA		2038		
ACAAA		2410		
AGGAA *		2146		
AGGCC		798		
AGTGG		1278		
ATTCA		2022		
ATTTA		2398		
CAACA		1852		
GATTA		1811		
GCTTA		1522		
TACTA		2002		
TGGAA		1915		

T=70 min

$\hat{k}_{CV}$	Run time	$\hat{\theta}_{full}$	$\hat{\theta}_{main}$	$\hat{\theta}_{full}/\hat{\theta}_{main}$
4	8.4 hrs	1296.356	1295.034*	1.001

Selected pentamers	$\sum_{i=1}^{n} X_i, \quad X_i = \{0, 1\}$
● 2-way interaction:	2590
AAAAG	2676
GAAAA [ECB]	2676
● 4-way interaction:	676
AAAAT	2679
AAACA *[STE 12]	2406
AAATA	2649
ACGCG [MCB]	774
● 11-way interaction:	103
AAAAG	2676
AAACA [STE 12]	2406
AATTG	2059
ACATG	1300
ATAAA	2609
ATACG	1480
ATATA	2434
CGCGA	743
GAATA	2213
GTTCA	1545
TCAAA	2326

T=70 min

$\hat{k}_{CV}$	Run time	$\hat{\theta}_{full}$	$\hat{\theta}_{main}$	$\hat{\theta}_{full}/\hat{\theta}_{main}$
4	8.4 hrs	1296.356	1295.034*	1.001

Selected pentamers	$\sum_{i=1}^{n} X_i, \quad X_i = \{0, 1\}$
● 15-way interaction:	49
AAAAT	2679
ACGCG [MCB]	774
AAACA [STE 12]	2406
AAATA	2649
AACAC	1670
AACAG	1895
AATCC	1370
AGGTG	1290
CAAAA	2538
CTTCA	1912
GATAA	2069
GGAAA	2312
GGGAA	1631
TATCA	2120
TGTAA	2102

T=110 min

$\hat{k}_{CV}$	Run time	$\hat{\theta}_{full}$	$\hat{\theta}_{main}$	$\hat{\theta}_{full}/\hat{\theta}_{main}$
1	11.8 hrs	1245.625	1232.567*	1.011

Selected pentamers	$\sum_{i=1}^{n} X_i, \quad X_i = \{0, 1\}$
● 9-way interaction:	178
AAAAT	2679
AAACA [STE 12]	2406
AAAGT	2349
AATAG	2247
ACGCG *[MCB]	774
AGAAG	2150
ATAAG	2095
GACGC	855
TCTCA	1695

PREDICTION OF SURVIVAL with CV-DSA

Let T be a log-survival time, and suppose that our goal is to estimate the optimal predictor $\psi_0(W) = E_0(T \mid W)$. However, due to right-censoring by a variable C, we only observe $O_i = (\tilde{T}_i \equiv \min(T_i, C_i), \Delta_i = I(T_i \leq C_i), W_i)$. Let $G(\cdot \mid T, W)$ be the conditional distribution of censoring C, given (T, W), and we assume that censoring is independent of survival time, given W: i.e., $G(\cdot \mid T, W) = G(\cdot \mid W)$.

The CV-DSA algorithm above for estimating the optimal predictor $\psi_0(W)$ based on the full (uncensored) data (T_i, W_i), $i = 1, \ldots, n$, is 100% driven by the squared error loss function $L(T, W, \psi)$. We can replace in the CV-DSA the squared error loss function $L(T, W, \psi)$ by a function of the observed data O with the same expectation.

The **Inverse Probability of Censoring Weighted** (IPCW) Squared Error Loss Function

$$L(O,\psi \mid G) \equiv L(T,W,\psi)\frac{\Delta}{P_G(\Delta=1 \mid W)} = (T-\psi(W))^2\frac{\Delta}{\bar{G}(T \mid W)}.$$

For the optimal (that is, minimal variance, and maximally robust) **double robust IPCW loss function**, we refer to van der Laan, Robins (2002).

Remark Given an estimator $G(P_n)$ (e.g., Kaplan-Meier, or Cox-proportional hazards) of the censoring distribution G, the cross-validation selector is now given by:

$$\hat{k} = \text{argmin}_k \sum_{i=1}^{n} I(S_n(i)=1)(T_i - \psi_k(W_i \mid P^0_{n,S_n}))^2\frac{\Delta_i}{\bar{G}(P^0_{n,S_n})(T_i \mid W_i)}.$$

CONCLUDING REMARKS

- Cross-validated DSA algorithms provide black-box algorithms for estimating parameters which minimize the expectation of a given loss function (e.g., regression, conditional density, conditional survival function).
- Simulations show that the DSA-algorithm is asymptotically surprisingly capable of truly minimizing the cross-validated/empirical risk function over all subsets of basis functions.
- In complex (i.e., genomic) studies we should let cross-validation make the choices: e.g, if we choose the parametrization/basis with cross-validation, then the estimator becomes adaptive to the truth.
- Any such algorithm is immediately generalizable to censored data by replacing the full-data loss function by the (double

robust) IPCW loss function.

SIMULATIONS

Consider the nonparametric polynomial regression (NPR) model for $E[Y|X]$, defined by the collection of sums of tensor-product polynomial basis functions:

$$Y = \sum_{s=1}^{size} \beta_s \prod_{j=1}^{d} X_j^{p_s(j)} + \varepsilon, \;\; E(\varepsilon|\vec{X}) = 0.$$

To assess the DSA algorithm's ability to minimize residual sum of squares over this NPR-model, we randomly generated true regressions in this NPR-model and set $\epsilon = 0$, and verified if the algorithm found the truth.

The true regression model is randomly generated as follows:

$$\begin{aligned}
&size \sim \mathcal{U}\{1,\ldots,5\}\\
&\sum_{j=1}^{d} p_s(j) \sim \mathcal{U}\{1,\ldots,5\}\\
&\vec{p}_s \sim \text{Multinomial}(\textstyle\sum_{j=1}^{d} p_s(j), d, (\frac{1}{d},\ldots,\frac{1}{d}))
\end{aligned}$$

After randomly choosing $size$ and $\vec{p}_s$, each formed $\prod_j X_j^{p_s(j)}$ tensor-product was ensured to be unique. The sum of these randomly generated unique terms and ε yielded the true response variable Y.

REPORTED QUANTATIES

The following quantities are represented in the tables summarizing simulation results:

- p: proportion of correctly fitted terms given the true model
- $\bar{p}$: average proportion of correctly fitted terms across the number of repetitions
- RSS: residual sum of squares of fitted model
- $\overline{RSS}$: average RSS across the number of repetitions

Simulation Results for Randomly Generated Polynomials

Zero error

X	n	d	nsims	$\bar{p}$	$\overline{RSS}$
$\mathcal{U}(0,1)$	1000	5	1000	1.000	0.0000
$\mathcal{U}(0,1)$	1000	100	500	1.000	0.0000
Bernoulli(p)	1000	5	100	0.996	0.0000
Bernoulli(p)	2000	10	100	0.921	0.0000
Bernoulli(p)	1000	25	100	0.884	0.0000
Bernoulli(0.6)	500	5	100	1.000	0.0000
Bernoulli(0.6)	500	25	100	1.000	0.0000

Simulations: Increase complexity of true regression

In the previous simulations, $size \sim \mathcal{U}\{1, \ldots, 5\}$, but now we increase both the size and the allowed sum of the powers of the polynomials within a tensor product as follows: $size \sim \mathcal{U}\{1, \ldots, 10\}$, $\sum_{j=1}^{d} p_s(j) \sim \mathcal{U}\{1, \ldots, 20\}$, with $d = 100$. In these simulations, $RSS \leq 0.000001$ was used as a stopping criterion.

The following three true regression models were generated:

$$E_1[Y|X] = X_1 X_{12} X_{13}^2 X_{22} X_{24} X_{54} X_{79} X_{83} X_{95} + X_{15} X_{18} X_{37} X_{42} X_{68} + X_6 X_{22} X_{33}^3 X_{40} X_{58} X_{75} X_{82} X_{87} + X_{15} X_{31}$$

$$E_2[Y|X] = X_7 X_{25} X_{31} X_{59} X_{63} X_{68} X_{70} X_{83} X_{88} X_{98} + X_0 X_{32} X_{47} X_{54} X_{66} X_{72} X_{73} X_{77} + X_{82} + X_7 X_{49} X_{55} X_{73} X_{80} + X_{33} X_{40} + X_{18} X_{21} X_{40} X_{56} X_{59} X_{71} X_{91} + X_9 X_{13} X_{18} X_{20} X_{41} X_{53} X_{69} X_{95} + X_3 X_{38} X_{78} X_{96} + X_0 X_{20} X_{64} X_{88} X_{91} X_{96} + X_2 X_6 X_{16} X_{37} X_{45} X_{46} X_{61} X_{68} X_{91} X_{95}$$

Simulation Results for Two Models

Zero error

$E[Y\|X]$	X	n	d	p	RSS
$E_1[Y\|X]$	$\mathcal{U}(0,1)$	1000	100	1.0	0.000000
$E_2[Y\|X]$	$\mathcal{U}(0,1)$	1000	100	1.0	0.000000

Comparing $\varepsilon = 0$ to $\varepsilon \sim \mathcal{N}(0,1)$

The following two models were generated, first with $\varepsilon = 0$ and then with $\varepsilon \sim \mathcal{N}(0,1)$.

$$E_3[Y|X] = X_0 X_1^2 X_2^2 + X_0 X_1 X_2^2 X_3 + X_2^3 + X_4^4$$

$$E_4[Y|X] = X_7 X_{25} X_{31} X_{59} X_{63} X_{68} X_{70} X_{83} X_{88} X_{98} + X_0 X_{32} X_{47} X_{54} X_{66} X_{72} X_{73} X_{77} + X_{82} + X_7 X_{49} X_{55} X_{73} X_{80} + X_{33} X_{40} + X_{18} X_{21} X_{40} X_{56} X_{59} X_{71} X_{91} + X_9 X_{13} X_{18} X_{20} X_{41} X_{53} X_{69} X_{95} + X_3 X_{38} X_{78} X_{96} + X_0 X_{20} X_{64} X_{88} X_{91} X_{96} + X_2 X_6 X_{16} X_{37} X_{45} X_{46} X_{61} X_{68} X_{91} X_{95}$$

The following quantities are used in the next table:

- RSS_n: $RSS/(n-k)$ represents the estimate of the variance of the error where k is the number of independent variables in fitted model
- RSS_0: the RSS of the true model
- *: indicates the model for which $\varepsilon \sim \mathcal{N}(0,1)$

Comparing $\varepsilon = 0$ to $\varepsilon \sim \mathcal{N}(0, 1)$*

$E[Y\|X]$	X	n	d	p	RSS_n	RSS_0
$E_3[Y\|X]$	$\mathcal{N}(5, 0.25)$	1000	5	1.0	0.0000	0.0000
$E_3[Y\|X]^*$	$\mathcal{N}(5, 0.25)$	10000	5	1.0	0.9886	0.9890
$E_4[Y\|X]$	$\mathcal{N}(5, 0.25)$	1000	100	1.0	0.0000	0.0000
$E_4[Y\|X]^*$	$\mathcal{N}(5, 0.25)$	10000	100	1.0	0.9955	0.9961

DSA ALGORITHM VERSUS stepAIC() R-FUNCTION

The DSA algorithm creates variables and therefore does not require enumeration of all potential variables.
The stepAIC() for linear regression does require enumeration of all variables. To compare the two black-box algorithms (data → model fit), we enumerated all main terms and two way interactions.

The following three true regression models were generated where $X_j \sim \mathcal{U}(1, 10)$, $j = 1, \ldots, d$, and $\varepsilon \sim \mathcal{N}(0, 1)$.

$$E_1[Y|X] = X_1 + X_2^2$$

$$E_2[Y|X] = X_1 X_3$$

$$E_3[Y|X] = X_1 X_3 + X_5^2 + X_7 X_{10}$$

The following quantities are represented in the table:

- $\hat{k}$: size of the final fitted model for each method
- $\hat{\theta}_{opt}$: estimate of the true risk, based on 20,000 independent observations, of the final model given by both methods

Comparing `stepAIC` to Cross-Validated Del/Sub/Add

$E[Y\|X]$	n	d	$\hat{k}_R$	$\hat{k}_{CV}$	$\hat{\theta}_{opt,R}$	$\hat{\theta}_{opt,CV}$
$E_1[Y\|X]$	5000	3	2	2	0.9963	0.9973
$E_2[Y\|X]$	5000	10	19	1	0.9995	0.9932
$E_3[Y\|X]$	5000	10	22	3	1.0174	1.0106

CROSS-VALIDATED ϵ-NET ESTIMATOR

Mark van der Laan

Joint work with Sandrine Dudoit, Peter Dimitrov.

Division of Biostatistics, University of California, Berkeley.

September 6, 2003

Department of Statistics, Neyman Seminar

©Copyright 2003, all rights reserved

SELECTION IN REGRESSION

Let $O_1 = (Y_1, W_1), \ldots, O_n = (Y_n, W_n)$ be n i.i.d. observations of $O = (Y, W) \sim P_0$, where Y denotes an outcome of interest and W is a d-dimensional vector of covariates. Let $\mathcal{M}$ be a model for P_0. Let $\psi_0(w) = E_{P_0}(Y \mid W)$ be the parameter (function) of interest, and let $\mathbf{\Psi} = \{E_P(Y \mid W) : P \in \mathcal{M}\}$ be the parameter space. Let $L(O, \psi)$ be the squared error loss function for a candidate ψ whose expectation is minimized by ψ_0:

$$\psi_0 \quad = \quad \text{argmin}_{\psi \in \mathbf{\Psi}} E_0 L(O, \psi \mid \eta_0).$$

Let P_n be the empirical distribution of $O_1, \ldots, O_n$. Let $\hat{\psi}_k(\cdot) = \psi_k(\cdot \mid P_n) \in \mathbf{\Psi}$, $k = 1, \ldots, K(n)$, be a collection of estimators (i.e., algorithms one can apply to data) of $\psi_0(\cdot)$.

The Selection Problem: Choose a data adaptive $\hat{k} = \hat{k}(P_n)$ so that

$$\begin{aligned} d_n(\hat{\psi}_{\hat{k}}, \psi_0) &\equiv \int \left\{ L(O, \psi_{\hat{k}}(\cdot \mid P_n)) - L(O, \psi_0) \right\} dP_0(O) \\ &= \int (\psi_{\hat{k}}(W \mid P_n) - \psi_0(W))^2 dP_0(W) \\ &\rightarrow 0, \text{ at asymptotically optimal speed.} \end{aligned}$$

THE OPTIMAL BENCHMARK SELECTOR

Let

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

This optimal benchmark selector (for each given data set) depends on the unknown data generating distribution P_0.

Asymptotic equivalence with benchmark selector: Given the $K(n)$ candidate estimators, a selector $\hat{k} = \hat{k}(P_n)$ is asymptotically equivalent with the optimal benchmark if

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

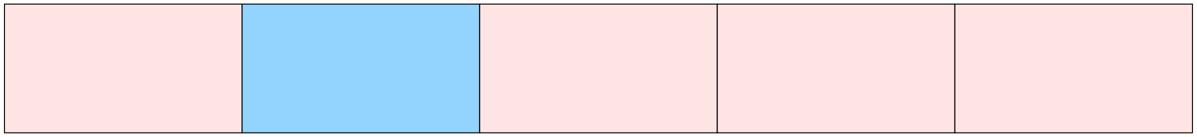
In particular, then it is asymptotically optimal.

THE CROSS-VALIDATION SELECTOR

Define random vector $S_n \in \{0,1\}^n$ for splitting the sample into a validation and a training sample.

$$S_{n,i} = \begin{cases} 0 & \text{if} \quad \text{i-th observation is in the training sample} \\ 1 & \text{if} \quad \text{i-th observation is in the validation sample} \end{cases}$$

Different choices of S_n cover all types of cross-validation including $V-$ fold cross-validation, monte carlo cross validation (bootstrap cross-validation): e.g. 5-fold cross-validation: S_n has 5 realizations.



Let $p = n_1/n$ be the proportion constituting the validation sample.

Let P^0_{n,S_n}, P^1_{n,S_n} be the empirical distributions of the training and validation sample, respectively.

The selector is defined by:

$$\begin{aligned} \hat{k} &\equiv \mathrm{argmin}_k E_{S_n} \int L(o, \psi_k(\cdot \mid P^0_{n,S_n})) dP^1_{n,S_n}(o) \\ &= \mathrm{argmin}_k E_{S_n} \sum_{i:S_n(i)=1} (Y_i - \psi_k(W_i \mid P^0_{n,S_n}))^2. \end{aligned}$$

FINITE SAMPLE RESULT

Define the distance function for estimators based on training samples of size $n(1-p)$:

$$\begin{aligned} d_{n(1-p)}(\hat{\psi}_k, \psi_0) &= E_{S_n} \int \left\{ L(o, \psi_k(\cdot \mid P^0_{n,S_n})) - L(o, \psi_0) \right\} dP_0(o) \\ &= E_{S_n} \int \left(\psi_k(W \mid P^0_{n,S_n}) - \psi_0(W) \right)^2 dP_0(W). \end{aligned}$$

$\hat{k}$ aims to minimize $k \rightarrow d_{n(1-p)}(\hat{\psi}_k, \psi_0)$. Denote the minimizer, i.e. the optimal comparable benchmark selector for $n(1-p)$ observations, with:

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

Suppose that the loss function $L(O, \psi)$ is uniformly bounded by a universal M_1, and

$$\mathrm{VAR}_0 \left\{ L(O, \psi) - L(O, \psi_0) \right\} \leq M_2 E_0 \left\{ L(O, \psi) - L(O, \psi_0) \right\}.$$

For any $\delta > 0$, we have for a specified constant
$C(M_1, M_2, \delta) = 2(1+\delta)^2 (M_1/3 + M_2/\delta)$

$$\begin{aligned} Ed_{n(1-p)}(\hat{\psi}(\hat{k}), \psi_0) \quad &\leq \quad (1+\delta) Ed_{n(1-p)}(\hat{\psi}(\tilde{k}_{n(1-p)}), \psi_0) \\ &\quad + \frac{C(M_1, M_2, \delta) \log K(n)}{np}. \end{aligned}$$

COROLLARY: ASYMPTOTIC OPTIMALITY

If $p = p(n) \to 0$ slowly enough with sample size, so that

$$\frac{\log(K(n))}{np(n)} \Big/ E d_n(\hat{\psi}(\tilde{k}_n), \psi_0) \to 0,$$

then

$$\frac{E d_n(\hat{\psi}(\hat{k}), \psi_0)}{E d_n(\hat{\psi}(\tilde{k}_n), \psi_0)} \to 1.$$

That is, the data adaptive selector $\hat{k}$ is asymptotically equivalent (and thus optimal) with the optimal benchmark selector.

THE ADAPTIVE ϵ-NET ESTIMATOR

SUB-PARAMETER SPACES Let $\boldsymbol{\Psi}_s \subset \boldsymbol{\Psi}$ be sub-parameter spaces indexed by $s \in \{1, \ldots, K_1(n)\}$. Let $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}$.

CONSTRUCT ϵ-NETS: For each subspace $\boldsymbol{\Psi}_s$, for a given $\epsilon > 0$, let

$$\{\psi_j^{\epsilon,s}, j = 1, \ldots, N_s(\epsilon)\} \subset \boldsymbol{\Psi}_s$$

be an ϵ-net of $\boldsymbol{\Psi}_s$. Here $N_s(\epsilon)$ can be chosen equal to the covering number of $(\boldsymbol{\Psi}_s, \| \cdot \|_{\boldsymbol{\Psi}})$.

MINIMIZE EMPIRICAL RISKS Let

$$\psi_{\epsilon,s}(\cdot \mid P_n) \equiv \text{argmin}_{\{\psi_j^{\epsilon,s}:j\}} \sum_{i=1}^{n} (Y_i - \psi_j^{\epsilon,s}(W_i))^2.$$

SELECT ϵ, s: Let $(\hat{\epsilon}, \hat{s})$ be the (ϵ, s) minimizing cross-validated empirical risk over a set of $K(n)$-values:

$$(\hat{\epsilon}, \hat{s}) \equiv \text{argmin}_{\epsilon,s} E_{S_n} \int L(Y, \psi_{\epsilon,s}(W \mid P^0_{n,S_n})) dP^1_{n,S_n}(Y, W).$$

The adaptive ϵ-net estimator is given by:

$$\psi(\cdot \mid P_n) = \psi_{\hat{\epsilon},\hat{s}}(\cdot \mid P_n).$$

FINITE SAMPLE RESULT FOR ϵ-NET ESTIMATOR

Let

$$\begin{aligned} B_0(\epsilon, s) &= \min_{j\in\{1,\ldots,N_s(\epsilon)\}} \int L(O, \psi_j^{\epsilon,s}) - L(O, \psi_0) dP_0(O) \\ &= \min_j \int \left(\psi_j^{\epsilon}(W) - \psi_0(W)\right)^2 dP_0(W). \end{aligned}$$

We have for any $\delta > 0$

$$E d_{n(1-p)}(\psi(\cdot \mid P_n), \psi_0) \leq$$

$$(1+2\delta)\min_{\epsilon,s}\left\{(1+2\delta)B_0(\epsilon,s) + 2C(M_1,M_2,\delta)\frac{1+\log(N_s(\epsilon))}{n(1-p)}\right\}$$

$$+2C(M_1,M_2,\delta)\frac{1+\log(K(n))}{np}.$$

Adaptivity: This finite sample inequality in terms of approximation errors of the ϵ-nets and the covering numbers $N_s(\epsilon)$ implies that the estimator is adaptive, that is, it achieves the optimal rate of convergence for the smallest subspace still containing the true ψ_0.

LARS/LASSSO VERSUS Epsilon-NET ESTIMATOR: SIMULATION

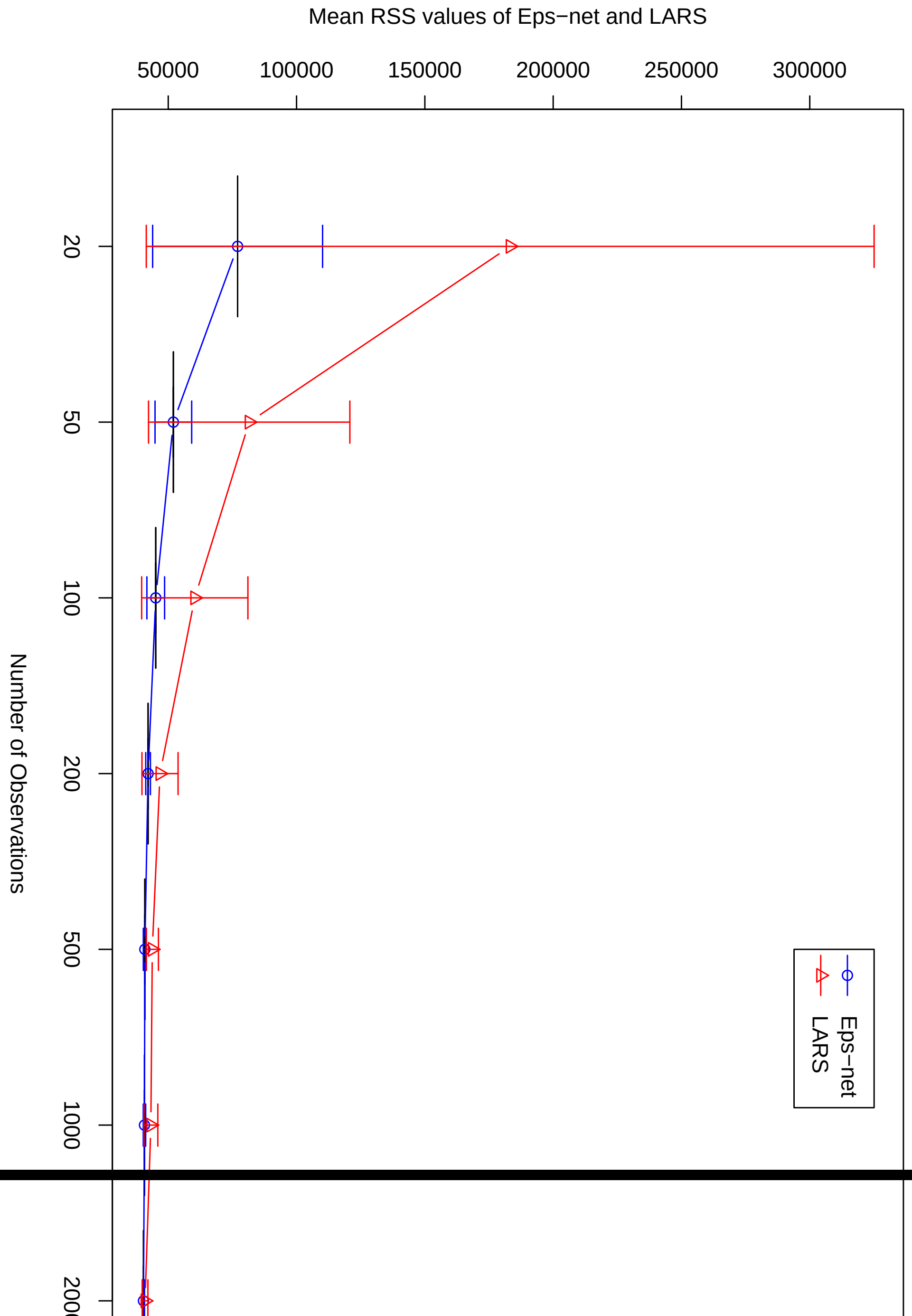
We simulate data sets from a linear regression $Y \sim \beta X + N(0, \sigma^2)$ with $X(j) \sim U(0,1)$, $j = 1, .., 10$, $\sigma^2 = 2$, and uniformly distributed regression coefficients β. We generated 2 simulated data sets of various sample sizes, and compared the ϵ-net linear regression estimator of β with the Least-Angle-Linear Regression estimator, (lars) based on residual sum of squares on an independent sample of 10,000 observations. “Lars” which has similar performance as the $L1$-penalized regression estimator (Lasso).

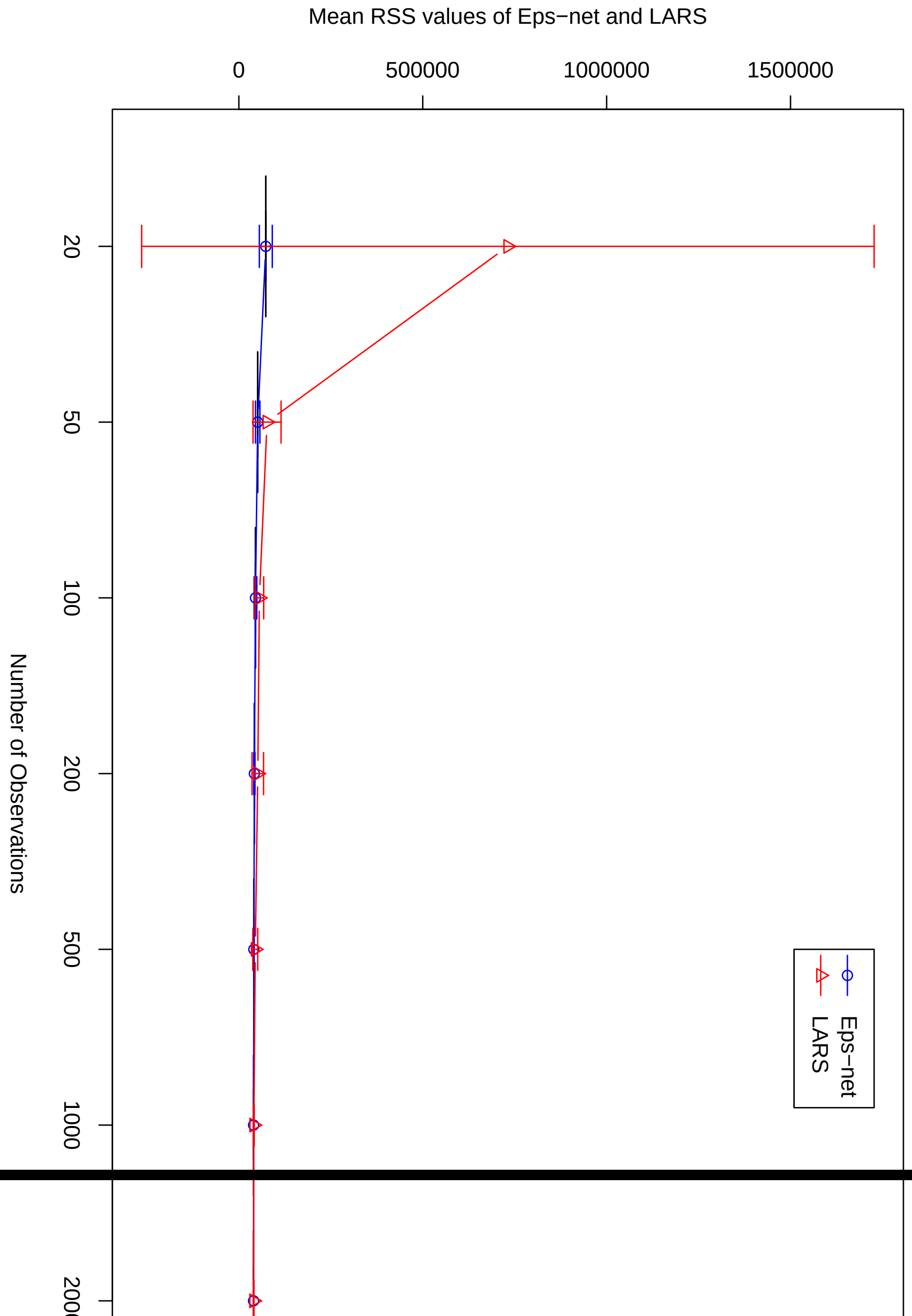
Sample size	eps-net	sd.eps-net	lars	sd.lars
20	77035.2	33114	183255.2	141810.6
50	52000	7143.2	81550.9	39228.6
100	45133.5	3447.5	60348.8	20702
200	42145.3	912.3	46798.5	7036.4
500	40886.5	586.3	43790.7	2394.7
1000	40738.6	471.8	43233.9	2702.7
2000	40348.2	400.6	40977.5	1095.3

Table 11: Sigma= 2

Sample size	eps-net	sd.eps-net	lars	sd.lars
20	73228	17601.2	731493.2	995771.5
50	51217.6	5992.5	76624.4	38065.6
100	45166.5	3690.1	55813.6	11746.9
200	42189.6	1754.1	51473.8	15664.4
500	40719.4	640.4	44636.3	6696.1
1000	40090.4	691	40724	943.8
2000	39987.2	573.7	40409.8	706.6

Table 12: Sigma= 2


Mean RSS values of Eps−net and LARS
50000
100000
150000
200000
250000
300000
Number of Observations
20
50
100
200
500
1000
200
Eps−net
LARS


Mean RSS values of Eps−net and LARS
0
500000
1000000
1500000
Number of Observations
20
50
100
200
500
1000
200
Eps−net
LARS

CLUSTERING ALGORITHMS and a STATISTICAL FRAMEWORK

Mark van der Laan

Division of Biostatistics, UC Berkeley

www.stat.berkeley.edu/~laan

www.bepress.com/ucbbiostat/

©Copyright 2003, all rights reserved

CLUSTERING

Consider a collection of n p-dimensional vectors. This can be represented as a $p \times n$-matrix.

As statisticians, we like to think as these n vectors as a random sample consisting of n independently and identically distributed observations of a random vector. For example, this random vector might represent the gene expression profile of a randomly drawn person from a population of cancer patients.

Clustering columns: For each pair of p-dimensional vectors compute a dissimilarity. Let D be the $n \times n$-distance/dissimilarity matrix.

Clustering rows: Construct a $p \times p$ dissimilarity matrix.

A (model free) clustering algorithm maps a distance matrix and a user supplied K into a n-dimensional (or p-dimensional) vector of cluster labels ranging in $\{1, \ldots, k\}$.

A clustering algorithm is defined by maximizing a performance criterian measuring the performance for a given clustering result, where the maximization is over an allowed set of possible cluster results.

Keep in mind, given K:

Different criterian $\Longrightarrow$ Different Clusters.
Different allowed set $\Longrightarrow$ Different Clusters.
Differen dissimilarity $\Longrightarrow$ Different Clusters.
Different criterian/dissimilarity/allowed set $\Longrightarrow$ Different VARIABILITY (across sample fluctuations) of clusters.

What dissimilarity matrix to use?
What clustering algorithm (defined by allowed set and criterian) to use?

Approach: 1) Understand the dissimilarity choice, 2) Understand the criterian and allowed set, 3) Understand variability and 4) Interpret results. Repeat 1–4 for different choices of dissimilarities and clustering algorithms.

BOOTSTRAP

What does variability of clusters mean?

Answer: In general, variance/variability of the sampling distribution of the clustering algorithm (this is a random vector or matrix) is a measure of spread of this sampling distribution around the true wished clustering result (one would have seen if the sample size is infinitely large).

Consequently, variance of clusters is calculated from a large sample of clustering results where each clustering result is obtained by resampling n vectors, and applying the clustering algorithm.

Measuring variability: There are a large number of ways of measuring the variance of these resampled clustering results depending on how one measures distance between a sampled clustering result and the aimed clustering result. Some specific proposals, such as cluster specific sensitivity, cluster specific positive predictive value, gene specific membership probabilities (with corresponding cluster-probabillity plot), are provided in van der Laan, Bryan (2001) (www.stat.berkeley.edu/ laan).

How to estimate variability? Resample from an estimate of the true data generating distribution. For example, resample n vectors from the empirical distribution of the n vectors which puts probability $1/n$ on each observation.

This statistical procedure, that is, *resampling with the purpose of estimating the variance of a data analytic result*, is called Bootstrap.

CLUSTERING OF MICROARRAY DATA

Clustering has important applications in the analysis of gene expression data. Consider a sample of n patients and suppose we collect a p-dimensional gene expression profile on each patient. Important results can be obtained by:

- Clustering of the p genes (n dimensional vectors).
- Clustering of the n patients (p dimensional vectors).
- Clustering genes, and within each cluster of genes, cluster patients.
- Clustering genes, reduce each patients' gene expression profile to the vector of cluster-specific medoids/centers. This vector of medoids can be used as a fingerprint and as a set of predictors of an outcome of interest (e.g. survival).

Give example of 3 cancer groups of patients. Clustering genes. What distance would show what clusters of genes? Different algorithms can still show different results. e.g. hierarchical with binary splits! is an example of a constraint allowed set. Also show clustering patients within clusters of genes, we have transparencies on that.

DISSIMILARITIES

Possible dissimilarities between a pair of vectors are:

- EUCLIDEAN DISTANCE
- 1 MINUS CORRELATION
- 1 MINUS ABSOLUTE CORRELATION
- 1 MINUS COSINUS ANGLE

VISUALIZATION OF DISTANCE MATRIX

Assign a color ranging (e.g.) from red (close) to blue (far) to each pairwise distance d_{ij} in the $n \times n$-distance matrix. Now, visualize the image.

VISUALIZING CLUSTERS:

1) order elements <u>within</u> clusters.
2) order clusters
3) visualize the <u>reordered</u> distance matrix.

Other visualisation tools: visualize elements in two dimensional plane by projecting on the space spanned by principal components, visualize reordered data matrix.

PARTITIONING ALGORITHMS

Possible partitioning algorithms are:

- PARTITIONING AROUND MEDOIDS(PAM). Choose K centers such that the sum of the distances to the closest center is minimal.
- PARTITIONING AROUND MEDOIDS MAXIMIZING AVERAGE SILHOUETTE. Given the cluster labels, for each element its silhouette is defined as the relative difference between average distance to its own cluster and average distance to the neighboring cluster: this is a number between -1 and 1 (Kaufman and Rousseeuw, 1990). Choose K centers (which define the clusters) so that the sum of the silhouettes is maximal.
- KMEANS. Choose K groups such that the sum of the distances to the closest cluster specific mean is minimal. The typical

implementation of KMEANS uses the Euclidean Distance.

- SELF-ORGANIZING MAPS. Similar as KMEANS, but it constraints the allowed set of partitions.
- HIERARCHICAL BINARY TOP-DOWN CLUSTERING. One splits the group in two clusters. Subsequently, one splits each of the two clusters in two to obtain 4 clusters and so on.

  Note, this restricts the class of allowed partitions: e.g., not each possible 4 groups is considered as an allowed clustering result.

Discuss a little simulation 3 groups of patients own groups of genes. Illustrate different dissimilarities are going to show different things. Show a picture of clustering results for PAM, PAMSIL.

HIERARCHICAL CLUSTERING

DOWN-TOP AGGLOMERATIVE CLUSTERING Start with single element clusters. Collapse the 2 closest clusters into one cluster and repeat this procedure till all elements are together. This produces a hierarchical tree. Each level correponds with a clustering result. Ordering of the clusters is completely determined by the initial ordering.

NON-BINARY HIERARCHICAL CLUSTERING Same, but allow partitioning in 2 or more clusters. If one orders the children of each parent cluster by their distance to closest uncle node, then running down the tree yields an ordered list. For details: **Hierarchical Ordered Partitioning and Collapsing Hybrid** (HOPACH) (van der Laan, Pollard, 2002).

SELECTION OF NUMBER OF CLUSTERS

A difficult (ill posed) problem!

Visualization of ordered distance matrix for different number of clusters K is a helpful tool to select number of clusters.

Formally, the idea is to come up with a criteria measuring strenght of a clustering result, which allows comparison of clustering results for different K, so that its maximum defines an "optimal" number of clusters.
The problem is: Different criteria give different "optimal clustering results".

A large collection of proposals have been made (for a overview of literature and new proposals, see papers on websites www.stat.berkeley/ dudoit and www.stat.berkeley/ laan)

CRITERIA REQUIRING RESAMPLING For example,

define optimal K in terms of 1) performance of cluster result **as classifier** (Dudoit, Frydland, 2002),

2) **variability** of clusters,

3) **statistical significance** of distance between clusters.

DIRECT CRITERIAS For example,

1) **average silhouette**,

2) **average of cluster specific homogeneities**.

STATISTICAL INFERENCE WITH MICROARRAY DATA

Mark J. van der Laan
UC Berkeley, Biostatistics

Fred Hutchinson Cancer Institute
March 31, 2000
Based on joint paper with
Jennifer Bryan.

NUMERICAL SUMMARY OF ONE MICROARRAY EXPERIMENT

Each microarray experiment yields a list X of p ratios representing the relative gene-expression profile.

TERMINOLOGY:

$X_j > 1$: gene j is overexpressed.
$X_j < 1$: gene j is underexpressed.
$X_j \neq 1$: gene j is differentially expressed.

PARTICULAR TYPE OF EXPERIMENT

EXPERIMENT: Randomly sample (e.g. colon, breast) cancer patients and for each patient

- Extract healthy and cancerous tissue.
- Carry out a microarray experiment to obtain the list of p ratios representing the relative gene-expression profile of cancerous versus healthy tissue for the p genes.

  Denote this list of ratios with $\mathbf{X}$. Let Y be the list of truncated log-ratios.

DATA SET: Our complete data set consists of n (samplesize) observations $Y_1, \ldots, Y_n$ of Y.

REMARK: Sample size n (e.g.100) is much smaller than number of genes p (e.g. 100,000)

SOME QUESTIONS ASKED

- What subset of the p genes cause cancer in an significant proportion of subjects, or at least are drug development targets?
- What groups of genes are dancing together.
- For the important findings, what is the probability that I can reproduce these findings?
- What sample size do I need?

SUBSET PARAMETERS

Let

$$\begin{aligned} \mu &\equiv EY \\ \Sigma &\equiv E\left\{(Y-\mu)(Y-\mu)^\top\right\} \\ \rho &= \text{CORRELATION MATRIX OF } \Sigma. \end{aligned}$$

Let

$$(\mu, \Sigma) \rightarrow \mathbf{S}(\mu, \Sigma) \in \{0, 1, \ldots, K\}^p$$

be a "subset rule" of interest.

DIFFFERENTIAL EXPRESSION RULES: Given user supplied

δ_1, δ_2

$$\begin{aligned}
\mathbf{S}(\mu) &= \{j : \mu_j > \delta_1\} \\
\mathbf{S}(\mu) &= \{j : \max(\mu_j, -\mu_j) > \delta_1\} \\
\mathbf{S}(\mu, \Sigma) &= \{j : \mu_j > \delta_1 - q_{0.7}\sigma_j\}
\end{aligned}$$

CLUSTERING RULES:

Given user supplied δ_1, δ_2

STEP 1: Apply *simple rule* to start with: e.g. Select all δ_1-differentially expressed genes.

STEP 2: Compute *distance matrix* $d = (d_{ij} : i, j)$ for remaining genes, using distance

$$d_{ij} = 1- \mid \rho_{ij} \mid .$$

Provide distance matrix to cluster program "Partitioning around Medoids" (PAM, Kaufman and Rousseeuw, 1990). This defines the clusters by the medoids and assigns a cluster-membership to each gene.

STEP 3: *Thin out* the clusters by deleting genes with links (or silhouette) weaker than δ_2. This also deletes False Positives.

SUPERVISED CLUSTERING

Find genes highly correlated with known master genes.

ESTIMATION AND CONSISTENCY

Let $(\mu_n, \Sigma_n, \rho_n)$ be the empirical counterparts of (μ, Σ, ρ). We estimate $S(\mu, \Sigma)$ with $S(\mu_n, \Sigma_n)$.

[Consistency]
Let $p = p(n)$ be such that $n/\log(p(n)) \to \infty$ as $n \to \infty$ and $M < \infty$. As $n \to \infty$, then

$$\sup_j |\mu_{n,j} - \mu_j| \to 0 \text{ in probability}$$

and

$$\sup_{ij} |\Sigma_{n,ij} - \Sigma_{ij}| \to 0 \text{ in probability.}$$

This implies $P(S(\mu_n, \Sigma_n) = S(\mu, \Sigma)) \to 1$ if $n \to \infty$ and $n/\log(p(n)) \to \infty$.

WHAT SAMPLE SIZE DO I NEED?

Let n^* be the sample size needed to make sure that with probability 0.95 the observed average expression level of EACH gene is within a DISTANCE ϵ of the TRUTH.

We can derive a closed form lower bound for this sample size in terms of maximal noise level, number of genes, wished precision ϵ.

It depends on the number of genes only through the logarithm of the number of genes!

Similarly, for the correlation matrix.

Thus data mining and fishing expeditions are allowed, just adjust sample size slightly

Put here the two sample size slides!

NONP. SAMPLE SIZE FORMULA

Let $\sigma = \max_j \sigma_j$. Define

$$n^*(p, \epsilon, \delta, M, \sigma^2) = \frac{1}{c(\epsilon, \sigma, M)} \{\log(p) + \log(2/\delta)\},$$

where

$$c = c(\epsilon, \sigma^2, M) \equiv \frac{\epsilon^2}{2\sigma^2 + 2M\epsilon/3}.$$

If $n > n^*(p, \epsilon, \delta, M, \sigma^2)$, then

$$P\left(\max_j |\mu_{n,j} - \mu_j| > \epsilon\right) < \delta$$

With this formula we can compute the sample size for which the probability that "low-differentially expressed genes make it into $S(\mu_n, \Sigma_n)$" is smaller than $\delta = 0.05$.

For example, $(\log(3) - \log(2) = 0.41)$

$$
\begin{aligned}
n^*(100000, \epsilon = 0.41, 0.1, 2, 0.5) &= 133 \\
n^*(5000, \epsilon = 0.1, 0.1, 2, 0.5) &= 1304 \\
n^*(5000, \epsilon = 0.5, 0.1, 2, 0.5) &= 77 \\
n^*(5000, \epsilon = 0.5, 0.01, 2, 0.5) &= 92 \\
n^*(5000, \epsilon = 1.0, 0.05, 2, 0.5) &= 28
\end{aligned}
$$

SIMULATION FOR UNIFORM DIFFERENCES

Noise Level: Suppose that all genes are independent with standard deviation $\sigma = 0.5$.

Sample size: Suppose that we have 150 subjects.

	10	1000	10000	100000
Max.Diff.Means	0.08	0.14	0.16	0.18
0.9-Quant	0.09	0.16	0.18	0.19
Max.Diff.Stdev	0.06	0.10	0.12	0.13
0.9-Quant	0.08	0.11	0.13	0.14

	p=10	p=100	p=1000
Max.Diff.Cor	0.2	0.31	0.39
0.9-Quant	0.23	0.33	0.41

If we set $n = 200, p = 1000$, then DIFCOR=0.34.

If we set $n = 1000, p = 1000$, then DIFCOR=0.15.

Suppose now that the true correlations between M independent pairs of variables is 0.8.

	10	1000	10000	100000
Max.Diff.Cor	0.07	0.15	0.18	0.22
0.9-Quant	0.09	0.16	0.2	0.24

Remark that in subset rule we only apply clustering and thus look at correlations for genes which make first cut of. e.g. 300 genes. Since mu_n is independent of Σ_n this is fine. So for knowing how good our correlation matrices for the clusters are we only have to set $p = 300$.

PARTITIONING AROUND MEDOIDS

- Define a distance matrix for the elements (genes or subjects)to be clustered: e.g. for genes we use correlation or absolute correlation distance.
- Provide distance matrix to cluster program "Partitioning around Medoids" (PAM, Kaufman and Rousseeuw, 1990) and specify number of clusters.

  This finds data adaptively the best centers (medoids) of the clusters and assigns a cluster-membership to each element.
- For each element it computes a silhouette measuring how strong it belongs to its cluster.
- Number of clusters is obtained by minimizing average silhouette.

put here ALL, AML clustering subjects plots

HOW RELIABLE???

How reliable is the observed structure or observed subset and clusters?

Examples of observed features:
4 genes in the same cluster.
Gene has correlation larger than 0.5 with a master gene.
Gene is more than 3-fold differentially expressed.

For example, if we repeat the experiment, how *likely* is it that one can reproduce the findings?

BOOTSTRAP: Simulate an approximation of the experiment many times and find out!

CONFIDENCE-LEVEL PARAMETERS OF INTEREST

Possible parameters of the distribution of $\widehat{S} \equiv \mathbf{S}(\mu_n, \Sigma_n)$ are:

Feature Probabilities:

Consider an observed feature: e.g.

1) gene j is in the subset estimate.

2) genes i, j were both in the subset estimate.

3) genes i, j were both in the subset estimate and in the same cluster. We can define the corresponding feature probability:

$$\begin{aligned} p_{j,n} &= P(\widehat{S}_j > 0) \\ P_{ij,n} &= P(\widehat{S}_i > 0, \widehat{S}_j > 0) \\ Q_{ij,n} &= P(\widehat{S}_i = \widehat{S}_j > 0) \end{aligned}$$

RESULT: If $n \to \infty$, then these probabilities converge uniformly to the features of the true subset $S = \mathbf{S}(\mu, \Sigma)$, even when $p = \infty$.

Performance measures:

“Sensitivity” and “Positive Predictive Value” of $\widehat{S}$:

$$\begin{aligned} sens_n &= E\left\{\frac{|S \cap \widehat{S}|}{|S|}\right\} \\ ppv_n &= E\left\{\frac{|S \cap \widehat{S}|}{|\widehat{S}|}\right\}. \end{aligned}$$

The distribution of the proportion of “Extreme False Positives” which make it into the subset estimate.

Uniform Distances. 0.95-quantiles of

$$\begin{aligned} \text{Maxdif.mean} &= \max_j \mid \mu_{nj} - \mu_j \mid \\ \text{Maxdif.cor} &= \max_{ij} \mid \rho_{n,ij} - \rho_{ij} \mid . \end{aligned}$$

PARAMETRIC BOOTSTRAP

RESAMPLING: We estimate distribution of $\widehat{S}$ by resampling from an estimated distribution of the true data generating distribution. Since $\widehat{S}$ only depends on the mean and covariance matrix, we want to resample data with the (asymptotically) the right mean and right covariance matrix (and we want NO TIES).

RESAMPLE FROM A MULTIVARIATE NORMAL DISTRIBUTION:

- Resample n observations $Y_1^{\#}, \ldots, Y_n^{\#}$ of $Y^{\#} \sim N_p(\mu_n, \Sigma_n)$. Construct estimate $\mathbf{S}(\mu_n^{\#}, \Sigma_n^{\#})$.
- Repeat: Obtain B i.i.d observations of $\mathbf{S}(\mu_n^{\#}, \Sigma_n^{\#})$.
- Computate relevant parameters of this empirical distribution of $\mathbf{S}(\mu_n^{\#}, \Sigma_n^{\#})$.

ASYMPTOTIC VALIDITY:

Nonparametrically, if $n/\log(p(n)) \to \infty$ and $M < \infty$, we have that
1) the bootstrap estimate of the distribution of $\sqrt{n}(\mu_n - \mu)$ is consistent and
2) $S(\mu_n^{\#}, \Sigma_n^{\#})$ converges to the degenerate distribution at $S(\mu, \Sigma)$.

Thus the estimated feature probabilities converge to the true features: e.g. $p_{j,n}^{\#} \to I(j \in S)$.

SIMULATING THE NULL DISTRIBUTION

To make sure that observed structures are not due to pure noise, one simulates from a multivariate normal distribution with either

1: No differential expression and no correlations or

2: No differential expression and observed covariance matrix.

SIMULATION STUDY

SAMPLE SIZE: 60.

NUMBER OF GENES: 1500

TRUE COVARIANCE MATRIX: block diagonal with three blocks of correlated genes.

SUBSET RULE: PAM-based subset rule with three clusters applied to δ_1-differentially expressed genes. We required sufficiently small distance between medoid or any previously included gene: see table.

TRUE SUBSET: Apply subset rule to true (μ, Σ).

TRUE FEATURE PROBABILITIES: p_j are the proportion of times gene j falls in subset estimate in the actual simulation.

TRUE SENSITIVITY, PPV etc: Similar.

SIMULATION: 1) Sample 60 subjects from true distribution. 2) Do the parametric bootstrap (200 resamples) to obtain estimates of the feature probabilities $p_{j,n}$ and other quantaties of interest. 3) repeat 1) and 2) 200 times.

Mean Cutoff	Correlation Cutoff
$\lvert\mu_j\rvert > \log 2.7 \approx 0.99$	$\lvert\rho_{ij}\rvert > 0.5$

$\lvert S\rvert$	avg $\lvert\widehat{S}\rvert$	avg $\lvert\widehat{S}^{\#}\rvert$
30	26.9	26.61

$p = 1500, n = 60$	True	Bootstrap
Sensitivity	0.73	0.78
Predictive Value	0.82	0.79
Prop. of Ext. False Pos.	0.00	0.00
Any Ext. False Pos.	0.00	0.00
Expected Lgst. Abs. Dev.	0.46	0.46

we need the table with p_j probabilities. show clustering subjects ALL, AML, show probability plot for 9 clusters and show one of cluster plots.