

Methods for analysis of gene expression data with a right-censored outcome

Katherine S. Pollard & Mark J. van der Laan

Division of Biostatistics, U.C. Berkeley

www.stat.berkeley.edu/~laan/

Motivation: Microarray Data

- Observe a matrix X whose columns are n copies of a p -dimensional vector of relative gene expression measurements.
- Each measurement is a ratio, calculated from the intensities of two fluorescently labeled mRNA samples cohybridized to an array spotted with known cDNA sequences.
- Data preprocessing may include background subtraction, normalization, log transformation.
- Also observe an n -dimensional vector Y , which is a (possibly censored) outcome for each sample.

e.g.: Tumor vs. healthy tissues of n cancer patients plus survival.

NOTE: Methodology also applies to gene chips, where each element of X is a quantitative expression level rather than a ratio.

Parameters of Interest

Gene Expression Only	Gene Expression with Outcome Y
Significantly differently expressed genes	Genes significantly associated with Y
Clusters with similar expression profiles	Clusters with similar association profiles
Subpopulations with similar expression	Subpopulations with similar association
Reduced dimension expression fingerprints	Reduced dimension (independent) effects on Y

Transformation Method

1. Model for marginal association of X_j with Y :

$$Y = m(X_j | \beta_j) + \epsilon_j, \quad j = 1, \dots, p$$

2. Transformation $\Phi(X_j, Y | P) = \tilde{X}_j, \quad j = 1, \dots, p$:

- Regression coefficients $\hat{\beta}$
- Residuals $\epsilon_j(\beta), \epsilon_j(\beta)/X_j$
- Influence functions

3. Subsetting and Clustering of \tilde{X} :

- $S(P) = \tilde{S}$
- Choice of distance metric defines D

Estimation and the Bootstrap

Let P_n be the empirical distribution of (X, Y) , then

$$\tilde{X}_n = \Phi(X, Y \mid P_n),$$

$$\tilde{S}_n = S(P_n).$$

The parametric or non-parametric bootstrap can be used to estimate the variability of an estimated clustering parameter \tilde{S}_n .

Refs: van der Laan & Bryan (2000), Pollard & van der Laan (2001).

Censoring

Suppose Y is right-censored for some subjects:

$$(X_i, Y_i, C_i), i = 1, \dots, n.$$

Inverse Probability of Censoring Weighted (IPCW) estimators for the mean μ and covariance $\Sigma = \{\sigma_{jl}\}$ of \tilde{X} :

$$\hat{\mu}_j = \hat{E}(\tilde{X}_j) = \frac{1}{n} \sum_{i=1}^n \frac{\tilde{X}_{ij} \Delta_i}{\bar{g}(Y_i | X)}, j = 1, \dots, p,$$

$$\hat{\sigma}_{jl} = \hat{Cov}(\tilde{X}_j, \tilde{X}_l) = \frac{1}{n} \sum_{i=1}^n \frac{\tilde{X}_{ij} \tilde{X}_{il} \Delta_i}{\bar{g}(Y_i | X)}, j = 1, \dots, p, l = 1, \dots, p,$$

where $\Delta_i = I[Y_i < C_i]$ and $\bar{g}(t | X) = P(C_i > t | X)$ is the probability that subject i was still at risk at time t given his/her gene expression profile.

Then, IPCW estimator of dissimilarity matrix is: $\hat{D} = D(\hat{\mu}, \hat{\Sigma})$.

Example: Data Generation

- $p = 1010$ genes, $n = 30$ patients
- Causal gene g_1 is $N(-9, 0.75)$
- Genes g_2, \dots, g_{10} are g_1 plus random $N(0, 0.05)$ noise
- Genes g_{11}, \dots, g_{1010} are 8 groups of 125 genes which are $N(m, 0.75)$, where $m \in \{-9, -8, -5, -4, 4, 5, 8, 9\}$.
- Multiplicative intensity model for survival time T :

$$\log(T) = \beta_0 + \beta_1 g_1,$$

where $\beta_0 = 0$ for half of the subjects and $\beta_0 = -2$ for the other half. For all subjects, $\beta_1 = -1$.

- Censoring time $\log(C)$ is $U(0, M/q)$, where $M = \max(\log(T))$ and q is the expected proportion of censored subjects.

Example: Method

1. Linear regression of each gene's expression on $\log(T)$
2. T statistics $\hat{\beta}_{1j}/se(\hat{\beta}_{1j})$ for significantly associated genes
3. Residual transformation for \tilde{X}
4. Euclidean distance matrix \hat{D}
5. Cluster genes with PAM using MSS to choose the number of clusters
6. Cluster patients within gene clusters
7. Censoring:
 - Cox proportional hazards model for the censoring mechanism $\bar{g}(t | X)$ to get IPCW estimates $\hat{\mu}, \hat{\Sigma}$
 - Cluster with $\hat{D} = D(\hat{\mu}, \hat{\Sigma})$

Example: Results

1. Without Censoring:

- T statistics significant ($\alpha = 0.05$ plus Bonferoni adjustment) for 10 genes, 7 of which are in g_1 's group.
- Two gene clusters: g_1 's group versus the other 1000 genes.
- Two patient clusters using only g_1 's group.

2. With Censoring:

- Gene cluster 1 contains g_1 's group plus 9 and 40 other genes, for 20% and 30% censoring, respectively.
- This is a problem with estimating \hat{D} ; when $n = 100$ subjects are used, cluster 1 contains only g_1 's group.

Summary

1. Methods for gene expression data applied to transformed data matrices.
2. Interesting gene and patient clusters.
3. IPCW estimators for right-censored data.
4. Additional results:
 - Easily adjust for covariates in regression models.
 - Bootstrap methods for estimating the variability of subsetting and clustering results.
 - Data Analysis: Gene expression and right-censored survival for 40 DLBCL patients (Alizadeh *et al.*, 2000)