Model-based Identification of Outliers in Gene Expression

Xiaogang Zhong, Giovanni Parmigiani

November 6, 2007

Background

- Cancer is a genetic disease caused by genomic mutations that confer an increased ability to proliferate and survive in a specific environment.

Background

- Cancer is a genetic disease caused by genomic mutations that confer an increased ability to proliferate and survive in a specific environment.
- Oncogenes that have heterogeneous activation patterns have been observed in the majority of cancer types.

Background

- Cancer is a genetic disease caused by genomic mutations that confer an increased ability to proliferate and survive in a specific environment.
- Oncogenes that have heterogeneous activation patterns have been observed in the majority of cancer types.
- Genes of interest are expected to be differentially expressed in a small subset of the samples, which, statistically, behave like "outliers".

Background

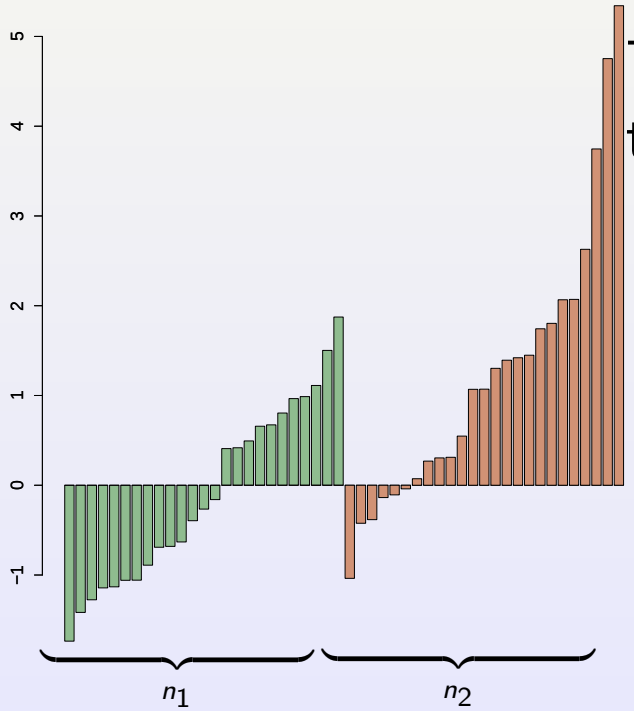
- Cancer is a genetic disease caused by genomic mutations that confer an increased ability to proliferate and survive in a specific environment.
- Oncogenes that have heterogeneous activation patterns have been observed in the majority of cancer types.
- Genes of interest are expected to be differentially expressed in a small subset of the samples, which, statistically, behave like "outliers".
- The detection may lead to implications in the development of carcinomas and the molecular diagnosis and treatment of cancers.

Outlier detection methods

Consider a two-class, for example, cancer/normal tissues, microarray data. Let X_{ij} be the observed expression values for samples $i = 1, \ldots, n$ and genes $j = 1, \ldots, g$.

Outlier detection methods

Consider a two-class, for example, cancer/normal tissues, microarray data. Let X_{ij} be the observed expression values for samples $i = 1, \ldots, n$ and genes $j = 1, \ldots, g$.



Traditional analytical methods, for example, t-statistic:

$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{s_j} \sqrt{\frac{n_1 n_2}{n}},$$

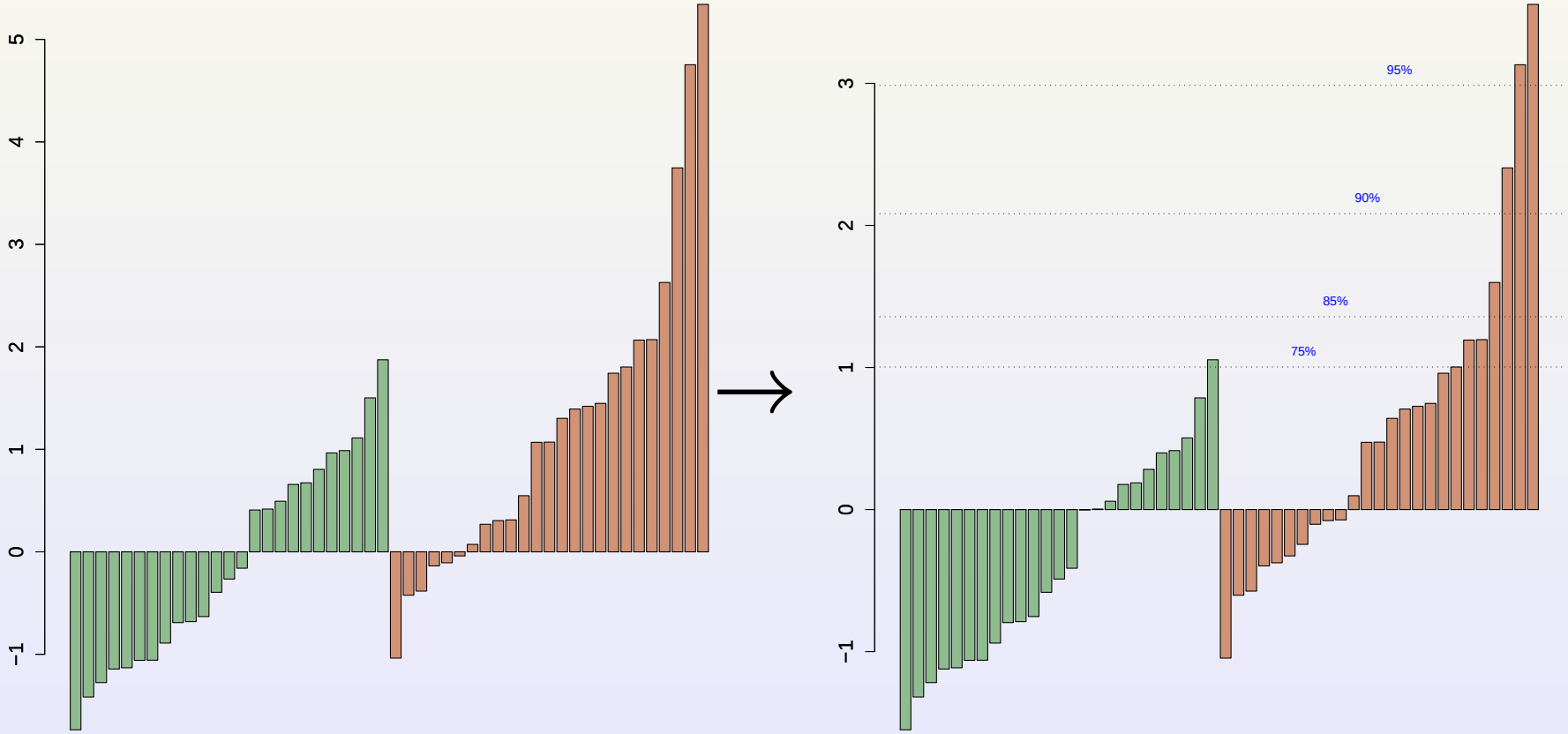
$$s_j^2 = \frac{\sum_{i \le n_1} (x_{ij} - \bar{x}_{1j})^2 + \sum_{i \ge n_2} (x_{ij} - \bar{x}_{2j})^2}{n - 2}.$$

Cancer Outlier Profile Analysis

1. Center and scale the data (on a row-wise basis) using the median and median average difference (MAD).
2. Select a common value for percentile as a cutoff for "outlier" status and apply this to all genes.
3. Look for pairs of genes that have a large number of mutually exclusive outlier (cancer) samples, but few or no normal outliers.
4. Rank the candidate gene pairs based on the sum of outlier samples for each pair.

Tomlins et al., Science, 2005

Cancer Outlier Profile Analysis (Cont'd)



COPA Generalization

- OS[1] (outlier sum statistics), centers and scales the genes in the same way as COPA. A threshold for outliers is defined as,

$$c_j = IQR(\tilde{x}_j) + q_{0.75}(\tilde{x}_j)$$

where $\tilde{x}_j$ is the expression of gene j after the normalization and IQR is the interquantile range.

- Outlier robust t-statistic[2] replaces the overall mean by the mean of the normal samples, and the overall MAD by the adjusted MAD,

$$c'_j = IQR(\tilde{x}_{ij}) + q_{0.75}(\tilde{x}_{ij}),\ i \in N,$$

where N is the set of normal samples.

[1] Tibshirani et al., Biostatistics, 2006
[2] Wu et al., Biostatistics, 2006

POE(Probability Of Expression)

- POE is an expression-based molecular classification method to discover novel biological classes and identify genes associated with them.
- The key idea of this method is that it models the gene expression using the latent categories that a gene is turned "on" or "off" compared to the baseline genes, and therefore estimates the probabilities of being differentially expressed.
- This approach defines three categories from which x_{ij} could have arised, and uses e_{ji} to represent them:

$$\begin{aligned} e_{ij} &= -1 && \text{gene } j \text{ has abnormally low expression} \\ e_{ij} &= 0 && \text{gene } j \text{ has baseline expression} \\ e_{ij} &= 1 && \text{gene } j \text{ has abnormally high expression.} \end{aligned}$$

POE (Cont'd)

For each j, the distribution of x_{ij} given e_{ij} follows probability distribution $f_{e;j}$,

$$x_{ij}|e_{ij} = e \sim f_{e;j}(\cdot), \qquad e \in \{-1, 0, 1\}.$$

The standard implementation of POE uses uniform distributions (U) for $f_{1,j}$ and $f_{-1,j}$ and a normal distribution Φ for $f_{0,j}$. More specifically,

$$\begin{aligned} f_{-1,j}(\cdot) &= U(-\kappa_j^- + \alpha_i + \mu_j, \alpha_i + \mu_j) \\ f_{0,j}(\cdot) &= \Phi(\alpha_i + \mu_j, \sigma_j) \\ f_{1,j}(\cdot) &= U(\alpha_i + \mu_j, \alpha_i + \mu_j + \kappa_j^+), \end{aligned}$$

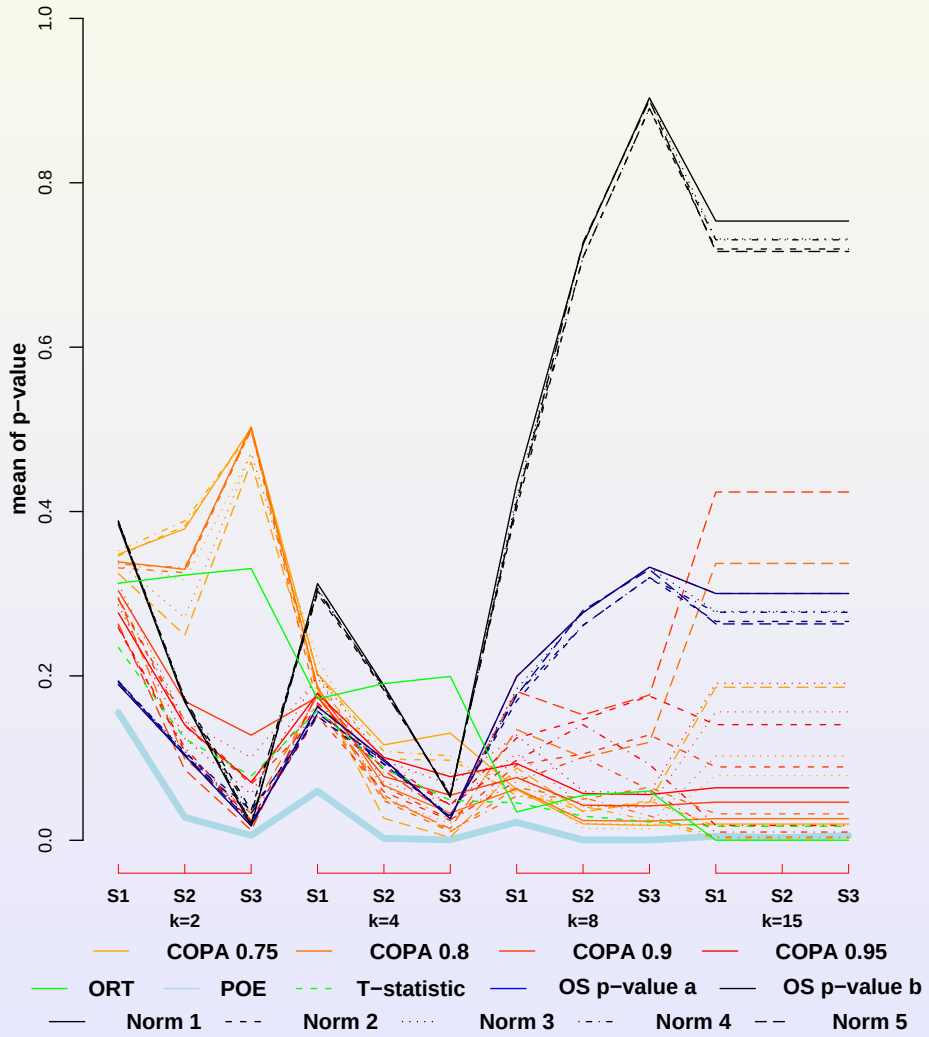
where U is the uniform distribution and Φ is the Gaussian (normal) distribution.

Simulation Studies

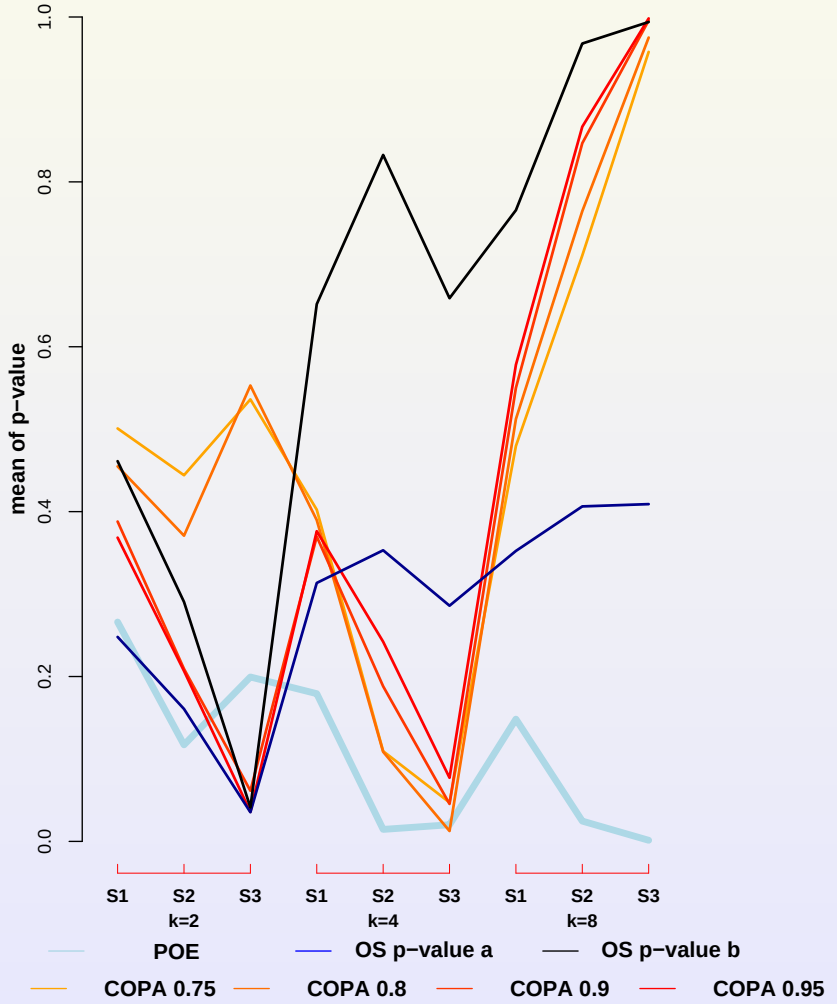
- Consider a 1000×30 expression matrix, in which the first 15 samples form the control group and the rest form the tumor group.
- For each simulation, the data is generated from the standard normal distribution independently for each gene and sample.
- The first gene is the one including the true outliers. For the first gene, two units are added to k samples, $k = 2, 4, 8, 15$.
- For each k and simulation method, 100 simulations are conducted for the four approaches.
- A p-value is calculated for the various configuration based on each simulation.

Simulation Studies Results

(a) The control group is known



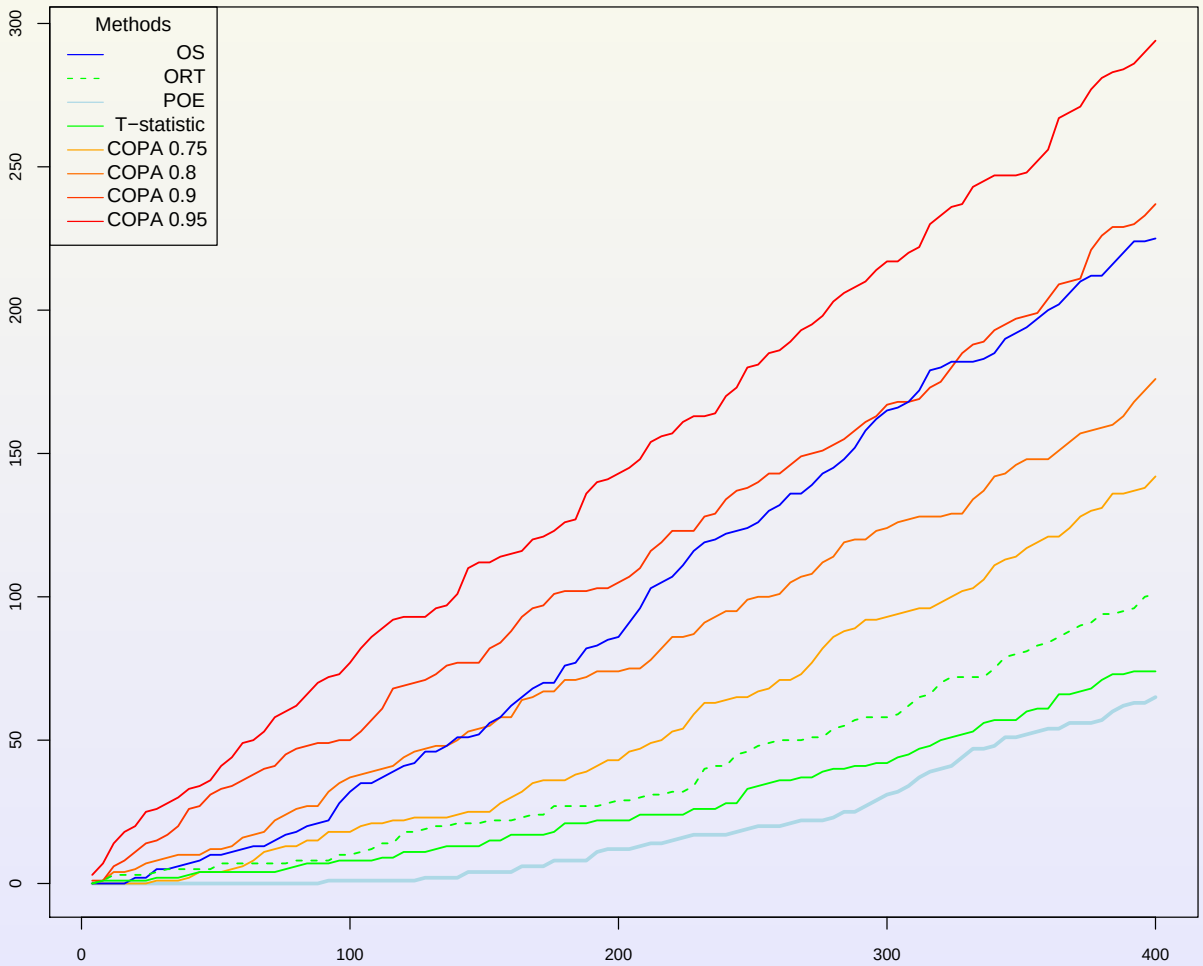
(b) The control group is not known



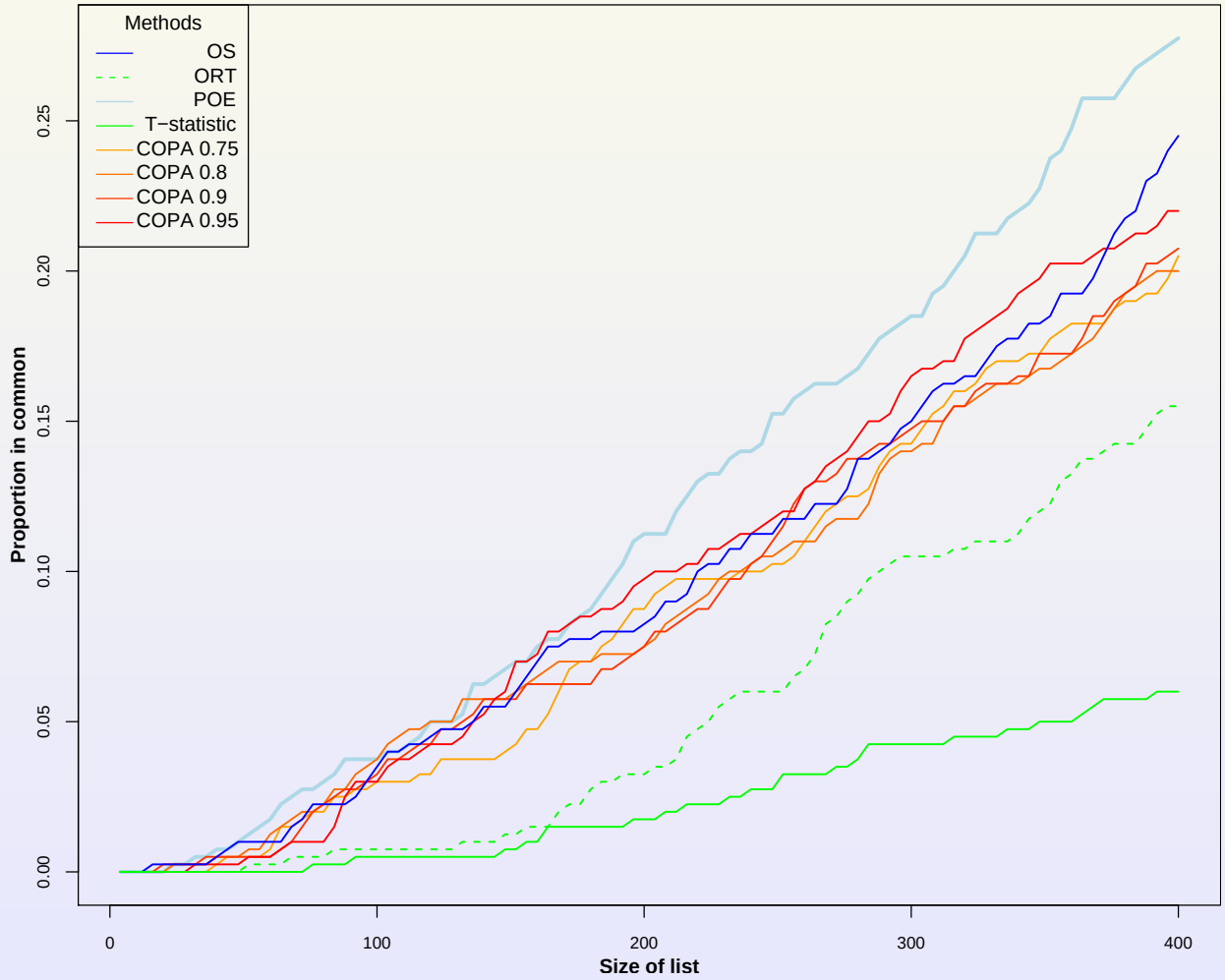
Application Studies

- We compare the five methods on a dataset from a study of transcriptional response of patients after radiation therapy. In our comparison, we consider the difference of expression between ionizing radiation treatment and a control.
- Different methods are compared by measuring the performance in controlling the false discovery rate and the consistency in discovering outlier genes across artificial splits of the data.

False Discovery Rate Control



Consistency of Detection



Summary

- Based on our experiments, POE had the best performance under almost all circumstances compared to the others.
- The benefit of this approach is that it borrows strength across genes using the the entire genomic distribution instead of fitting a separate, independent model for each gene.
- POE can be extended to provide probabilistic statements about the assignment of tumors to molecular profiles, so that it gives hope for more effective molecular prognosis and treatment of cancer.