

# Rank-Based Classification of Gene Expression Profiles

Daniel Q. Naiman<sup>‡</sup>

Collaborators:

Donald Geman<sup>†‡</sup>, Christian d'Avignon<sup>†§</sup> & Raimond L. Winslow<sup>†§</sup>

<sup>‡</sup>Department of Applied Mathematics and Statistics

<sup>†</sup>Center for Cardiovascular Bioinformatics and Modeling, Whitaker Biomedical Engineering Institute

<sup>§</sup>Department of Biomedical Engineering

Johns Hopkins University

Baltimore, MD

# Basic approach to classification using gene expression

Use pairwise comparisons between gene expression levels in pairs as a *feature* for classification.

## Motivations

- the small sample dilemma
- parsimony/interpretability
- transparency - invariance to normalization
- experimental evidence

# Microarray Data Analysis

## Expression data:

$G \times n$  matrix with labeled columns

$G$  = number of genes/EST's

$n$  = number of samples (tissues)  
obtained under various  
biological conditions

column labels indicate *class* of samples e.g.

- tumor/normal
- disease/non-disease

# Typical Experimental Objectives

**Clustering** – group genes or samples in meaningful ways

**Modeling** – describe statistical behavior of expression levels

- marginal behavior for individual genes
- joint behavior for multiple genes

**Classification** (the focus of this talk) – predict classes e.g.

- cancerous tumor vs. normal tissue
- treatment outcome (success/failure)
- disease type

# Statistical Perspective: Small Sample Dilemma

- **Problem:** Small number of experiments ( $n$ ), typically tens, relative to the number of genes ( $G$ ), typically thousands.
- **Example:**  $n = 34$  samples, and  $G = 7,129$  genes.
- **Consequence:** Standard methods in machine learning algorithms are “tuned” (outside of the CV loop!!!) often lead to over-fitting and inflated estimates of performance.

# The Bias Variance Tradeoff

- **Machine learning community mantra:**

Complex models lead to **low bias/high variance**.

Simpler models give rise to **high bias/low variance**.

- **Consequence:** Minimization of error rates can result from choosing models in a smaller class.

# Biological Perspective: Interpretability/Parsimony Dilemma

- **Problem:** The decision boundary generated by standard classifiers can often be highly complex
- **Examples:** Support-vector machines, neural networks, random forests, logitboost, nearest neighbors.
- The manner in which decisions are made too much resembles a *black box*, and decision rules are lacking in transparency.
- We seek **transparent classifiers** involving small numbers of genes.

# Mathematical Formulation

- **Expression random variables:**  $X = (X_1, \dots, X_G)$ .
- **Class random variable:**  $Y \in \{1, 2\}$
- **Classifier:**  $f : \mathbb{R}^G \rightarrow \{1, 2\}$
- **Training data:**  $L$  a matrix consisting of  $n = n_1 + n_2$  columns (expression profiles) where  $n_k$  of the columns are iid samples of  $X$  given  $Y = k$  for  $k = 1, 2$ .
- **Learning algorithm:** Mapping  $S$  that assigns a classifier  $f_L$  for every choice of training data  $L$ .
- **Generalization error:**  $e(f) = P[f(X) \neq Y]$  the probability of making an error on a future profile (depends on  $L$  and the distribution of  $(X, Y)$ ).
- **Estimated error rate:** An estimate of  $e(f)$  from data.

# Pairwise Comparison

Focus on detecting “marker gene pairs”  $(i, j)$  whose expression values *invert* in going from class 1 to class 2, that is, for which

$$p_{ij}(k) := P[X_i < X_j | Y = k]$$

changes considerably when changing from  $k = 1$  to  $k = 2$ . These probabilities are estimated by relative frequencies of occurrences of

$$X_i < X_j,$$

*within* profiles and over samples.

# “Scoring” Gene Pairs

Define a “score” associated with each gene pair  $(i, j)$

$$\Delta_{ij} = \left| p_{ij}(1) - p_{ij}(2) \right|$$

We seek pairs  $(i, j)$  with high scores  $\Delta_{ij}$ .

# Gene Pair Score Example

	$X_i < X_j$	$X_i > X_j$	
class 1	17	4	21
class 2	4	35	39

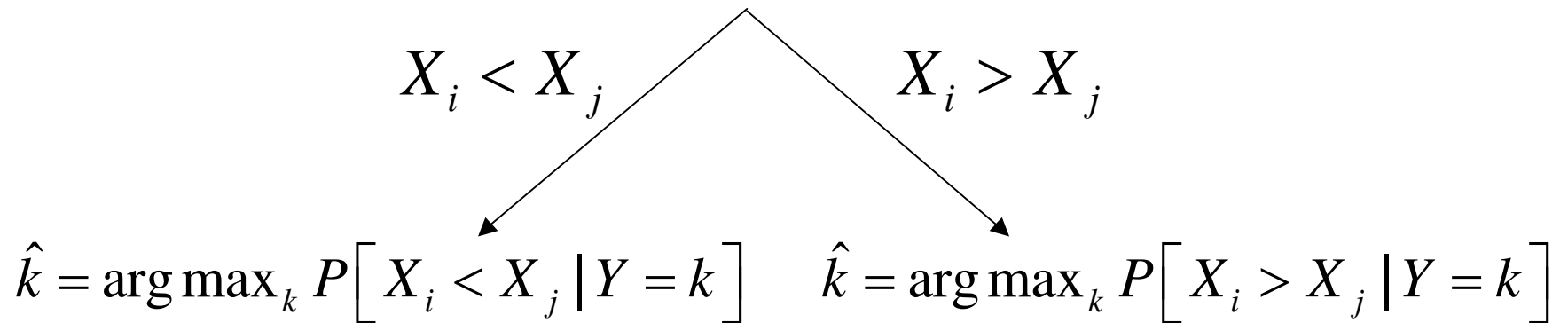
$$n_1 = 21 \quad \hat{p}_{ij}(1) = 17/21$$

$$n_2 = 39 \quad \hat{p}_{ij}(2) = 4/39$$

$$\hat{\Delta}_{ij} = |17/21 - 4/39| = .707$$

# Interpretation of the Score

Consider classification “**stump**” based on the *feature* defined by the indicator  $I(X_i < X_j)$ :



$$\begin{aligned} \text{Sum of error probs} &= P[\hat{k} = 2 | k = 1] + P[\hat{k} = 1 | k = 2] \\ &= 1 - \Delta_{ij} \end{aligned}$$

# Gene Pair Selection

- **Estimate**  $\Delta_{ij}$  for all gene pairs  $(i,j)$  .
- **Rank** all pairs  $(i,j)$  based on  $\hat{\Delta}_{ij}$ .
- **Select** all of the pairs  $(i,j)$  attaining the maximum score (ties are common).

# The Top Scoring Pair (TSP) Classifier

- **Pair selection** results in a family  $\mathcal{P}$  of **distinguished top scoring pairs**.
- We seek **classification decisions that are easily interpreted**.
- **Voting** is an example of an easy to interpret algorithm.
- Let each pair  $(i, j) \in \mathcal{P}$  **vote using the maximum likelihood scheme** described above.
- Make a **majority rules decision**.

# Voting and Maximum Likelihood

Under the following assumptions, the **majority rules procedure** can be interpreted as a *maximum likelihood* estimate of the class:

- all *informative* pairs are included
- individual comparisons are conditionally independent given the class  $k$
- for some  $p$  we have either

$$p_{ij}(k) = p \quad \text{or} \quad p_{ij}(k) = 1 - p$$

for all  $(i, j) \in \mathbb{I}^{\mathcal{P}}$  and for all classes  $k = 1, 2$

# Miscellaneous Remarks

- The TSP classifier is **rank-based** hence **invariant** to a large class of **normalization methods** (monotone transformations)
- **NO PARAMETERS TO TUNE** in TSP leading to **HONEST ERROR RATES**.
- Natural generalization to **k-TSP** where we choose the **k top scores**
  - k determined inside a **cross-validation loop** (double CV)
  - method remains **rank-based**, hence **invariant** as above
- Bø and Jonassen (2002) introduced an indirect approach to selecting gene pairs involving profile classification, linear discriminant analysis, and nearest neighbors.

# Miscellaneous Remarks (cont.)

- Another approach to selection is possible, where, first attention is restricted to differentially expressed genes
  - possible to miss certain gene pairs when both are not significantly differentially expressed
  - loss of invariance to normalization
- A gene may appear in more than one TSP, and this typically occurs

# Class Prediction Problems

- **Cardiac study:** Classifying tissue samples of patients diagnosed with idiopathic dilated cardiomyopathy (IDCM) vs. control.

3 publicly available studies from the Kent Ridge Bio-medical Data Set Repository

- **Survival study:** Predicting outcomes of treatment for tumors of the central nervous system.
- **Leukemia study:** Classifying profiles into leukemia subtypes
- **Prostate study:** Distinguishing prostate cancers from normal profiles.

# Data Set Parameters

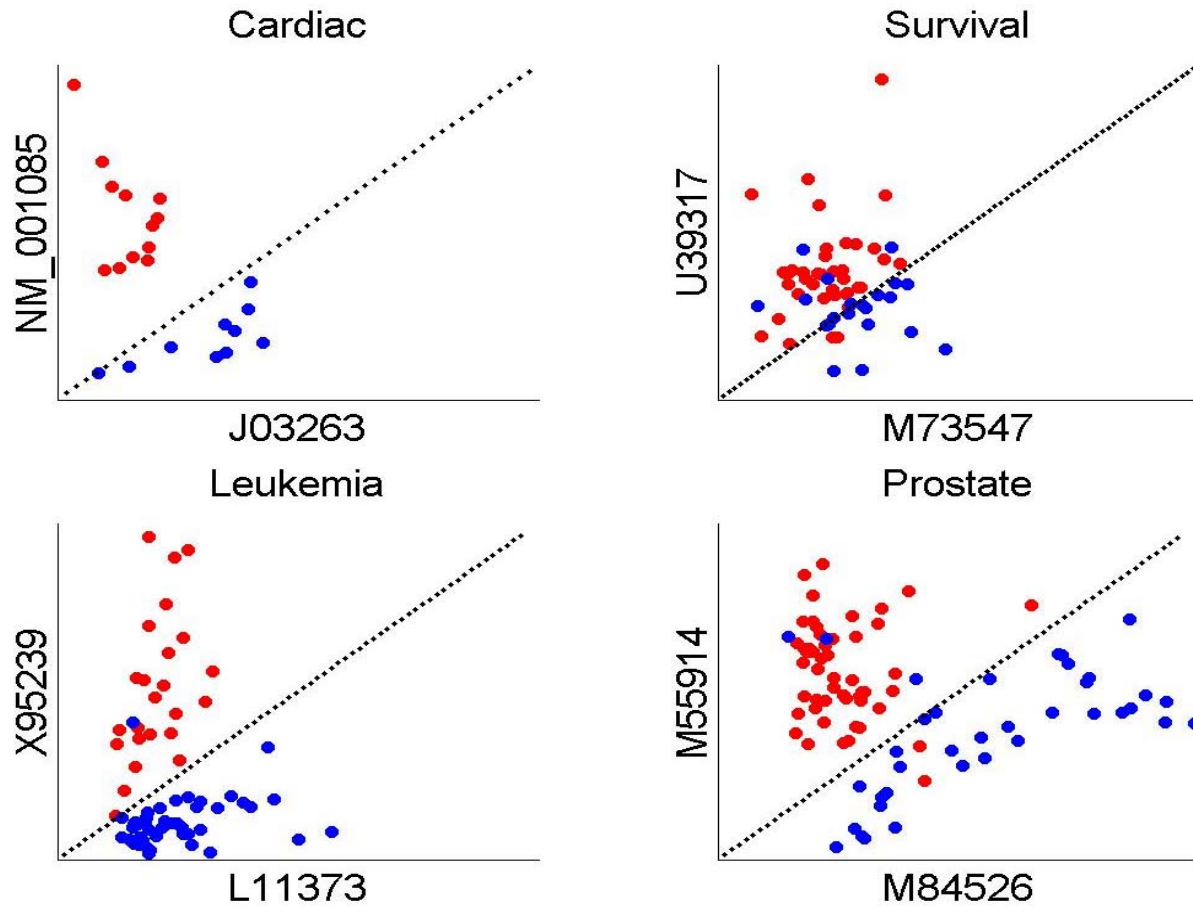
Study	$G$	$n$	class 1	class 2
Cardiac	22,283	22	10 normal	12 IDC
Survival	7,129	60	21 non-survivor	39 survivor
Leukemia	7,129	72	47 ALL	25 AML
Prostate	12,600	102	52 tumors	50 normal

# Numbers of Top Scoring Pairs

Generally, the larger the sample size is large relative to the number of genes the fewer TSPs we expect to see.

Study	Number of TSPs
Cardiac	2,460
Survival	1
Leukemia	3
Prostate	1

# TSP Classification



# Performance Comparisons (Classification Rates by LOOCV)

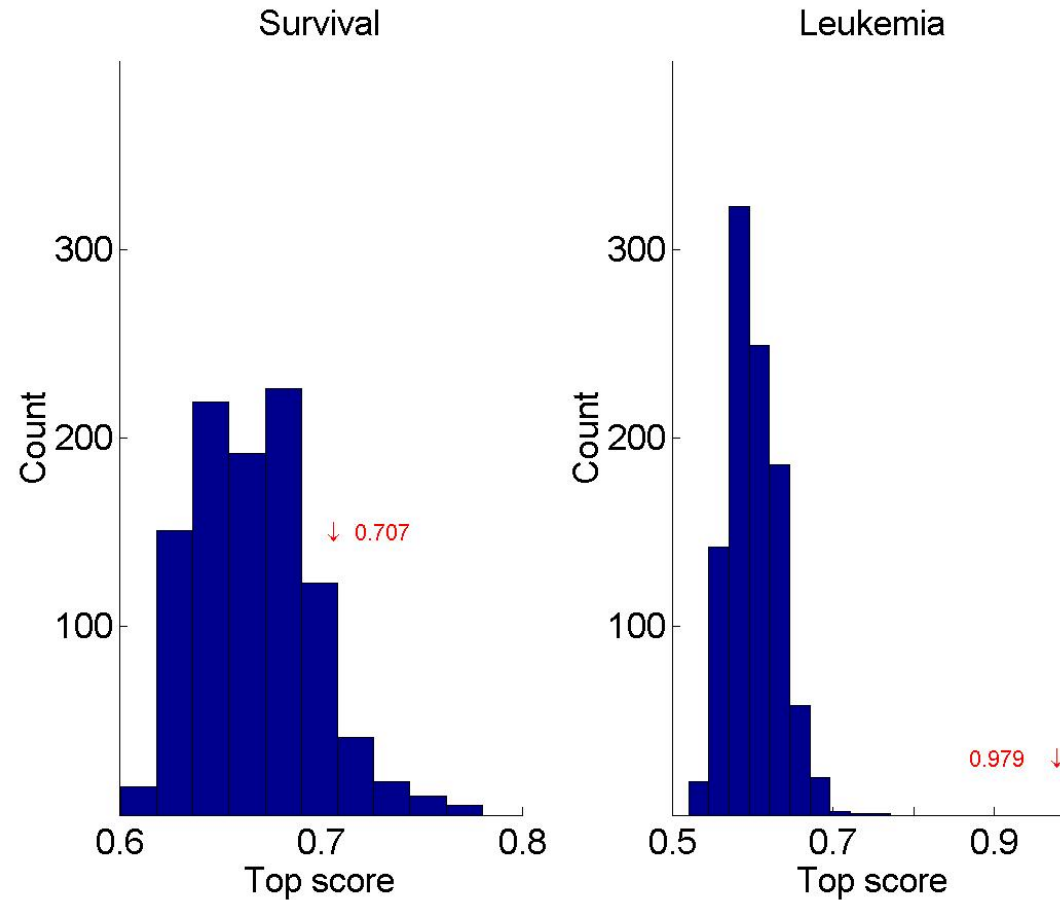
Study	TSP	Previous results
Cardiac	100%	100%
Survival	83%	47%-77%
Leukemia	94%	85%, 95%
Prostate	95%	86%-92%

# Significance by Permutation Analysis

Create artificial data sets by random permutations of column labels

- maintain sample sizes of the two classes
- preserve statistical dependency structure among genes
- resulting top scores in artificial data are indicative of scores obtained when attempting to classify based on profile labels that cannot be predicted from expression values

# Histograms of Simulated TSPs



# Permutation Analysis Results

Study	Simulated p-value
Cardiac	large
Survival	.10
Leukemia	0
Prostate	0

(Based on 1,000 permutations)

# Conclusions from Permutation Analysis

Prostate/Leukemia studies → Clear statistical significance of TSPs

Survival study → Ambiguous

Cardiac study → Insignificant\*

\***Note:** Despite this, there must be informative pairs since otherwise, random voting in the LOOCV would lead to poor classification results.

# Individual t-Statistics of TSPs

Study	Score	Genbank ID 1	t-stat1	Genbank ID 2	t-stat2
Survival*	.707	M73547	2.82	U39317	3.23
Prostate	.902	M84526	7.46	M55914	4.13
Leukemia	.979	L11373	1.99	X95735	10.92
Leukemia	.979	D86976	1.60	X95735	10.92
Leukemia	.979	J05243	7.87	M23197	6.62

\*Neither gene for the TSP in the survival study shows significant differential expression by itself.

# Conclusions

TSP classifier appears to have many desirable properties:

- Simple model /Easily interpretable
- Competitive statistical performance
- Invariant to normalization
- Generalizes to more complex (but still simple) k-TSP classifier