

A GENE SELECTION CRITERION FOR DISCRIMINANT MICROARRAY DATA ANALYSIS BASED ON EXTREME VALUE DISTRIBUTIONS

Wentian Li¹, Ivo Grosse^{2,3}

1. Center for Genomics and Human Genetics, North
Shore LIJ Research Institute, Manhasset, NY.

email: wli@nslj-genetics.org

2. Cold Spring Harbor Lab, Cold Spring Harbor, NY

3. Bioinformatics Center Gatersleben Halle, Leibniz
Institute for Plant Genetics and Crop Plant Research,
Gatersleben, Germany.

email: grosse@ipk-gatersleben.de

Purpose of Discriminant Microarray Analysis

Identify genes that could distinguish different phenotypes, affection status, source of tissues, etc. These are “ **differentially expressed**” genes. The presence of differentially expressed genes make **classification, discrimination, prediction** of phenotypes,... possible. Genes can also be considered **jointly** to make better classifications.

Many many classifiers that use combined expressions of multiple genes

- Bayes' rule, K-nearest-neighbor,...
- multivariate logistic regression, Fisher's linear discrimination, support vector machine,...
- Decision/classification tree, recursive partitioning,...
- Neural networks, ...
- Model averaging, committee machine, bagging, boosting...

Many ACTUAL discriminant microarray analyses are done by only ONE gene. WHY?

- It is easier to explain the biology of one gene
- the concept of linear combinations of gene expressions (e.g. PCA) is confusing to biologists
- Quite often, one or a few genes can classify a majority of samples (e.g., Li and Yang, CAMDA'2000).

Back to univariate (single-gene) classifiers

- genes selected by fold-change
 - genes selected by t -test
- univariate logistic regression

Fold Change

definition: $R = \mu_1 / \mu_0 = \frac{\sum_{i \in (y_i=1)} x_i / N_1}{\sum_{j \in (y_j=0)} x_j / N_0}$

genes with “large enough” or “small enough” R 's are “good” classifiers

pro: simple. easy to explain.

con: statistically not sound. R obtained from a smaller dataset is less reliable than the same value from a larger dataset.

***t*-test**

definition: t-statistic, $t = \frac{\mu_1 - \mu_0}{\sqrt{s_1^2/N_1 + s_0^2/N_0}}$

pro: sample size effected is considered

con: the validity depends on whether normal distribution is a good description of sample values in both groups. only in normal distribution (or other “nice” distributions), can we use the means μ_1, μ_0 to summarize the whole set of data.

Univariate Logistic Regression

definition: for each sample i , gene j ,

$$L_i = P(y_i = 1 | x_{ij}) = \frac{1}{1 + e^{-a - b_j x_{ij}}}$$

estimate a , b value from the data. classifier performance can

be measured by the likelihood: $\hat{L} = \prod_i \hat{L}_i^{y_i} (1 - \hat{L}_i)^{1 - y_i}$

pro: no normal distribution is assumed. an extreme data can affect μ_1, μ_0 and a t -test result. but for logistic regression, as long as these extreme data are on the correct side, it's not a problem.

con: unreliable if the model used is "inappropriate".

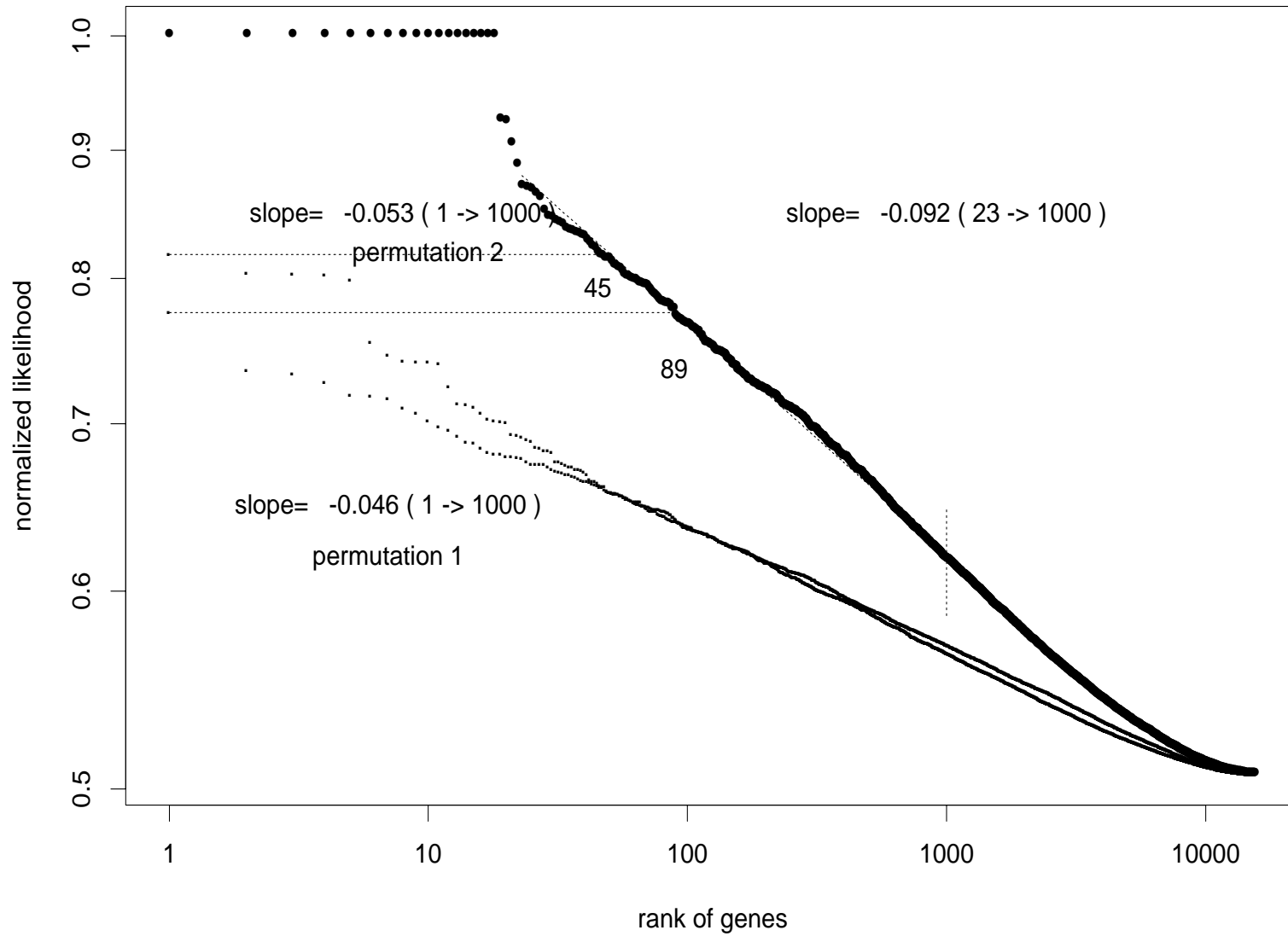
For any univariate classifier, a selection criterion is used

- **fold-change:** arbitrary. e.g. $F > 2$ (or $F < 1/2$)
- ***t*-test:** standard criteria, e.g. p -value $< 0.05, 0.01, \dots$. to some extent, it's also arbitrary. many publications on multiple-testing, Bonferroni correction, false discover rate,..
- **logistic regression:** ?

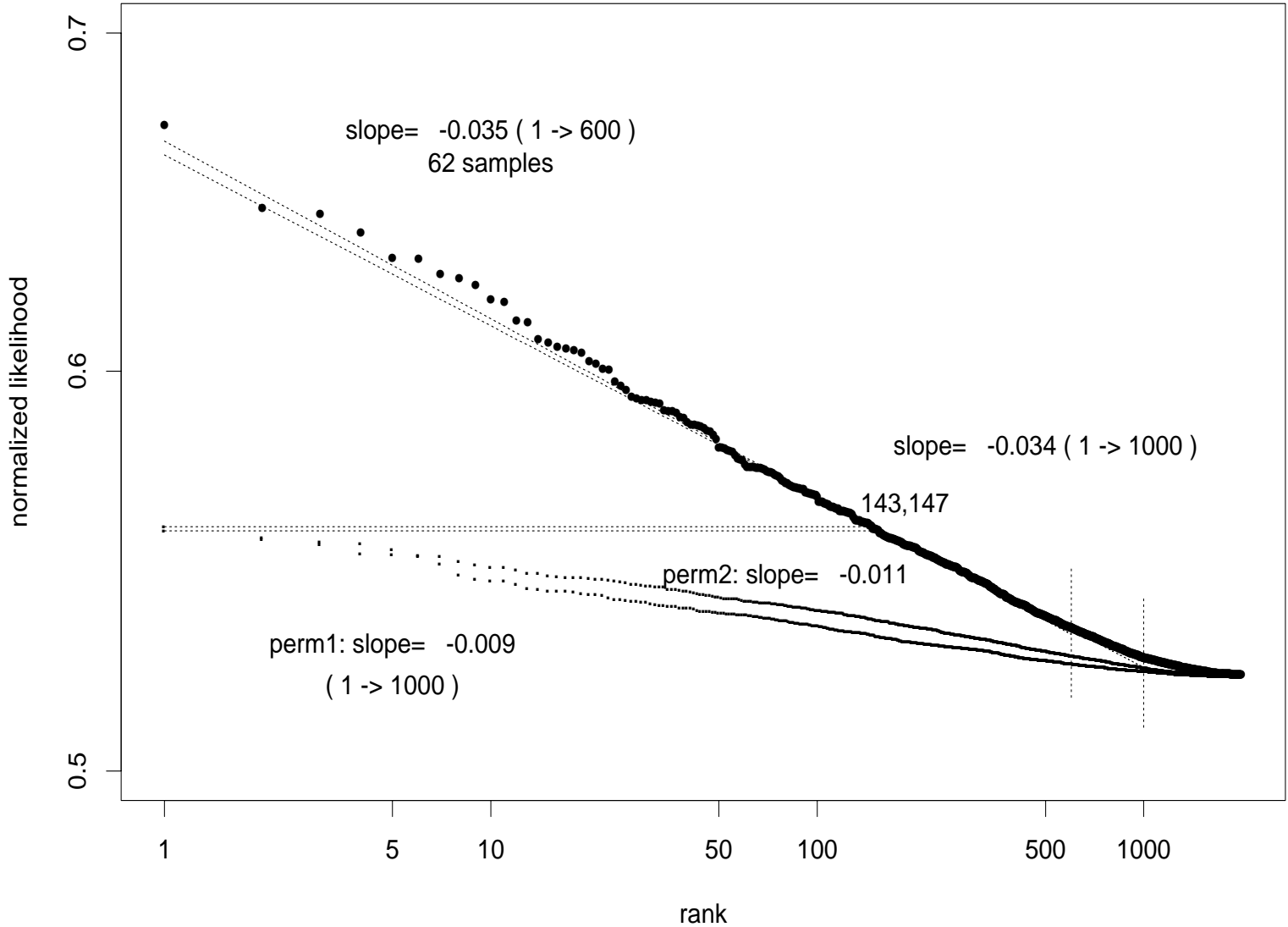
For logistic regression, a conceptually simple, but computationally expensive method for classifier (gene) selection is to scramble the label y_i and repeat the same logistic regression analysis.

For each permutation, N likelihoods for N genes are obtained: $\hat{L}_1, \hat{L}_2, \dots, \hat{L}_N$. These likelihoods can be ranked in descending order: $\hat{L}_{(1)} \geq \hat{L}_{(2)} \geq \dots \geq \hat{L}_{(N)}$

N+S vs. I+M classification (N=19)



classifying colon cancer vs normal



- For both the real data and permuted data, sorted (ranked) likelihoods $\hat{L}_{(r)}$ decrease with rank r as a power-law (algebraic) function: $\hat{L}_{(r)} = c/r^\alpha$ [**because** $\log(\hat{L})$ **vs.** $\log(r)$ **is a straight line**]
- It is called a **Zipf's law** [more accurately, generalized Zipf's law, because the slope is not -1] (Li and Yang, J. Theo. Biol, 2002)
- But the slopes and the top likelihoods are different between the real and permuted data. $(\hat{L}_{(1)}(\text{real}) \gg (\hat{L}_{(1)}(\text{null}))$

The Purpose of This Work

Obtain a (semi) analytical estimation of the extreme value distribution of $\hat{L}_{(1)}(\text{null})$, its mean $E[\hat{L}_{(1)}(\text{null})]$, and use it as the threshold for selecting classifiers (genes).

why use rank-1: less arbitrary

why adopt a (semi) analytical approach: saving tremendous computing time

notations:

D_0 - permuted data/scrambled data/null data;

M - (logistic regression) model with parameters estimated from the data;

M_0 - a simpler (trivial) model

The key trick to the solution of this problem is to use a trivial model whose likelihood can be written easily, then relate the maximum-likelihood of the logistic regression model and that of the trivial model by the likelihood-ratio test.

We choose the “proportion guess” as the null model M_0 :

$P(y_i = 1|x_{ij}) = c$, with one parameter (c)

the parameter value is estimated as $\hat{c} = N_1/N$

Maximum-likelihood of this model is

$$\hat{L}(D_0|M_0) = \hat{c}^{N_1}(1 - \hat{c})^{N-N_1}$$

log-max-likelihood is related to entropy

$$H = -N_1/N \log(N_1/N) - (N - N_1) \log((N - N_1)/N)$$

$$\text{by: } \log \hat{L}(D_0|M_0) = -NH$$

the “random guess” is: $P(y_i = 1|x_{ij}) = 0.5$

(2-log-max) Likelihood-ratio test:

using two models on the same dataset. the ratio of the two max likelihoods of the two models, in the asymptotic limit, is:

$$2 \log \frac{\hat{L}_j(D_0|M)}{\hat{L}_j(D_0|M_0)} \sim \chi_{df}^2 \quad df = d(M) - d(M_0)$$

or

$$\log \hat{L}_j(D_0|M) - \log \hat{L}_j(D_0|M_0) = \frac{t(\chi^2)}{2} + \dots$$

and

$$\max_j \log \hat{L}_j(D_0|M) = -NH + \max(t_1, t_2, \dots, t_p)/2 + \dots$$

the left-hand-side of the equation is exactly what we need!

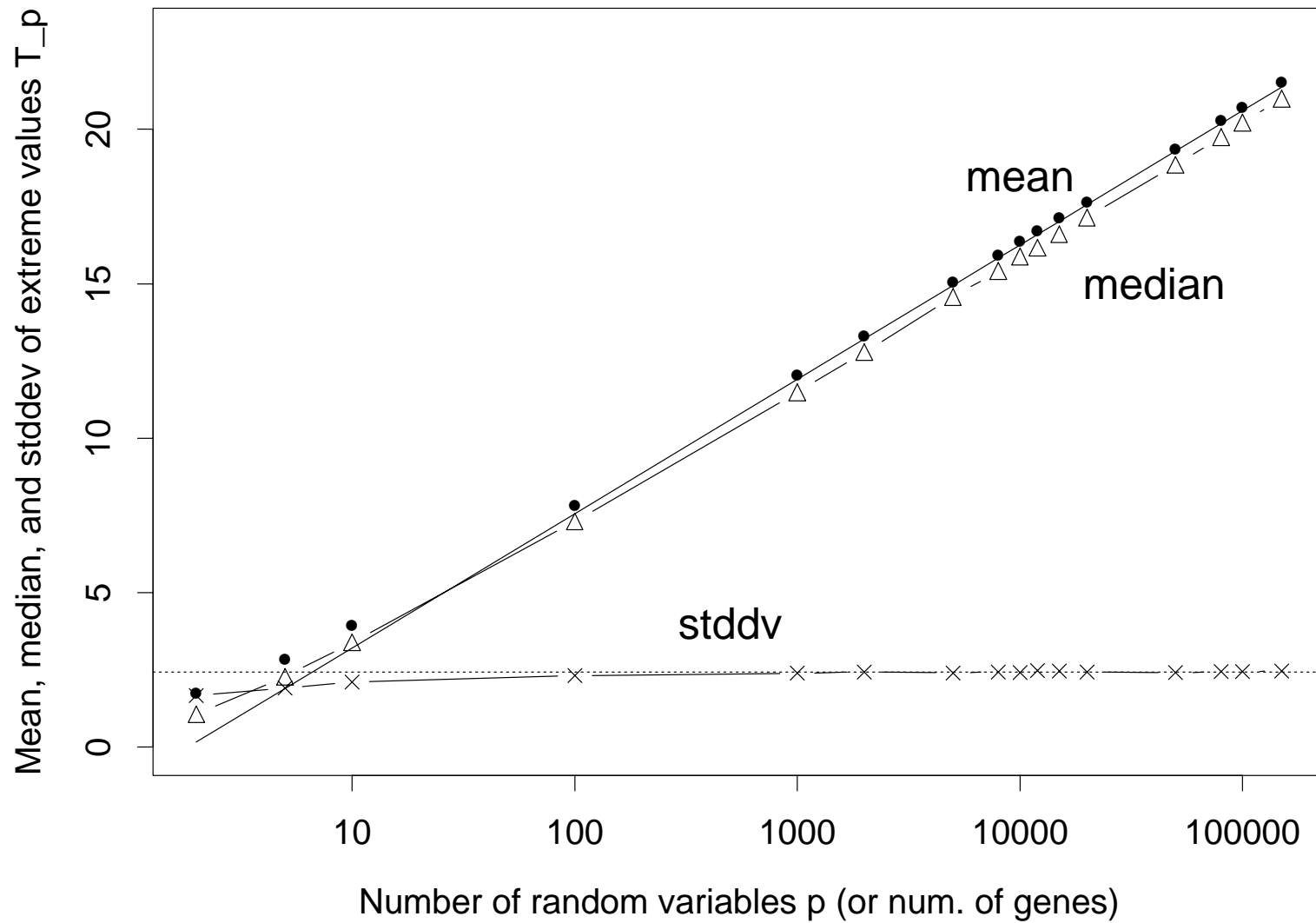
Maximum Value (Extreme Value) of p Values Sampled from the χ^2 Distribution

$$T_p \equiv \max(t_1(\chi^2), t_2(\chi^2), \dots, t_p(\chi^2))$$

$$P(T_p)? E[T_p]? \text{Median}[T_p]? \sigma^2[T_p]?$$

we rely on a numerical calculation (semi-analytical!).

Extreme values of p chi-square distributed samples



Maximum Value of p Values Sampled from the χ^2 Distribution (cont.)

With our range of p values, the mean of the extreme value increases with $\log(p)$ **linearly**:

$$E[T_p] \approx -1.1385 + 1.8881 \log(p)$$

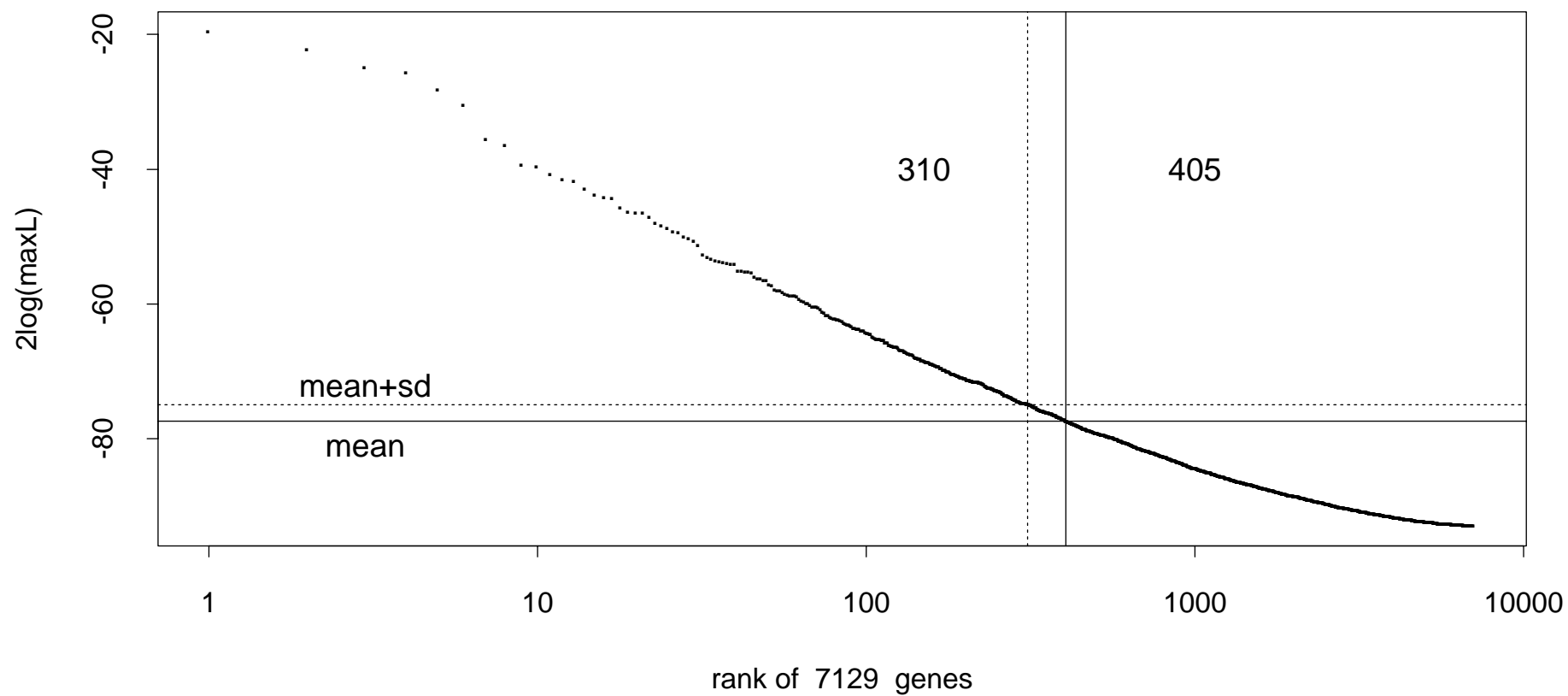
if p values are sampled from the exponential distribution, there is an exact solution: $E[\max(t_1, t_2, \dots, t_p)] = 0.5772 + \log(p)$

Our Gene Selection Criterion for Univariate Logistic Regression

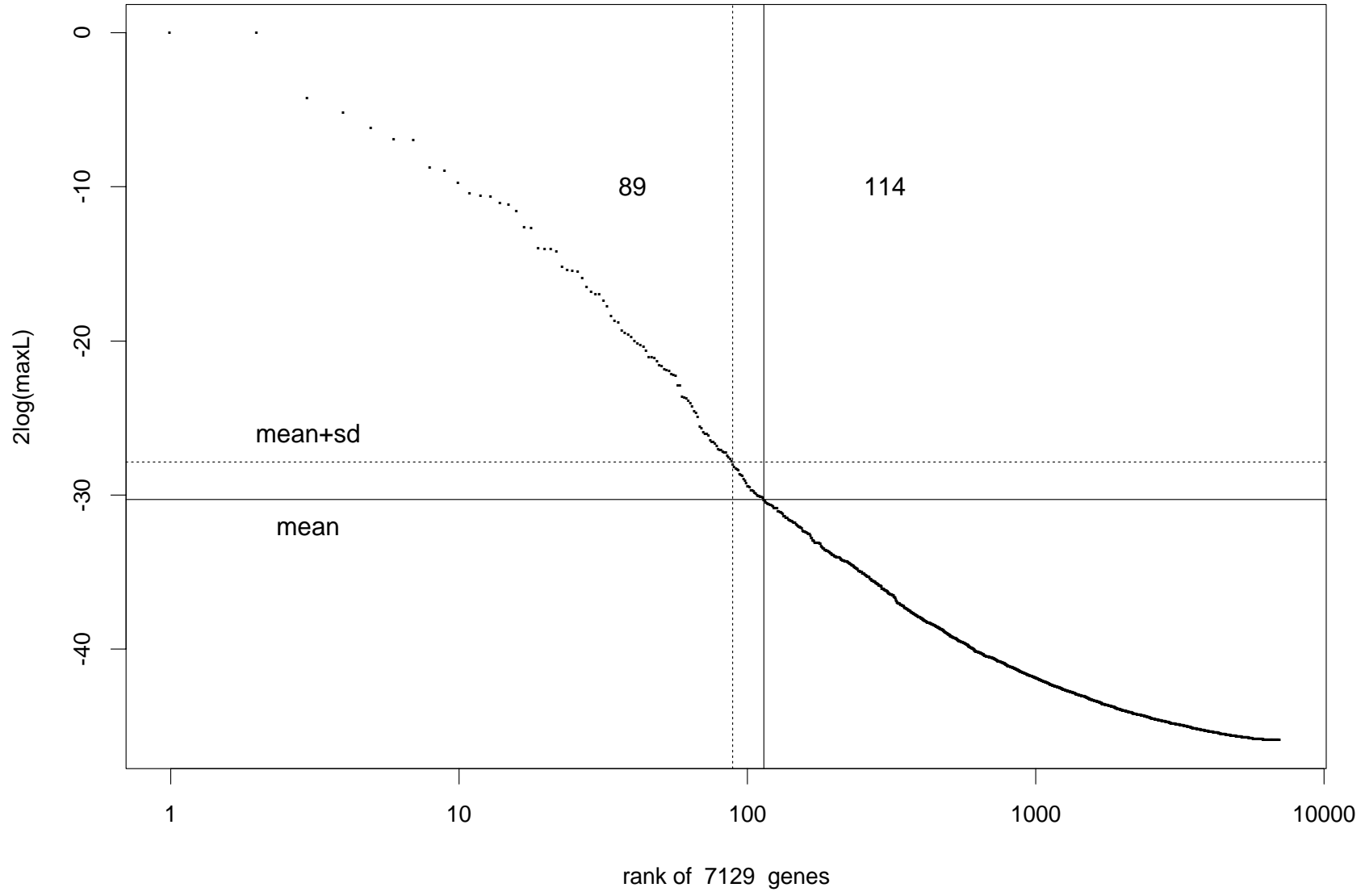
Any gene whose (2-log-max) likelihood is larger than:

$$2 \log \hat{L}(D|M) > -2NH - 1.1385 + 1.8881 \log(p)$$

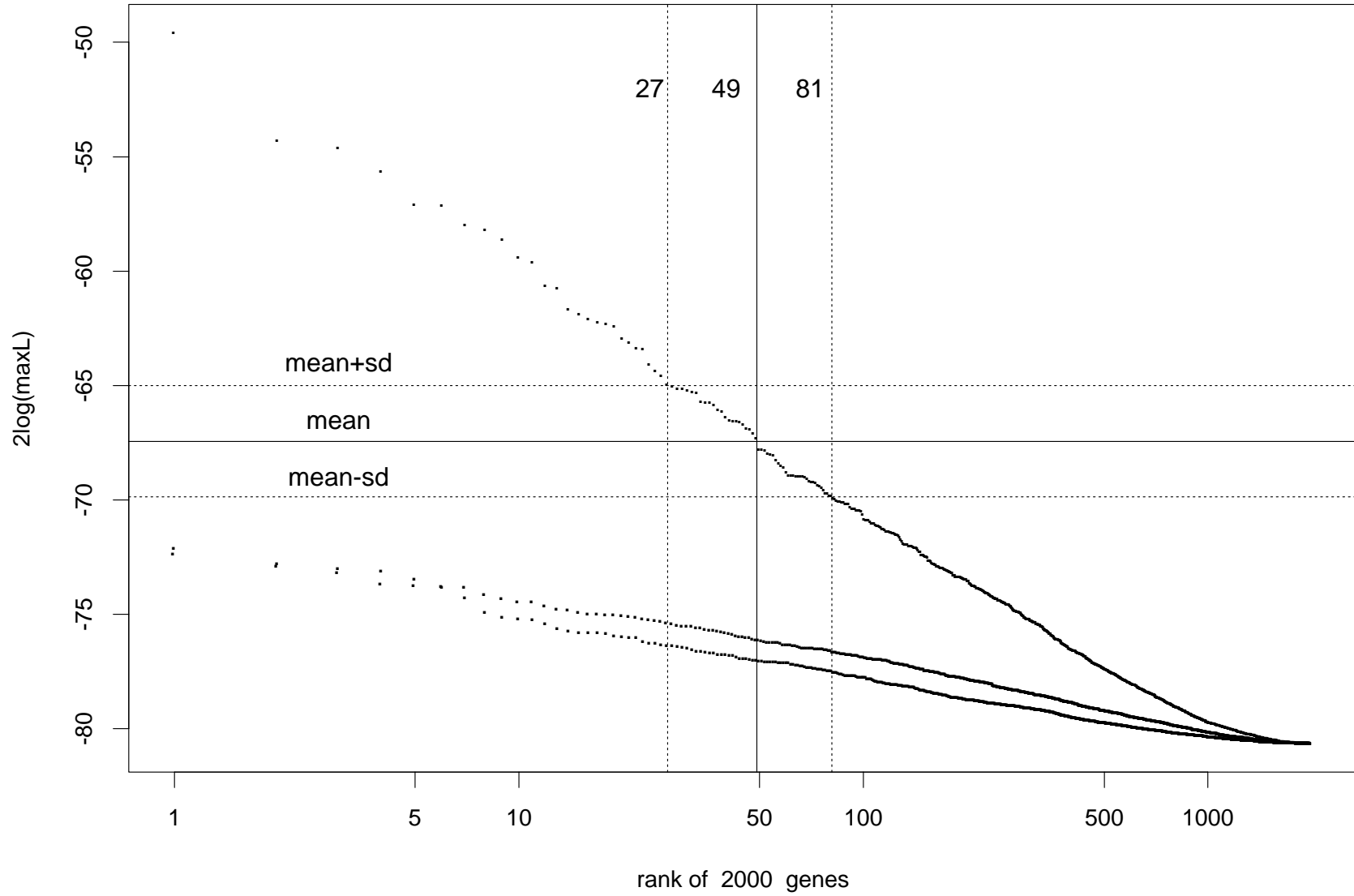
leukemia: ALL vs AML(N=72)



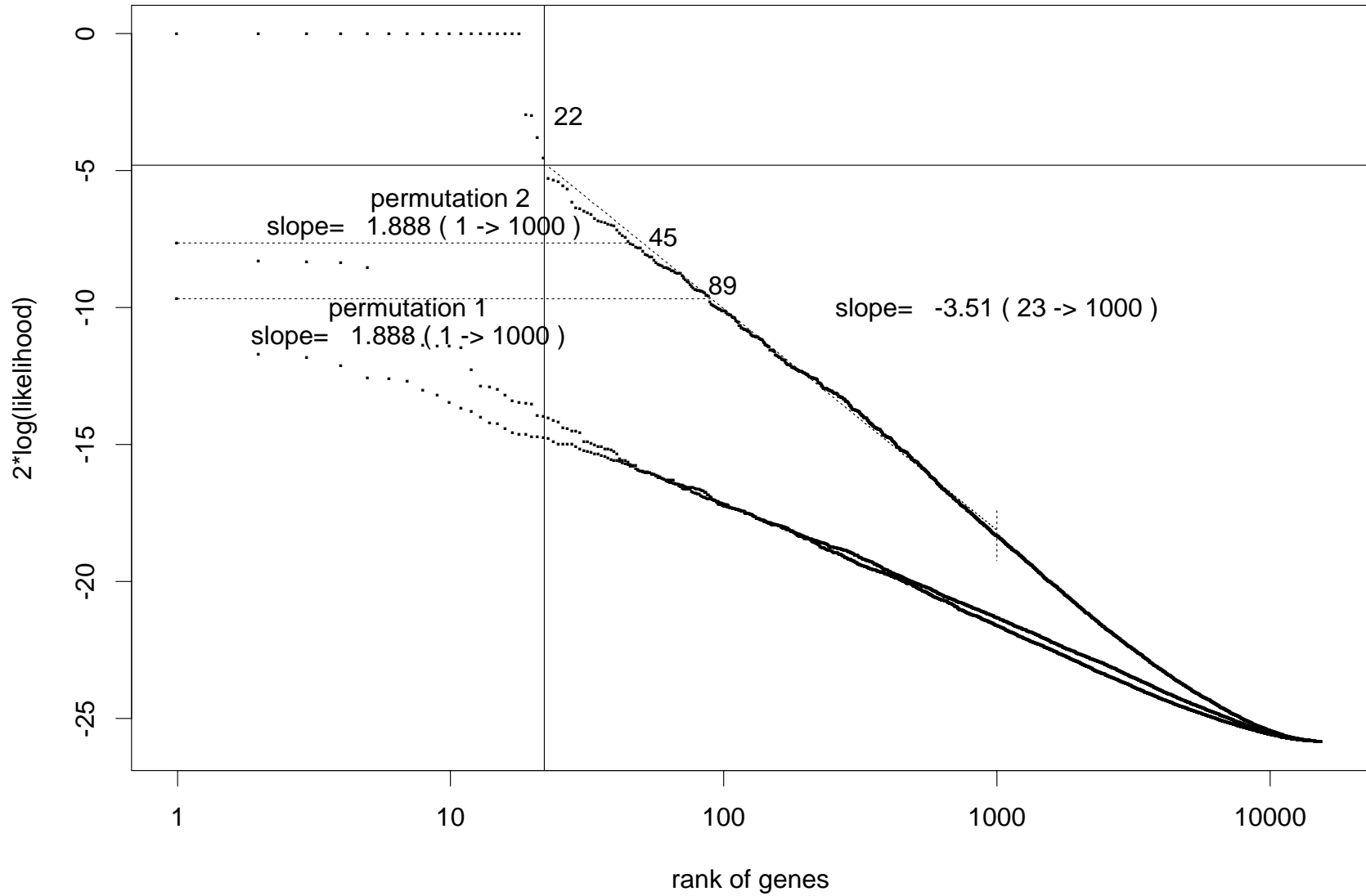
ALL T vs B-cell(N=47)



colon cancer vs normal(N=62)



N+S vs. I+M classification (N=19)



Conclusions

We have derived a semi-analytical equation of the mean of the extreme (top) likelihood value for the null data. This equation can be used to select genes in model-based discriminant microarray analysis (such as logistic regression).

This equation is an approximation, and partially based on numerical simulation of χ^2 distributed random values. Further tests on how good this equation is are needed in the future.

References

- **[on much fewer genes are needed than apparent numbers]**

W Li, Y Yang (2002), “How many genes are needed for a discriminant microarray data analysis”, in *Methods of Microarray Data Analysis. Papers from CAMDA'00*, eds SM Lin and KF Johnson (Kluwer Academic), 137-150

- **[on power-law decay of max-likelihood (Zipf's law)]**

W Li, Y Yang (2002), “Zipf's law in importance of genes for cancer classification using microarray data”, *Journal of Theoretical Biology*, 219(4):539-551.

- **[an application to bladder cancer data]**

M Sanchez-Carbayo, N Socci, JJ Lozano, W Li, T Belbin, M Preystowski, A Ortiz, G Childs, C Cordon-Cardo (2003), “Gene discovery in bladder cancer progression using cDNA microarrays”, *American Journal of Pathology*, to appear.