

How Many Genes Are Needed for a Discriminant Microarray Data Analysis?

Wentian Li and Yaning Yang

Lab of Statistical Genetics
Rockefeller University

contact information:
wli@linkage.rockefeller.edu
<http://linkage.rockefeller.edu/wli/>

In Golub et al. paper, 50 genes are used to classify cancer types on 38 samples.

At a first glance, the number of parameters should be smaller than the number of sample points, otherwise the solution is not unique.

To address this issue, we need to distinguish **model selection** and **model averaging**.

This issue determines the answer to the questions on how many genes are needed for a discriminant analysis.

MODEL SELECTION

We have a series of possible models M_1, M_2, \dots with p_1, p_2, \dots parameters respectively.

The naive maximum likelihood framework would select a model that leads to the best fit of the available data.

This naive solution is not fair since models with more parameters tend to fit a data set better than models with fewer parameters. A penalty on the model complexity is needed.

fitting the data first, penalizing model complexity later

1. data-fitting: adjusting the parameter values to maximize the likelihood $L = P(D|M)$ (D for data, M for model):

$$\hat{L} = L(D|\hat{\theta}) = \max_{\theta} L(D|\theta, M)$$

2. model-penalizing: log-likelihood subtracts a term proportional to the number of parameters in the model (K):

$$\log(\hat{L}) - \alpha K$$

$\alpha = 2$ (Akaike) or $\alpha = \log(N)$ (Bayesian) (N is the sample size)

3. change the sign, and multiplied by two:

$$AIC = -2 \log(\hat{L}) + 2K$$

$$BIC = -2 \log(\hat{L}) + \log(N)K$$

Akaike information criterion and Bayesian information criterion

MODEL SELECTION BY MINIMIZING AIC OR BIC

1. comparing AIC and likelihood-ratio test (LRT): LRT is limited to comparison of two nested models, whereas AIC is more general.

If degrees of freedom difference of two models is 1, AIC model selection is equivalent to LRT with the significance level of 0.157. But when the difference of degrees of freedom of two models is large, the corresponding significance level is much smaller.

2. comparing BIC and AIC: BIC selects models simpler than those selected by AIC. In the large sample limit $N \rightarrow \infty$, both LRT and AIC are not “dimensionally consistent” (false positive rate remains non-zero). BIC, on the other hand, is “dimensional consistent”: a model is selected correctly in $N \rightarrow \infty$ limit.

MODEL SELECTION IN DISCRIMINANT MICROARRAY DATA ANALYSIS

1. data: $(x_1, x_2, \dots, x_{7129}, y)$, where y is the cancer type (0,1), x 's are the (log) mRNA expression levels.
2. select a model M for discriminant analysis ($y = M(x_1, x_2, \dots, x_p)$ with $p < 7129$), such as logistic regression:

$$p(y = 1) = \frac{1}{1 + e^{-a_0 - \sum_i a_i x_i}}$$

3. adjusting $p+1$ parameters a_0, a_1, \dots, a_p to best fit the data (maximizing likelihood).
4. penalizing models with more parameter (minimizing AIC/BIC): $K=p+1$. **number of genes included is determined by the optimal number of parameters include in the model by model selection criterion.**

a schematic result

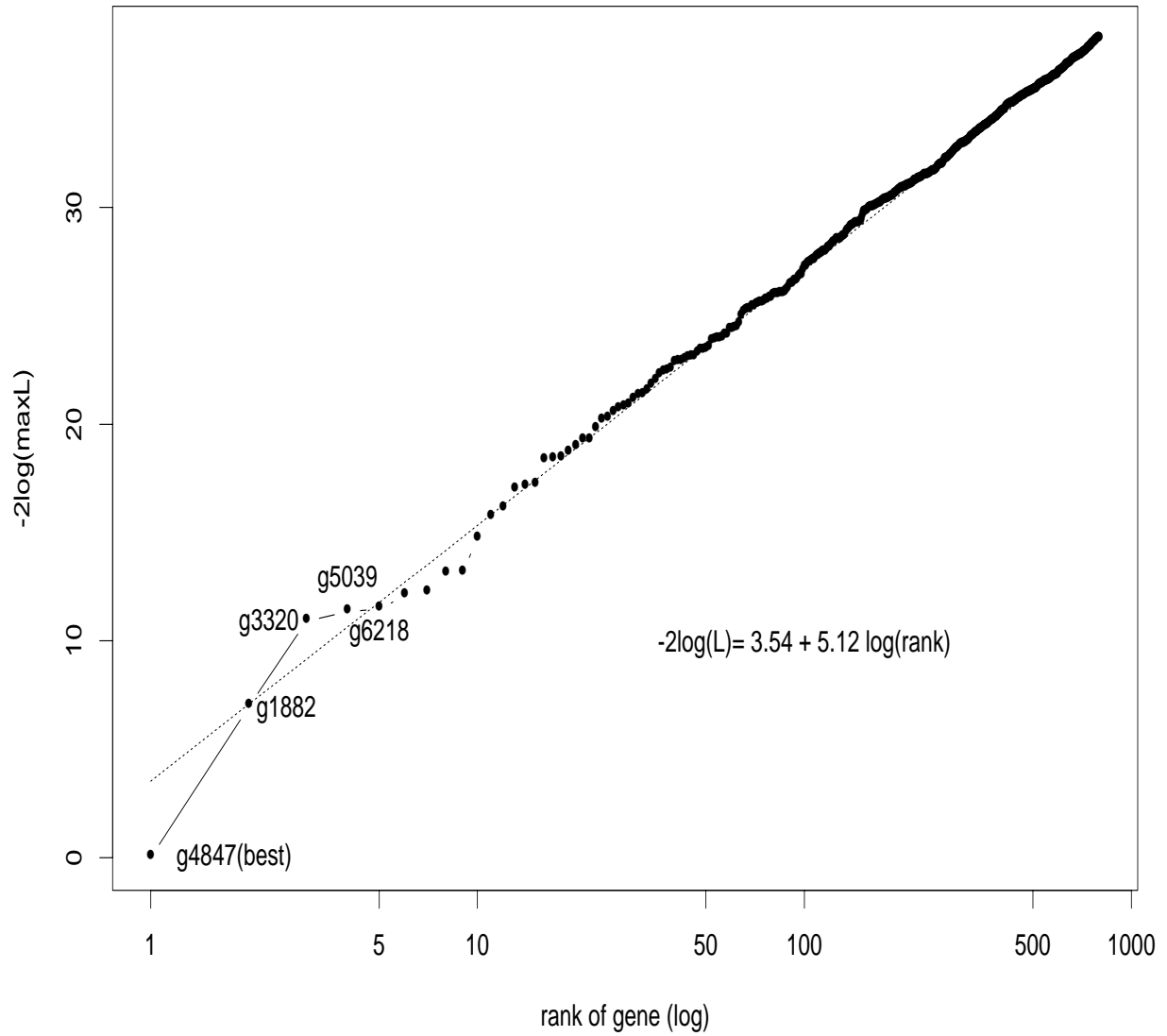
model	K	$-2 \log(L)$	AIC/BIC
g1	2		
g7000	2		
g1+g3	3		
g1+g3 +g1*g3	4		
g1+g3+g56	4		
best model			(lowest)
some genes	$2 \leq K \leq 38$		
all genes	7130	(impossible)	
no gene (null)			
fixed prob	1		
random guess	0	(highest)	

It is impossible to include all 7129 genes in one model, because the sample size is 38.

A two-step strategy is usually adopted: single-variable logistic regression is carried out for all 7129 genes, then the top 37 genes are selected. Further selection of genes is carried out by a stepwise variable selection.

Ranking of single-variable logistic regression

$-2\log(L)$ of each gene on training set



We can achieve perfect fitting/prediction on the training set with 37 genes, 22 genes, 10 genes, 5 genes, ... and 2 genes.
 $(-2 \log(\hat{L}) \rightarrow 0)$.

type	K	$-2\log(\hat{L})$	AIC	Δ AIC
g1834+g2267	3	0.004	6.004	2.004
g5039+g5772	3	0.008	6.008	2.008
sum of top 2 (g4847+g1882)	3	0.029	6.029	2.029
sum of top 5	6	0.011	12.011	8.011
sum of top 10	11	0.002	22.002	18.002
sum of top 22	23	0.001	46.001	42.001
sum of top 37	38	0.001	76.001	72.001
fixed prob	1	45.728	47.728	43.728
random guess	0	52.679	52.679	48.679

type	BIC	Δ BIC	p_{train}	p_{test}
g1834+g2267	10.917	3.642	38/38	22/34
g5039+g5772	10.921	3.646	38/38	26/34
sum of top 2 (g4847+g1882)	10.942	3.667	38/38	32/34
sum of top 5	21.837	14.562	38/38	24/34
sum of top 10	40.016	32.741	38/38	31/34
sum of top 22	83.666	76.391	38/38	27/34
sum of top 37	138.229	130.954	38/38	21/34
fixed prob	49.365	42.090	27/38	20/34
random guess	52.679	45.404	19/38	17/34

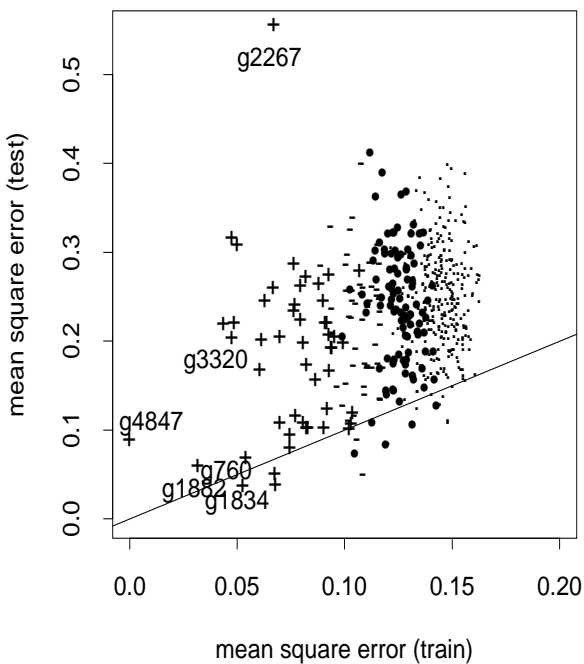
one single-variable model also achieves prediction (gene 4847, zyxin)

type	K	$-2\log(\hat{L})$	AIC	Δ AIC
#1 g4847 (zyxin)	2	≈ 0	4.000	0
#2 g1882 (CST3 cystatin C)	2	6.973	10.973	6.973
#3 g3320 (leukotriene..)	2	10.914	14.914	10.914
#4 g5039 (LEPR leptin rec)	2	11.355	15.355	11.355
#5 g6218 (ELA2 elastatse 2)	2	11.459	15.459	11.459
#6 g2020 (FAH ..)	2	12.103	16.103	12.103
#7 g1834 (CD33 antigen)	2	12.226	16.226	12.226
#8 g760 (cystatin A)	2	13.104	17.104	13.104
#9 g1745 (LYN v-yes-1..)	2	13.151	17.151	13.15
#10 g5772 (c-myb)	2	14.723	18.723	14.723
#100 g2833(AF1q)	2	27.215	31.215	27.21
#200 g3312(ATR)	2	30.841	34.841	30.841
fixed prob	1	45.728	47.728	43.728

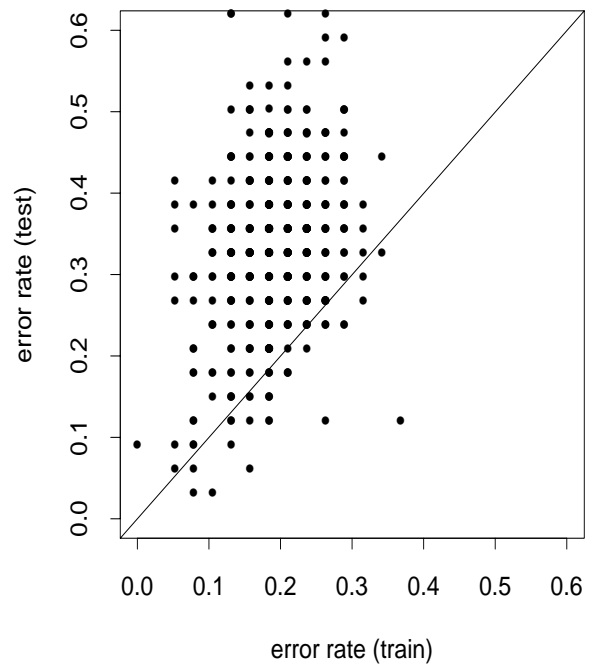
type	BIC	Δ BIC	p_{train}	p_{test}
#1 g4847 (zyxin)	7.275	0	38/38	31/34
#2 g1882 (CST3 cystatin C)	14.248	6.973	36/38	32/34
#3 g3320 (leukotriene..)	18.190	10.915	35/38	27/34
#4 g5039 (LEPR leptin rec)	18.630	11.355	36/38	22/34
#5 g6218 (ELA2 elastatse 2)	18.734	11.459	34/38	22/34
#6 g2020 (FAH ..)	19.378	12.103	36/38	25/34
#7 g1834 (CD33 antigen)	19.501	12.226	35/38	31/34
#8 g760 (cystatin A)	20.379	13.104	35/38	32/34
#9 g1745 (LYN v-yes-1..)	20.426	13.151	33/38	28/34
#10 g5772 (c-myb)	21.998	14.723	35/38	27/34
#100 g2833(AF1q)	34.490	27.215	30/38	28/34
#200 g3312(ATR)	38.117	30.842	29/38	21/34
fixed prob	49.365	42.090	27/38	20/34

what about the testing set?

test vs train (mean square error)



test vs train (0/1 error)



$$\text{mean squared error} = (1/N) \sum_i^N (P(y|x_i, \hat{\theta}) - y_i)^2$$

$$\text{prediction (0/1) error} = (1/N) \sum_i^N \mathbf{I}(|P(y|x_i, \hat{\theta}) - y_i| > 0.5)$$

How many genes are needed for a discrimination in this data set?

Answer 1:

Within the model selection framework – one model is used for discrimination, while (log) mRNA expressions are added – **one gene is enough.**

Two-gene models are doing quite well also.

MODEL AVERAGING

With the available data (training set), a series of possible models M_1, M_2, \dots are evaluated. For a new data point (testing set), a prediction is made by (weighted) averaging predictions from different models.

Bayesian framework is ideal for describing model averaging (D for available data, \tilde{D} for new data):

$$\begin{aligned} p(\tilde{D}|D) &= \sum_M p(\tilde{D}|M)p(M|D) \\ &= \sum_M p(\tilde{D}|M) \cdot \frac{p(D|M)p(M)}{p(D)} \\ &\propto \sum_M p(\tilde{D}|M) \cdot e^{-BIC/2} \end{aligned}$$

$\exp(-BIC/2)$ is proportional to the maximum likelihood if all models have the same number of parameters.

MODEL AVERAGING IN DISCRIMINANT MICROARRAY DATA ANALYSIS

1. models are single-variable logistic regression.

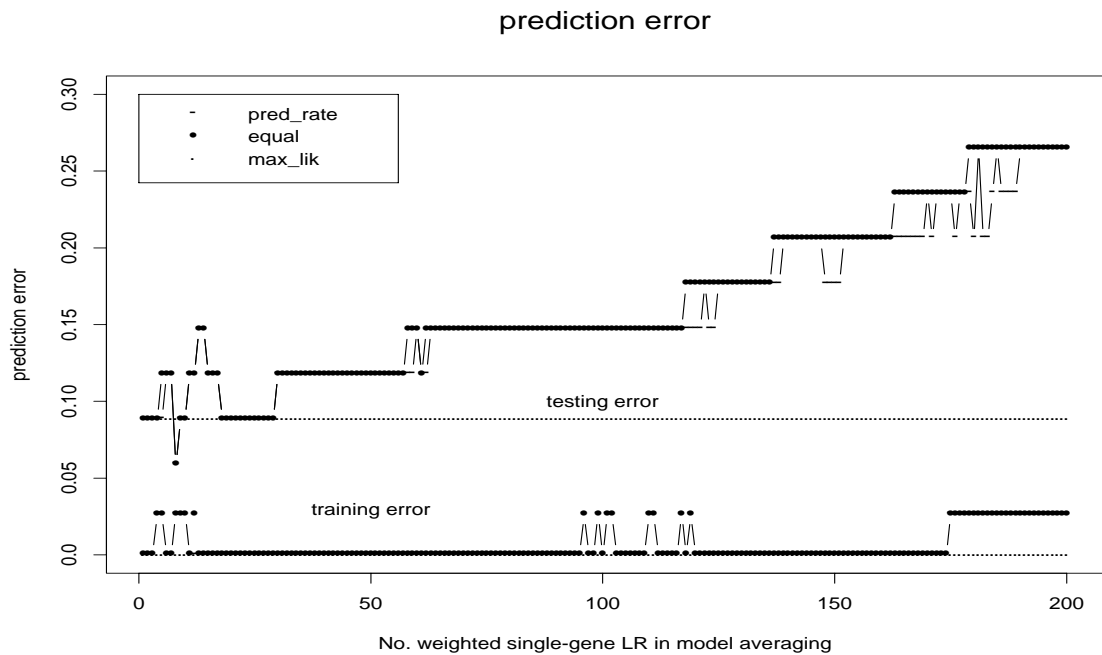
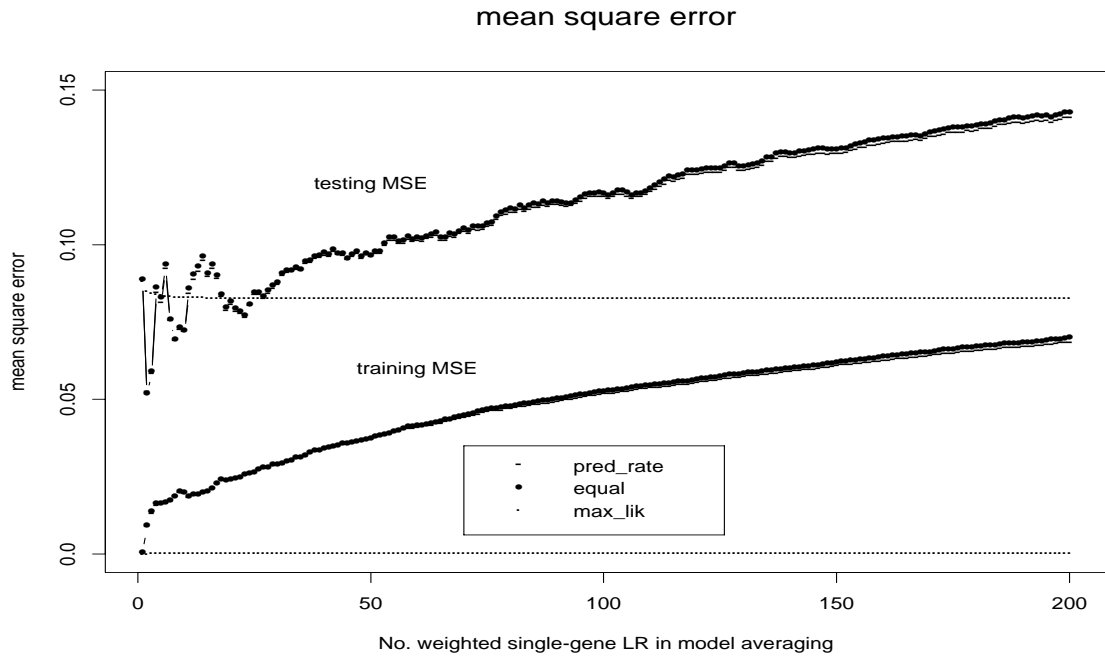
$$p(y = 1)_j = \frac{1}{1 + e^{-a_0 - a_j x_j}}$$

2. these models are averaged:

$$p(y = 1) = \sum_j w_j \left(\frac{1}{1 + e^{-a_0 - a_j x_j}} \right)$$

3. three different weighting schemes are tried: (1) maximized likelihood (same as $\exp(-BIC/2)$ and $\exp(-AIC/2)$ here); (2) prediction rate on the training set (prediction confidence information is removed); (3) equal weight.

With the first weighting scheme, error does not change much with the number of models included. On the other hand, with the second and the third weighting scheme, error increases.



$$e^{-BIC/2} \text{ (or } \hat{L} \text{) weight}$$

The best model (g4847) is too good!

$-2 \log(\hat{L})$ of the top ten genes (logistic regression predictors) are: 0, 6.973, 10.914, 11.355, 11.459, 12.103, 12.226, 13.104, 13.151, 14.723, which lead to the relative weights w_j/w_1 of 1, 0.031, 0.0043, 0.0034, 0.0032, 0.0024, 0.0022, 0.0014, 0.0014, and 0.0006. w_1/w_2 is already 32.67.

Treating the dominant term as 1, the weighted number of terms (up to top 10) is only 1.05!

Effective number of genes used in this weighting scheme is less than 2.

p_{train} weight

For this data set, it might be possible for perform better on the testing set when the number of terms is less than ~ 25 .

equal weight

Similar to p_{train} weight, but worse. In general, it is not a good weighting scheme.

How many genes are needed for a discrimination in this data set?

Answer 2:

Within the model averaging framework – predictions from many models are averaged – the answer depends on the weighting scheme, and depends on whether the apparent or effective number of genes is considered. **From ~ 2 to ~ 25.**

Beyond this data set

model selection:

$$0 + 2(p + 1) < AIC(null)$$

$$0 + \log(N)(p + 1) < BIC(null)$$

an upper limit on the number of genes included in a logistic regression can be obtained. (for this data set, if we use the “no gene” fixed prediction model as null, the upper limits obtained this way are 22 and 12.)

model averaging:

any number of single-gene models can be included, but the effective number of terms can be much smaller.

end