

# Robust and Accurate Cancer Classification with Gene Expression Profiling

---

(Computational Systems Biology, 2005)

Author: Haifeng Li, Keshu Zhang, Tao Jiang



# Outline

---

- Background
- LDA (linear discriminant analysis) and small sample size problem
- Previous methods and Generalized Linear Discriminant Analysis (GLDA)
- Experimental results
- Conclusion



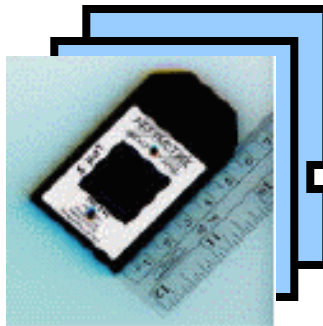
# Background

---

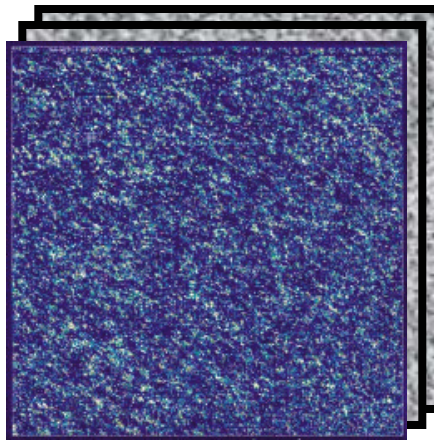
- Accurate diagnosis of human cancer is essential in cancer treatment. Gene expression profiling could be a precise and systematic method for cancer classification.
- Microarray technology enables us to simultaneously observe the expression levels of many thousands of genes on the transcription level.
- Microarray:
  - chip with a matrix of thousands of spots printed on to it
  - Each spot binds to a specific gene

# Background

Microarray chips

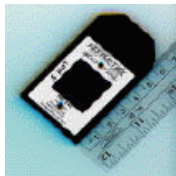


Images scanned by laser



Gene	Value
D26528_at	193
D26561_cds1_at	-70
D26561_cds2_at	144
D26561_cds3_at	33
D26579_at	318
D26598_at	1764
D26599_at	1537
D26600_at	1204
D28114_at	707

New sample



Prediction:  
Cancer or Normal

Classifiers

Datasets

	D26528	D63874	D63880	...
Sample1	193	4157	556	
Sample2	129	11557	476	
Sample3	44	12125	498	
Sample4	218	8484	1211	
Sample5	109	3537	131	
Sample6	106	4578	94	
Sample7	211	2431	209	
...				



# Background

---

- **Two Major Challenges :**  
high dimension of data (too many columns (genes), usually  $> 1,000$ )  
the number of samples is small. (too few samples, usually  $< 100$ )
- **“Peaking Phenomenon”:** a large number of features may degrade the performance of classifiers if the number of training samples is small relative to the number of features.
- It needs at least **5 – 10** times as many training samples per class as the number of features to obtain well-trained (robust) classifiers (A.K.Jain *et al.* 1982)
- Consequently, **dimensionality reduction** is essential to cancer classification.



# Dimension Reduction

---

- **Feature Selection:** choose a “best” subset of features from a large initial set.
- **Advantage:** the selected features retain their original biological interpretation.
- **Method:** Single-gene-ranking, gene-pair-ranking, GA/KNN, SVM with RFE (recursive feature elimination)
- **Disadvantage:** At least 50 features need be chosen in general. This number is far from the 5-10 times ratio of samples to features.



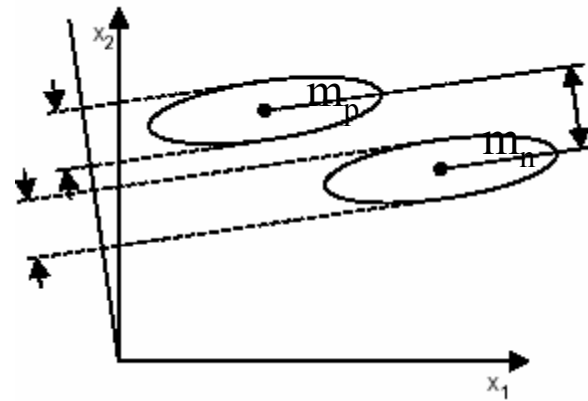
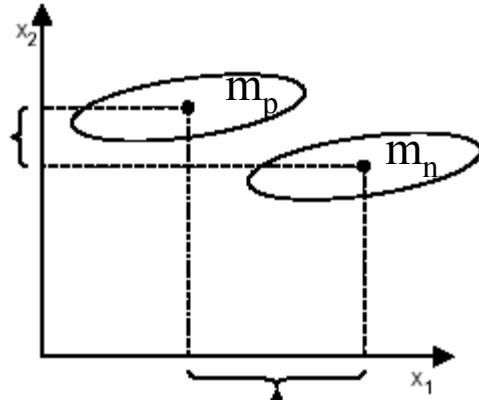
# Dimension Reduction

---

- **Feature Extraction:** transform the original features space into a reduced feature space.
- **Advantage:** provides a better discriminative ability than feature selection; meets the recommended 5-10 times ratio of samples to features per class.
- **Disadvantage:** the new features generated by feature extraction may not have a clear biological meaning.

# LDA (Linear Discriminant Analysis)

- What kind of projection is desirable ?  
(two dimensional and two classes example)



- Intuition:  
After projection, the points from same classes are clustered while points from different classes are as far away as possible.  
(The distance between the projected sample and the corresponding projected mean is as short as possible and the distance between the projected means is as long as possible)

# Linear Discriminant Analysis- two classes

- Projection:  $y = W^T x$
- Between-Class Scatter Matrix

$$S_b = (m_p - m_n)(m_p - m_n)^T$$

$$S_b^{(y)} = W^T S_b W$$

- Within-Class Scatter Matrix

$$S_w = \sum_{x_i \in P} (x_i - m_p)(x_i - m_p)^T + \sum_{x_j \in N} (x_j - m_n)(x_j - m_n)^T$$

$$S_w^{(y)} = W^T S_w W$$

- Total Scatter Matrix

$$S_t = \sum_{x_i \in P+N} (x_i - m)(x_i - m)^T = S_b + S_w$$



# Linear Discriminant Analysis

---

- In the projected low dimensional space we try to maximize:

$$\frac{|S_b^{(y)}|}{|S_w^{(y)}|} = \frac{|W^T S_b W|}{|W^T S_w W|}$$

- Solution for the projection:

$$W_{opt} = \arg \max_W \frac{|W^T S_b W|}{|W^T S_w W|}$$

- $W$  is the eigenvectors with the largest eigenvalues for

$$S_w^{-1} S_b$$

# Linear Discriminant Analysis- multiple classes

- Extension of two-class LDA (Fisher Discriminant Analysis) to Multi-Class Case with assumption of  $C$  classes.

- Formulas have to be rewritten

1) Between-Class Scatter Matrix

$$S_b = \sum_{i=1}^C P_i (m_i - m)(m_i - m)^T$$

2) Within-Class Scatter Matrix

$$S_w = \sum_{i=1}^C P_i \sum_{x_i \in \text{Class}_i} (x_i - m_i)(x_i - m_i)^T$$

3) Target function

$$W_{opt} = \arg \max_W \frac{|W^T S_B W|}{|W^T S_W W|}$$

- $W$  is the eigenvectors with the largest eigenvalues for  $S_w^{-1} S_b$



# Small Sample Size Problem with LDA:

---

- Small Sample Size Problem:
  - The number of samples ( $n$ ) is smaller than the dimensionality of samples ( $D$ ).
- $\text{Rank}(S_w) < D$ . So  $S_w$  is singular and  $S_w^{-1}$  doesn't exist.
- We can not get the optimal  $W$ , which is the eigenvectors with the largest eigenvalue for  $S_w^{-1}S_b$



# Previous methods

---

- 1) First reduce the dimensionality with some other feature selection/extraction method and then apply LDA on the dimensionality-reduced data.
- 2) Pseudo-inverse  $S_w^+$ : replace  $S_w^{-1}$  with  $S_w^+$ .  $W$  is the largest eigenvectors of  $S_w^+ S_b$ .
- 3) Regularization: add small quantities to the diagonal elements of  $S_w$  to make it nonsingular.

For example: PDA (Penalized discriminant analysis).  $S_w$  is replaced with  $S_w + \lambda \Omega$  and then LDA proceeds as usual, where  $\Omega$  is a symmetric and nonnegative definite penalty matrix. The choice of  $\Omega$  depends on the problem.

# Generalized Linear Discriminant Analysis (GLDA)

- **Theorem:** The column vectors of  $W$  are the eigenvectors of  $S_t^{-1} S_b$
- **Definition 1 (Moore-Penrose Inverse)** : A matrix  $A^+$  satisfying the following conditions is unique and is called the Moore-Penrose inverse of  $A$ :

$$\begin{aligned} AA^+A &= A & A^+AA^+ &= A^+ \\ (A^+A)^T &= A^+A & (AA^+)^T &= AA^+ \end{aligned}$$

- **Lemma 2:**  $S_t S_t^+ (x - m) = x - m$
- **Lemma 3:**  $S_t^{-1} S_t^{-1} S_w = S_w$
- **Lemma 4:**  $S_t^{-1} S_t^{-1} S_w = S_w$

$$J(W) = \text{tr} \left( \frac{W^T S_b W}{W^T S_t W} \right) = \text{tr} \left( (W^T S_t W)^{-1} (W^T S_b W) \right)$$

- Diagonalize symmetric matrices  $W^T S_b W$  to  $\Lambda$  and  $W^T S_t W$  to  $I$   
 $P^T (W^T S_b W) P = \Lambda$  ,  $P^T (W^T S_t W) P = I$  where  $P$  is a  $d \times d$  nonsingular matrix.
- $J(W) = J(WP)$



# Generalized Linear Discriminant Analysis (GLDA)

- Denoting  $\mathbf{K} = \mathbf{S}_t \mathbf{W} \mathbf{P}$ ,  $\mathbf{S}_b \mathbf{S}_t^+ \mathbf{K} = \mathbf{K} \mathbf{A}$

$$\begin{aligned} J(\mathbf{S}_t^+ \mathbf{K}) &= J(\mathbf{S}_t^+ \mathbf{S}_t \mathbf{W} \mathbf{P}) = J(\mathbf{S}_t^+ \mathbf{S}_t \mathbf{W}) \\ &= \text{tr}((\mathbf{W}^T \mathbf{S}_t \mathbf{S}_t^+ \mathbf{S}_t \mathbf{S}_t^+ \mathbf{S}_t \mathbf{W})^{-1} (\mathbf{W}^T \mathbf{S}_t \mathbf{S}_t^+ \mathbf{S}_b \mathbf{S}_t^+ \mathbf{S}_t \mathbf{W})) \\ &= \text{tr}((\mathbf{W}^T \mathbf{S}_t \mathbf{W})^{-1} (\mathbf{W}^T \mathbf{S}_b \mathbf{W})) = J(\mathbf{W}) \end{aligned}$$

In other words, the optimal transformation  $\mathbf{W}$  is

$$\mathbf{W} = \mathbf{S}_t^+ \mathbf{K}$$

$$\mathbf{S}_t^+ \mathbf{S}_b \mathbf{S}_t^+ \mathbf{K} = \mathbf{S}_t^+ \mathbf{K} \mathbf{A}$$

$$\mathbf{S}_t^+ \mathbf{S}_b \mathbf{W} = \mathbf{W} \mathbf{A}$$

- **The column vectors of  $\mathbf{W}$  are the eigenvectors of  $\mathbf{S}_t^+ \mathbf{S}_b$ .**



# Generalized Linear Discriminant Analysis (GLDA)

---

- Both  $\mathbf{S}_t$  and  $\mathbf{S}_b$  is  $D \times D$ . Since  $D$  is usually many thousands, it is very time and memory consuming to calculate  $\mathbf{S}_t^+$  and the eigenvectors of  $\mathbf{S}_t^+ \mathbf{S}_b$  using common computational methods.
- Devise a fast algorithm to efficiently calculate  $\mathbf{S}_t^+$  and the eigenvectors of  $\mathbf{S}_t^+ \mathbf{S}_b$  via singular value decomposition (SVD).



# A fast Algorithm

---

$$W = S_t^+ K = S_t^{+\frac{1}{2}} S_t^{+\frac{1}{2}} K$$

$$S_t = U \Lambda U^T \Rightarrow S_t^+ = U \Lambda^{-1} U^T \Rightarrow S_t^{+\frac{1}{2}} = U \Lambda^{-\frac{1}{2}} U^T$$

$$S_t = X X^T \text{ with } X = \frac{1}{\sqrt{n}} [(x_1 - m), \dots, (x_n - m)]$$

Thus, we can obtain the eigenvalues and corresponding eigenvectors through the SVD of X. Note that dimensionality of X is  $D \times n$ , where n is the number of samples  $\ll D$

The column vectors of  $S_t^{+\frac{1}{2}} K$  are the eigenvectors of  $S_t^{+\frac{1}{2}} S_b S_t^{+\frac{1}{2}}$

Recall  $S_b = M M^T$  with  $M = [\sqrt{P_1}(m_1 - m), \dots, \sqrt{P_c}(m_c - m)]$ , we

can obtain  $S_t^{+\frac{1}{2}} S_b S_t^{+\frac{1}{2}} = (S_t^{+\frac{1}{2}} M)(S_t^{+\frac{1}{2}} M)^T$

Thus, we can obtain  $S_t^{+\frac{1}{2}} K$  by the SVD of  $S_t^{+\frac{1}{2}} M$

Since  $S_t^{+\frac{1}{2}} M$  has the dimensionality  $D \times c$  and the number of classes c is  $\ll D$ .



# A fast Algorithm

## ALGORITHM GENERALIZED LINEAR DISCRIMINANT ANALYSIS

**Input:** A gene expression dataset containing  $n$  samples with corresponding labels from  $c$  classes.

**Output:** The mapping matrix  $W$ .

**Method:**

- 1: Calculate  $M = [\sqrt{p_1}(\mathbf{m}_1 - \mathbf{m}), \dots, \sqrt{p_c}(\mathbf{m}_c - \mathbf{m})]$  and  $X = \frac{1}{\sqrt{n}}[(\mathbf{x}_1 - \mathbf{m}), \dots, (\mathbf{x}_n - \mathbf{m})]$ .
- 2: Perform the SVD  $X = U\Lambda^{\frac{1}{2}}V^T$
- 3:  $S_t^{+\frac{1}{2}} = U\Lambda^{-\frac{1}{2}}U^T$
- 4: Perform the SVD  $S_t^{+\frac{1}{2}}M = \tilde{U}\tilde{\Lambda}^{\frac{1}{2}}\tilde{V}^T$
- 5:  $W = S_t^{+\frac{1}{2}}\tilde{U}$



# Experiments

---

- Test GLDA on seven public datasets and compare it with PDA, Random Forests, SVM, DLDA, KNN with feature selection of Dudoit and GA/KNN.
- For each dataset, randomly divide it into three part, of which two parts are used for training (both feature selection/extraction methods and classifiers) and the last part is kept for test.
- This procedure is repeated for 200 times and the averages and standard deviations of error rates are listed in Table 1.



# Results

**Table 1. Average classification error rates and standard deviations on seven public datasets based on 200 runs. In the table,  $c$  is the number of classes,  $n$  is the number of samples,  $D$  is the number of genes, and RandFor stands for random forests.**

	Features	Leukemia	Colon	Prostate	Lymphoma	SRBCT	Brain	GCM
$c$		2	2	2	3	4	5	14
$n$		72	62	102	62	63	42	190
$D$		3571	2000	6033	4026	2308	5597	16063
GLDA	$c - 1$	$3.1 \pm 2.8$	$14.5 \pm 5.7$	$7.6 \pm 3.7$	$0.05 \pm 0.47$	$1.9 \pm 2.6$	$16.6 \pm 8.8$	$17.9 \pm 3.8$
PDA	$c - 1$	$3.3 \pm 2.7$	$14.0 \pm 5.7$	N/A	$0.17 \pm 0.88$	$1.7 \pm 2.4$	N/A	N/A
RandFor	200	$3.0 \pm 3.2$	$14.6 \pm 6.2$	$8.0 \pm 4.5$	$0.76 \pm 2.31$	$2.0 \pm 2.8$	$23.6 \pm 8.3$	$28.1 \pm 4.5$
SVM	200	$2.0 \pm 2.5$	$13.7 \pm 6.4$	$8.6 \pm 4.5$	$1.07 \pm 2.26$	$2.3 \pm 3.1$	$22.5 \pm 9.7$	$32.8 \pm 4.4$
DLDA	200	$2.8 \pm 2.9$	$14.4 \pm 6.6$	$15.5 \pm 7.7$	$1.71 \pm 2.57$	$2.1 \pm 2.7$	$24.0 \pm 9.7$	$31.6 \pm 5.0$
$k$ NN	200	$3.7 \pm 3.1$	$18.2 \pm 6.4$	$11.2 \pm 4.6$	$1.12 \pm 2.24$	$0.9 \pm 1.9$	$23.0 \pm 8.3$	$44.1 \pm 4.6$
RandFor	50	$3.8 \pm 3.6$	$14.8 \pm 6.0$	$7.9 \pm 4.8$	$2.29 \pm 3.63$	$1.6 \pm 2.9$	$25.3 \pm 9.8$	$34.1 \pm 4.6$
SVM	50	$3.2 \pm 3.4$	$13.5 \pm 5.8$	$8.4 \pm 4.5$	$2.07 \pm 3.29$	$1.8 \pm 3.2$	$25.7 \pm 10.1$	$39.3 \pm 4.6$
DLDA	50	$2.8 \pm 2.9$	$13.6 \pm 6.4$	$12.3 \pm 6.5$	$3.86 \pm 3.92$	$1.8 \pm 3.6$	$25.8 \pm 10.5$	$38.2 \pm 5.2$
$k$ NN	50	$4.2 \pm 3.4$	$19.3 \pm 7.2$	$12.6 \pm 4.8$	$2.12 \pm 3.46$	$1.8 \pm 2.6$	$26.6 \pm 9.6$	$46.8 \pm 5.8$
RandFor	10	$4.6 \pm 3.8$	$15.9 \pm 6.4$	$8.8 \pm 4.3$	$3.93 \pm 4.33$	$17.0 \pm 10.3$	$39.2 \pm 12.5$	$45.4 \pm 5.0$
SVM	10	$3.5 \pm 3.5$	$13.3 \pm 5.7$	$8.2 \pm 4.2$	$4.76 \pm 5.01$	$19.0 \pm 10.6$	$38.8 \pm 12.1$	$49.8 \pm 5.8$
DLDA	10	$3.2 \pm 3.5$	$12.9 \pm 6.0$	$11.0 \pm 5.6$	$7.76 \pm 5.85$	$17.4 \pm 10.1$	$37.3 \pm 14.0$	$49.6 \pm 6.0$
$k$ NN	10	$3.6 \pm 3.5$	$18.6 \pm 6.8$	$11.0 \pm 4.7$	$5.00 \pm 5.16$	$21.6 \pm 12.0$	$43.9 \pm 12.9$	$54.0 \pm 4.7$



# Results

---

**Table 2. The averages and standard deviations of the error rates of GA/KNN given 10, 50, or 200 top ranked genes.**

Datasets	Runs	10	50	200
Prostate	100	$7.3 \pm 3.5$	$8.1 \pm 4.1$	$8.6 \pm 4.4$
SRBCT	200	$3.0 \pm 3.4$	$1.3 \pm 2.3$	$1.4 \pm 2.3$
Brain	200	$31.5 \pm 9.9$	$22.1 \pm 8.2$	$21.0 \pm 8.2$
GCM	25	$55.6 \pm 5.0$	$41.7 \pm 5.0$	$36.3 \pm 3.6$



# Results

**Table 3. The leave-one-out cross validation accuracy on the GCM training dataset. The left part is the results on genes selected by RFE that were obtained by Ramaswamy *et al.* [33]. The right part is the results on genes selected by GLDA. In the table,  $s$  is the number of selected genes per classifier.**

$s$	RFE		GLDA		
	$k$ NN OVA	SVM OVA	Avg. Genes	$k$ NN	GLDA
30	65.3%	70.8%	212	67.4%	78.5%
92	68.0%	72.2%	461	65.3%	81.9%
281	65.7%	73.4%	1132	66.0%	81.9%
1073	66.5%	74.1%	3972	66.7%	85.4%
3276	66.3%	74.7%	9821	67.4%	85.4%
6400	64.2%	75.5%	14512	67.4%	84.7%
All	N/A	78.0%	All	67.4%	84.7%



# Conclusion

---

- Gene expression profiling has great potential for accurate cancer diagnosis. It also brings machine learning researchers two challenges, the curse of dimensionality and the small sample size problem.
- In this paper, the author has presented a novel method GLDA to solve these two problems.
  - 1) They give us a mathematical proof to show that the column vectors of  $W$  are the eigenvectors of  $S_t + S_b$ .
  - 2) They provide a fast algorithm to compute the  $W$ .
  - 3) Extensive experiments on seven public datasets demonstrate that the method is able to classify tumors robustly with a high accuracy.
- Besides cancer classification, GLDA may also find applications in other areas where the small sample size problem and the curse of dimensionality arise.