Classification of Acute Leukemia Via Partial Least Squares [a]

Danh V. Nguyen

Department of Statistics

and

David M. Rocke

Department of Applied Science

University of California, Davis

Davis, CA 95616

December, 2000

[a] Critical Assessment of Techniques for MicroArray Data Analysis (CAMDA'00), Dec. 18-19, 2000, Duke University, Durham, N.C.

Outline

1. **Introduction**
2. **Results**
   - Comparing to original results of Golub et al
   - Comparing PLS to classical method of PCA
   - Assessing component construction and classification methods
3. **Potential Applications of PLS**
4. **Computational (Real) Time**
5. **Just Some Details** (if time available)
   - Thresholding algorithm for data re-scaling/normalization
   - Dimension reduction methods: PCA, PLS, & others
   - References

Introduction

1. **Data**
   - A matrix of gene expression values: $\mathbf{X}_{N \times p}$, $N << p$
   - A vector of responses: $\mathbf{y}$, $y = 1$ (AML), $y = 0$ (ALL)
   - Original data: 38 training samples, 34 test samples
2. **Classification of acute leukemia/class prediction**
   - How to predict leukemia classes (AML, ALL), $\mathbf{y}$, based on gene expression data, $\mathbf{X}$?
   - How to make use of classical methods: logistic discrimination (LD), quadratic discriminant analysis (QDA)?
3. **Dimension reduction**
   - Reduce $p \rightarrow N > p^*$
   - Methods: PCA, PLS, and other variants PLSM2, PLSM1
4. **Re-scaling/normalization**

- **Results**: **Gene not selected for predictive ability**

	% correct, train		% correct, test	
Gene Set	PLS	PC	PLS	PC
		LD		
Set A, $p = 1,554$	100.00	84.21	91.18	73.53
Set B, $p = 1,076$	100.00	81.58	91.18	73.53
Set C, $p = 864$	100.00	84.21	91.18	73.53
Set D, $p = 662$	100.00	81.58	91.18	73.53
Set E, $p = 246$	100.00	76.32	79.41	64.71
		QDA		
Set A	100.00	84.21	91.18	82.35
Set B	100.00	84.21	94.12	82.35
Set C	100.00	81.58	91.18	82.35
Set D	100.00	81.58	91.18	88.24
Set E	100.00	57.89	71.05	50.00

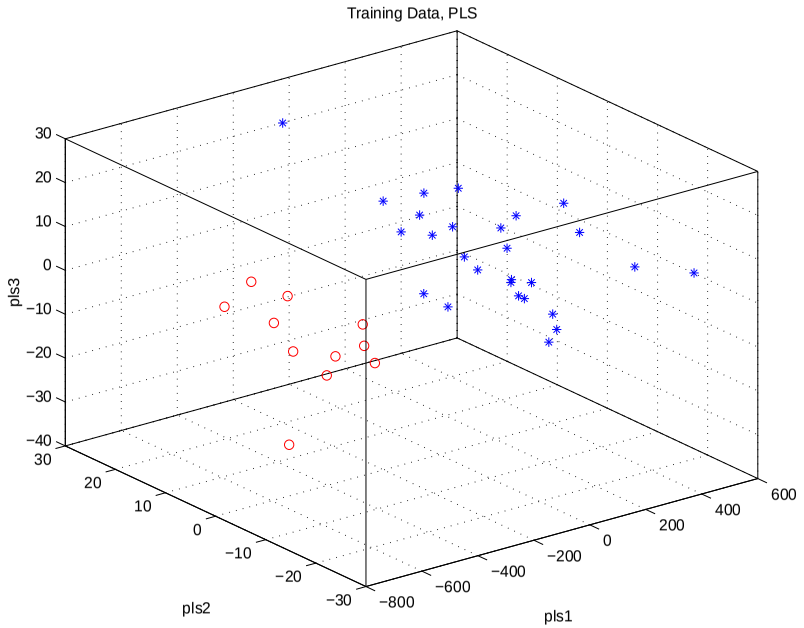
- **Comparison**: **Gene set A to results of Golub et al.**
- **38 Training Samples (CV): 27 ALL, 11 AML**
  1. Golub et al.: Overall 36/38 (97.4%) w/ 2 uncertain (# 12^*, ALL, $PS = 0.20$ #35, AML, $PS = 0.21$)
  2. QDA and LD w/ PLS: 100% w/ $p = prob(correct|data) = 1.00$
  3. PCs did poorly 32/38 (84.2%) in both QDA & LD
- **34 Test Samples: 20 ALL, 14 AML**
  1. Golub et al.: Overall 32/34 (94.1%) w/ 5 uncertain (# 54, 57, 60, 66^*, 67^*)
  2. QDA and LD w/ PLS: 31/34 (91.2%) (# $54^*, 60^*, 66^*$)
  3. 2 samples, # 57, #67 ($PS = 0.22, 0.15$); correctly w/ $p = 0.99, 1.00$ (QDA) & $p = 0.65, 0.99$ (LD)
  4. #71, $PS = 0.3$; correcly w/ $p = 1.00$
  5. PCs only 73.5% (LD), 82.4% (QDA)

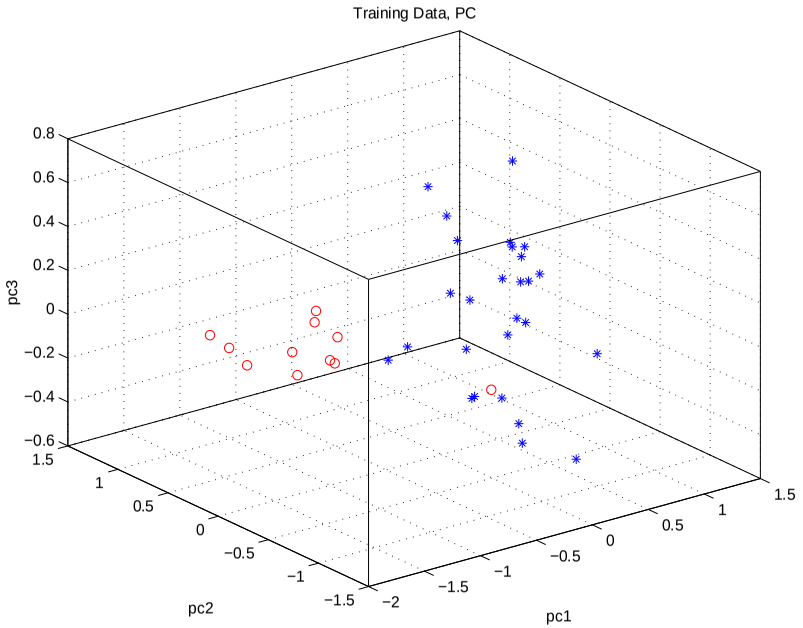
Gene Selection, Thresholding Algorithm & PLS

- **Results**: **50 Gene selected for predictive ability**
- **38 Training Samples (CV): 27 ALL, 11 AML**
  1. QDA and LD w/ PLS: 100% w/ $p = 1.00$ as before
  2. PCs improved drastically: 36/38 (94.7%), (12, 35, both low PS)
- **34 Test Samples: 20 ALL, 14 AML**
  1. LD w/ PLS: 33/34 (97.1%) (66^*)
  2. 5 test samples (# $54, 57, 60^*, 67, 71$) w/ low PS =0.23, 0.22, 0.06, 0.15, 0.30 were correcly classified w/ p =0.97, 1.00, 0.98, 0.89, 1.00
  3. Remaining sample # 66 by PLS and Golub et al.
  4. QDA w/ PCs: improved 91.2%

Separability of Leukemia Classes

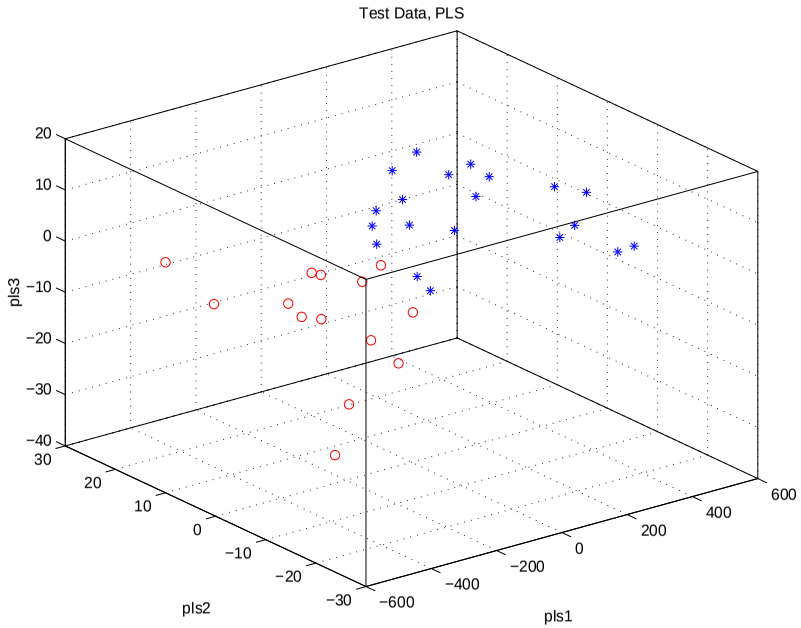
- **Training components based on 50 genes**

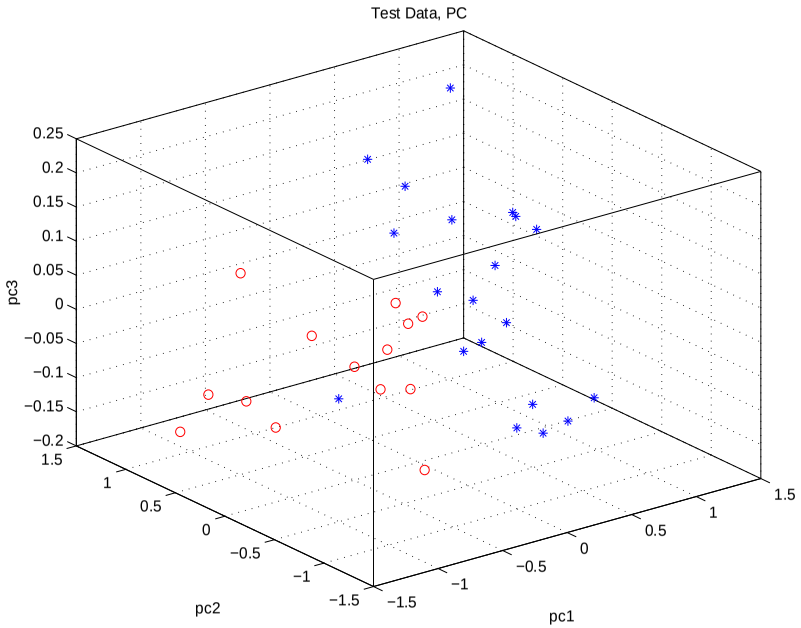




Separability of Leukemia Classes

- **Test components based on 50 genes**





Re-randomizations

- 36 training samples/36 test samples split

	LD		QDA	
	PLS	PC	PLS	PC
Training Data	99.56	96.44	99.56	97.00
Test Data	95.94	94.17	96.44	95.44

- **Training**: PLS components performed at least as well as PCs **50/50** (QDA or LD)
- **Test**: PLS components performed at least as well as PCs **42/50** (QDA or LD)
- Performance of PLS observed in original data (38/34) holds in re-randomization
- Major improvement w/ PCs

Some Explanations

- **Thresholding algorithm**
  1. Array re-scaling $\rightarrow$ expressions equivalent across arrays.
  2. Retain genes above threshold $q\%$ of arrays–genes not selected for predictive ability.
- **Gene selection**
  1. Expressed genes may not be good predictors of leukemia classes.
  2. $\rightarrow$ genes expressed differentially between leukemia classes
  3. $\rightarrow$ 50 genes of Golub et al.
- **Dimension reduction**: Why PLS perform better than PC?
  1. PC: objective? attained?
  2. PLS: objective? attained?

Response and Predictor Variation Explained

	PLS				PC			
	Predictor		Response		Predictor		Response	
K	Prop.	Tot.	Prop.	Tot.	Prop	Tot.	Prop.	Tot.
1	26.47	26.47	50.02	50.02	**44.46**	44.46	2.35	**2.35**
2	27.19	53.67	26.03	76.05	10.57	55.03	38.27	40.62
3	5.06	**58.72**	17.75	**93.79**	5.32	**60.35**	14.68	**55.30**
4	3.89	62.61	3.82	97.61	4.06	64.42	0.09	55.40
5	1.96	64.58	1.89	99.50	3.27	67.69	8.90	64.30
1	46.26	46.26	86.19	86.19	**46.31**	46.31	84.94	**84.94**
2	14.74	61.00	3.42	89.62	19.34	65.65	0.74	85.68
3	7.23	**68.23**	4.44	**94.05**	5.36	**71.02**	0.16	**85.84**
4	3.62	71.85	1.50	95.56	3.28	74.30	0.33	86.17
5	2.73	74.58	1.24	96.80	2.89	77.19	0.80	86.97

Other Potential Applications of PLS

- prediction of the expression of a target gene
- relationship between gene expression patterns of cell lines and their sensitivity to drug therapy
- prediction of patient survival times based on gene expression patterns

Computational Time

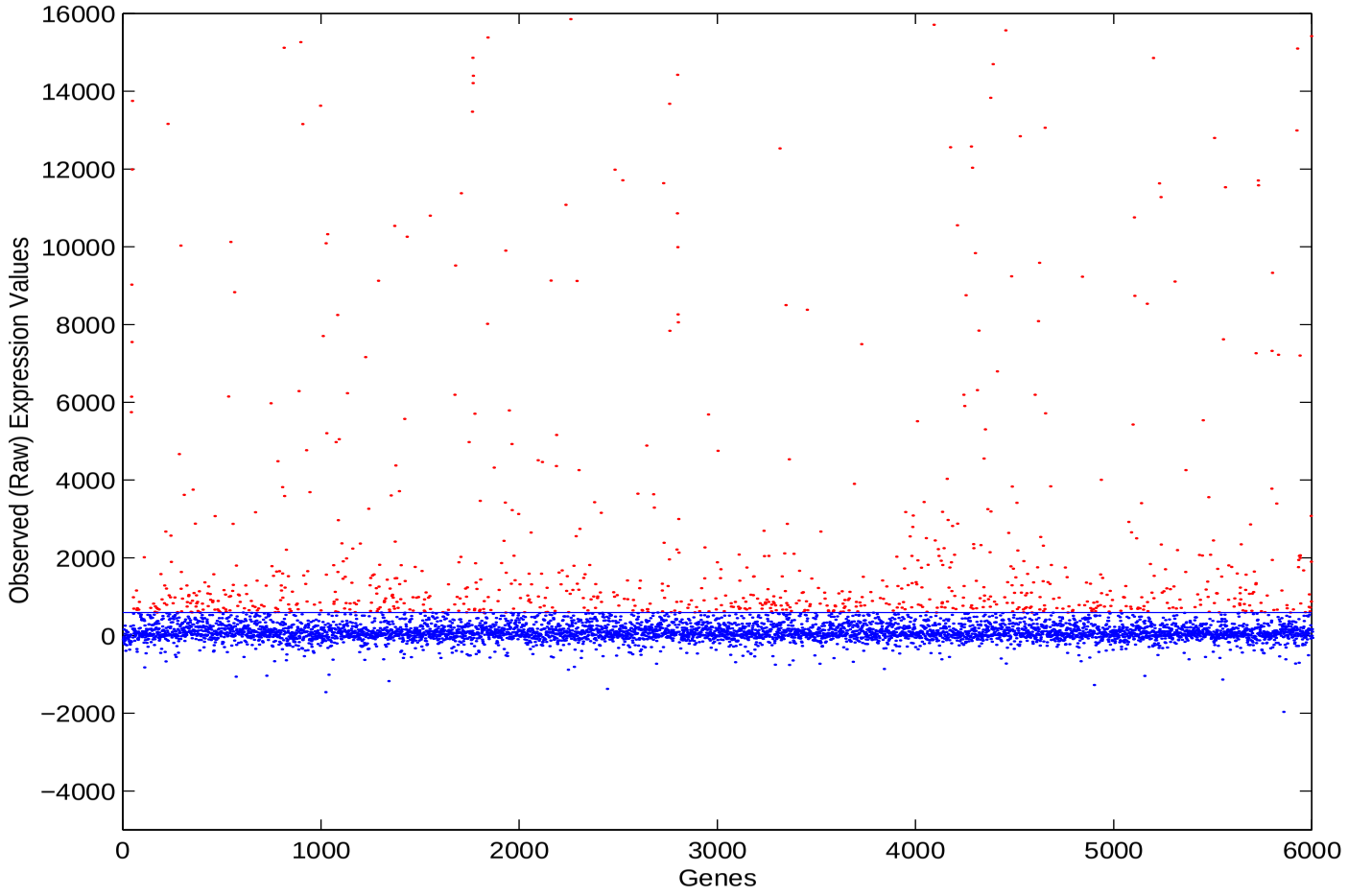
- A blink, a cup, a lunch, a night, a week?
- PLS components $\rightarrow$ a couple of blinks, since $N < p, K = 1, ...N = \max \rightarrow$ no problem
- First few PCs $\rightarrow$ a dozen blinks or more (a cup?), K large $\rightarrow$ problem w/ time
- Classification: LD, QDA–blinks
- Thresholding algorithm: a few dozen blinks

Thresholding Algorithm

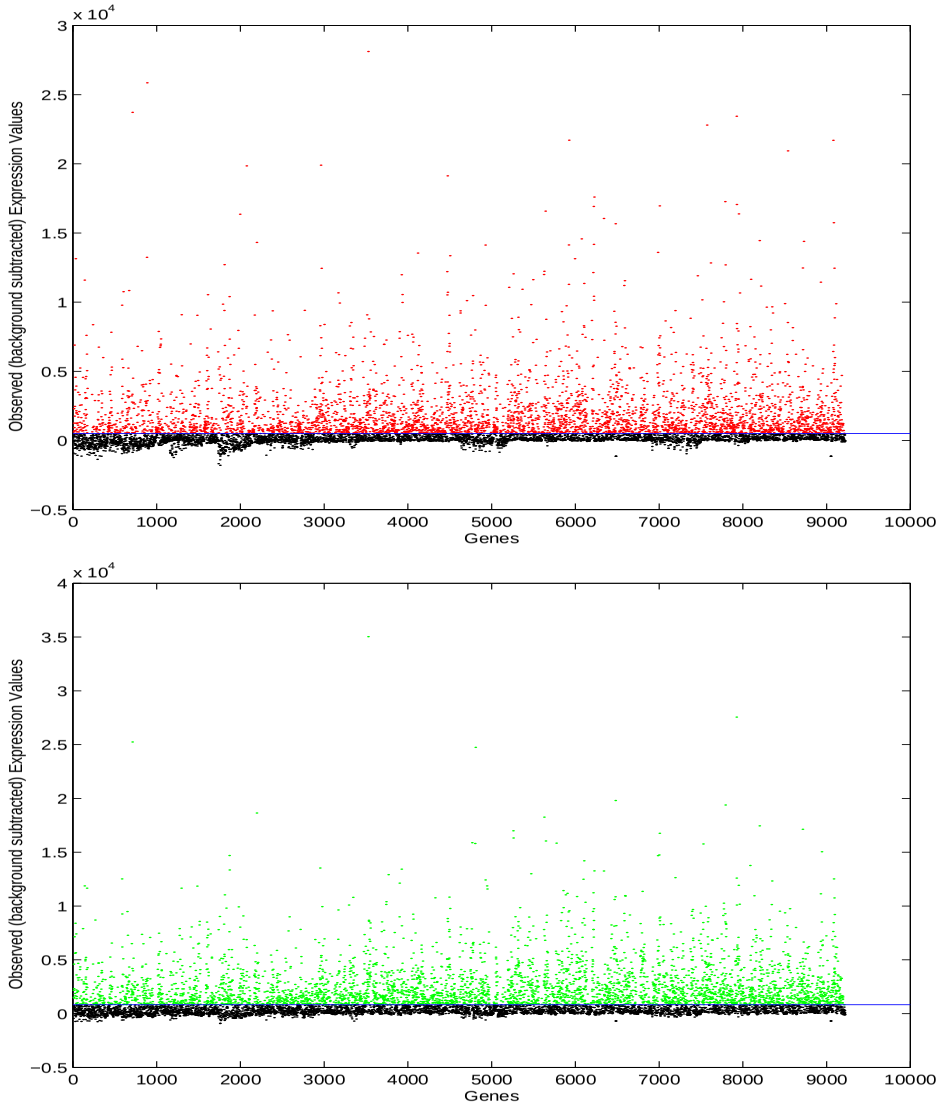
1. $\{x_j\}_{j=1}^{p} \leftarrow$ sort$(\{x_j\}_{j=1}^{p})$
2. Select $q\%$ of lowest values $\rightarrow$ initial set: $A_0 = \{x_1, \ldots, x_{n_0}\}$.
3. Calculate median of the initial set: $m_0 = \text{median}\{x_j\}_{j=1}^{n_0}$.
4. Calculate median of the absolute deviations about the median:
$$MAD_0 = \text{median}\{|x_j - m_0|\}_{j=1}^{n_0} \quad \text{of} \quad A_0.$$
5. Calculate the cutoff point: $u_0 = m_0 + c \times s_0$,
$s_0 = MAD_0/0.6745$ and $c = 2, 2.5$ or 3.
6. Determine the new set: $A_1 = \{\text{all } x_j < u_0\}$.
7. Repeat steps 3-6 and stop when $n_k = n_{k-1}$ (convergence).
8. Repeat steps 2-7 for each array, $i = 1, \ldots, N$.

At convergence (step 7), a set of n_i genes w/ expression levels below cutoff point u_i is A_{n_i}.

Example: typical leukemia array thresholding



Example: Typical cDNA array thresholding



Dimension Reduction

- Optimization and some statistical methods

Method	Objective Function
OLS	$f(\mathbf{c}) = \text{corr}^2(\mathbf{Xc}, \mathbf{y})$
PCA	$f(\mathbf{c}_k) = \text{var}(\mathbf{Xc}_k)$
RR	$f(\mathbf{c}) = \text{corr}^2(\mathbf{Xc}, \mathbf{y})\{\text{var}(\mathbf{Xc})/[\text{var}(\mathbf{Xc}) + \theta]\}$
CCA	$f(\mathbf{c}_k, \mathbf{d}_k) = \text{corr}^2(\mathbf{Xc}_k, \mathbf{Yd}_k)$
PLS (univariate)	$f(\mathbf{c}_k) = \text{cov}^2(\mathbf{Xc}_k, \mathbf{y})$
PLS (multivariate)	$f(\mathbf{c}_k, \mathbf{d}_k) = \text{cov}^2(\mathbf{Xc}_k, \mathbf{Yd}_k)$

- Dimension Reduction
  1. PCA: max **var** of LC of predictors (genes)
  2. UPLS: max **cov** of LC of predictors and response (leukemia classes)
  3. MPLS: max **cov** of LC of predictors and LC of responses.

References

- Helland, I. S. (1988), "On the Structure of Partial Least Squares," *Communications in Statistics–Simulation and Computation*, 17, 581-607.
- Garthwaite, P. H. (1994), "An Interpretation of Partial Least Squares," *Journal of the American Statistical Association*, 89, 122-127.
- Geladi, P. and Kowalski, B. R. (1986), "Partial Least Squares Regression: A Tutorial," *Analytica Chimica Acta*, 185, 1-17.
- Frank, I. E. and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," (with discussion), *Technometrics*, 35, 109-148.