

Classification of High Dimensional Data By Two-way Mixture Models

Jia Li

Statistics Department
The Pennsylvania State University

Outline

- Goals
- Two-way mixture model approach
 - Background: mixture discriminant analysis
 - Model assumptions and motivations
 - Dimension reduction implied by the two-way mixture model
 - Estimation algorithm
- Examples
 - Document topic classification (Discrete)
 - * A mixture of Poisson distributions
 - Disease-type classification using microarray gene expression data (Continuous)
 - * A mixture of normal distributions
- Conclusions and future work

Goals

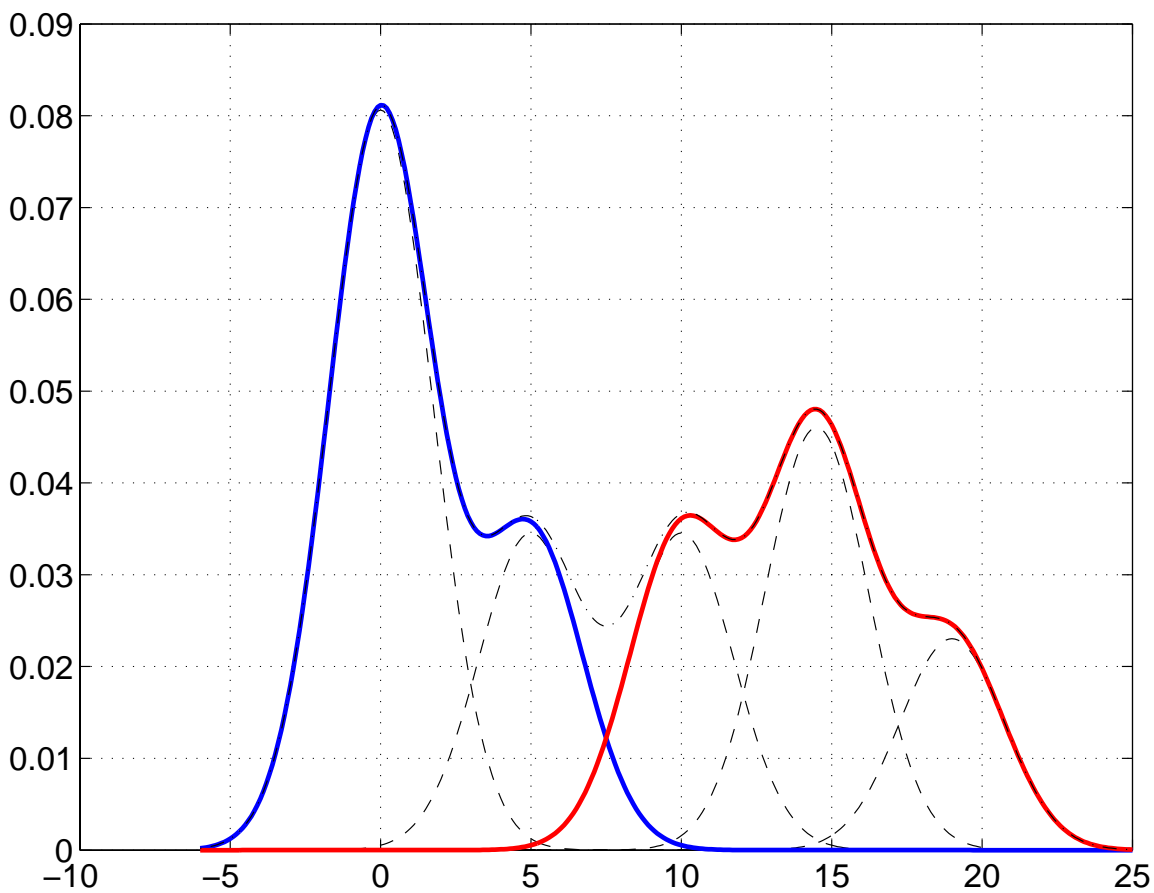
- Achieve high accuracy for the classification of high dimensional data.
 - Document data:
 - * Dimension: $p > 3400$.
 - * Training sample size: $n \approx 2500$.
 - * Number of classes: $K = 5$.
 - * The feature vectors are sparse.
 - Gene expression data:
 - * Dimension: $p > 4000$.
 - * Training sample size: $n < 100$.
 - * Number of classes: $K = 4$.
- Attribute (variable, feature) clustering may be desired.
 - Document data: which words play similar roles and do not need to be distinguished for identifying a set of topics?
 - Gene expression data: which genes function similarly?

Mixture Discriminant Analysis

- Proposed as an extension of linear discriminant analysis.
- T. Hastie, R. Tibshirani, “Discriminant analysis by Gaussian mixtures,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 155-176, 1996.
- The mixture of normals is used to obtain a density estimation for each class.
- Denote the feature vector by X and the class label by Y .
- For class $k = 1, 2, \dots, K$, the within-class density is:

$$f_k(x) = \sum_{r=1}^{R_k} \pi_{kr} \phi(x | \mu_{kr}, \Sigma)$$

- A 2-classes example. Class 1 is a mixture of 3 normals and class 2 a mixture of 2 normals. The variances for all the normals are 3.0.



- The overall model is:

$$\begin{aligned}
 P(X = x, Y = k) &= a_k f_k(x) \\
 &= a_k \sum_{r=1}^{R_k} \pi_{kr} \phi(x | \mu_{kr}, \Sigma)
 \end{aligned}$$

where a_k is the prior probability of class k .

- Equivalent formulation:

$$P(X = x, Y = k) = \sum_{m=1}^M \pi_m \phi(x | \mu_m, \Sigma) q_m(k)$$

where q_m is a pmf for the class label Y within a mixture component.

- Here we have $q_m(k) = 1.0$ if mixture component m “belongs to” class k and zero otherwise.
- The ML estimation of a_k is the proportion of training samples in class k .
- EM algorithm is used to estimate π_{kr} , μ_k , and Σ .
- Bayes classification rule:

$$\hat{y} = \arg \max_k a_k \sum_{r=1}^{R_k} \pi_{kr} \phi(x | \mu_{kr}, \Sigma)$$

Assumptions for the Two-way Mixture

- For each mixture component, the variables are independent.
 - As a class may contain multiple mixture components, the variables are NOT independent in general given the class.
 - To approximate the density within each class, the restriction on each component can be compensated by having more components.
 - Convenient for extending to a two-way mixture model.
 - Efficient for treating missing data.
- Suppose X is p -dimensional, $x = (x_1, x_2, \dots, x_p)^T$. The mixture model is:

$$P(X = x, Y = k) = \sum_{m=1}^M \pi_m q_m(k) \prod_{j=1}^p \phi(x_j | \theta_{m,j})$$

We need to estimate parameter $\theta_{m,j}$ for each dimension j in each mixture component m .

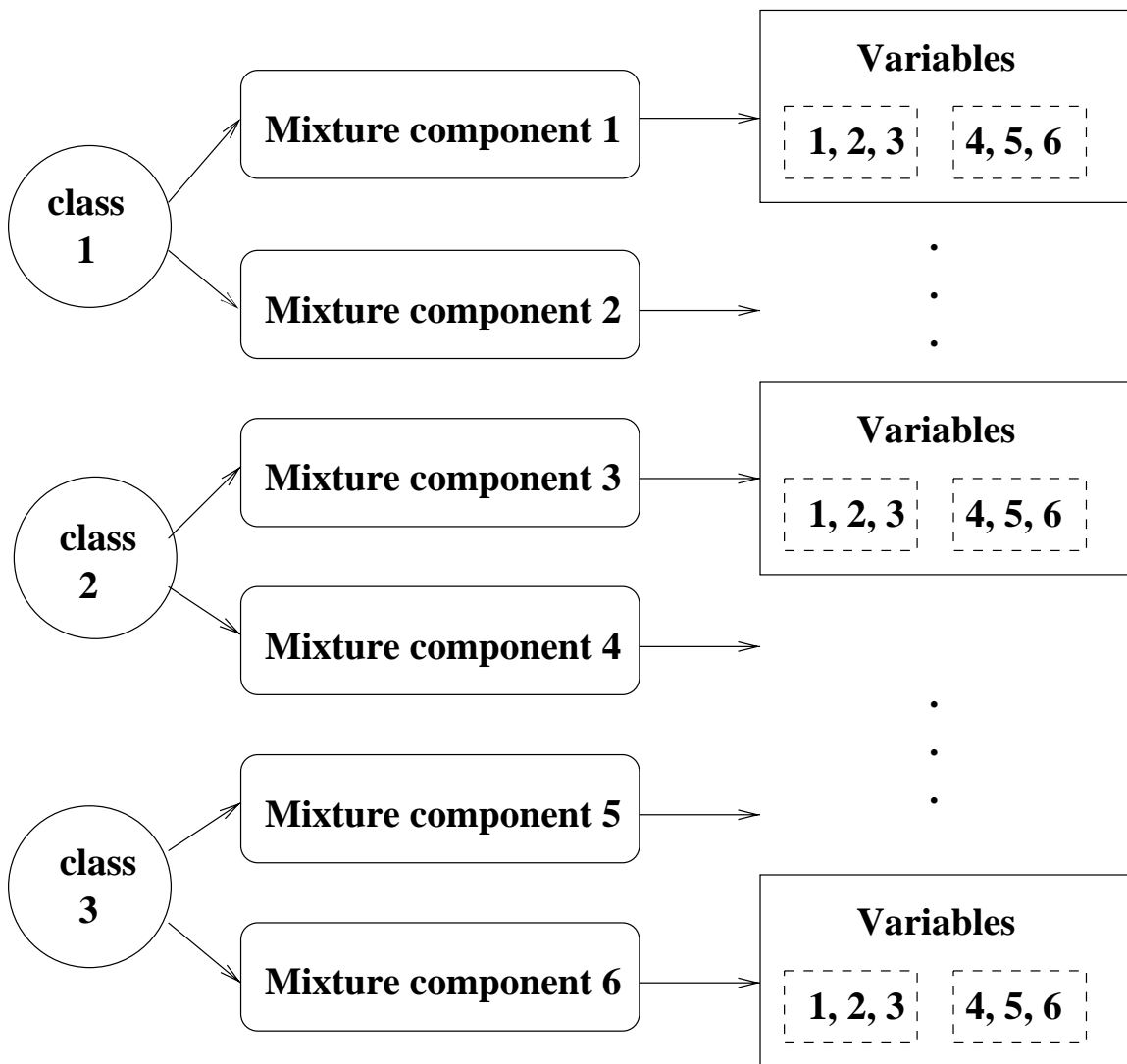
- When the dimension is very high (sometimes $p \gg n$), we may need an even more parsimonious model for each mixture component.
- Clustering structure on the variables:
 - Assume that the p variables belong to L clusters. Two variables j_1, j_2 , in the same cluster have $\theta_{m,j_1} = \theta_{m,j_2}$, $m = 1, 2, \dots, M$.
 - Denote the cluster identity of variable j by $c(j) \in \{1, \dots, L\}$.
 - For a fixed mixture component m , only need to estimate L θ 's.
 - The $\theta_{m,j}$'s are shrunk to L $\theta_{m,c(j)}$'s.

$$P(X = x, Y = k) = \sum_{m=1}^M a_m q_m(k) \prod_{j=1}^p \phi(x_j | \theta_{m,c(j)})$$

- This way of regularizing the model leads to variable clustering.

- Properties of variable clusters:

- Variables in the same cluster have the same distributions within each class.
- For each cluster of variables, only a small number of statistics are sufficient for predicting the class label.



Dimension Reduction

- Within each mixture component, variables in the same cluster are i.i.d. random variables.
- For i.i.d. random variables sampled from an exponential family, the dimension of the sufficient statistic for the parameter θ is fixed w.r.t. the sample size.
- Assume the exponential family to be:

$$p_{\theta}(x_j) = \exp \left(\sum_{s=1}^S \eta_s(\theta) T_s(x_j) - B(\theta) \right) h(x_j)$$

Proposition: For X_j 's in cluster l , $l = 1, \dots, L$, define

$$\bar{T}_{l,s}(x) = \sum_{j:c(j)=l} T_s(x_j) \quad s = 1, 2, \dots, S.$$

Given $\bar{T}_{l,s}(x)$, $l = 1, \dots, L$, $s = 1, \dots, S$, the class of a sample Y is conditionally independent of X_1, X_2, \dots, X_p .

- **Dimension reduction:** “sufficient information” for predicting Y is of dimension LS . We often have $LS \ll p$.

- **Examples:**

- **Mixtures of Poisson:** $S = 1$.

$$\bar{T}_{l,1}(x) = \sum_{j:c(j)=l} x_j$$

- **Mixtures of normal:** $S = 2$.

$$\bar{T}_{l,1}(x) = \sum_{j:c(j)=l} x_j$$

$$\bar{T}_{l,2}(x) = \sum_{j:c(j)=l} x_j^2$$

Equivalently:

$$\text{Sample mean: } \bar{T}_{l,1}(x) = \frac{\sum_{j:c(j)=l} x_j}{\sum_j I(c(j) = l)}$$

$$\text{Sample variance: } \bar{T}_{l,2}(x) = \frac{\sum_{j:c(j)=l} (x_j - \bar{T}_{l,1}(x))^2}{\sum_j I(c(j) = l)}$$

Model Fitting

- We need to estimate the following:
 - Mixture component prior probabilities a_m , $m = 1, \dots, M$.
 - Parameters of the Poisson distributions: $\theta_{m,l}$, $m = 1, \dots, M$, $l = 1, \dots, L$.
 - The variable clustering function $c(j)$, $j = 1, \dots, p$, $c(j) \in \{1, \dots, L\}$.
- Criterion: Maximum likelihood estimation.
- Algorithm: EM.
 - E-step: compute the posterior probability of each sample coming from each mixture component.
 - M-step:
 - * Update the parameters $a_m, \theta_{m,l}$.
 - * Update the variable clustering function $c(j)$ by optimizing $c(j)$ individually for each j , $j = 1, \dots, p$ with all the other parameters fixed.
- Computational perspective:
 - E-step: a “soft” clustering of samples into mixture components, “row-wise” clustering.
 - M-step: 1) update parameters; 2) a clustering of attributes, “column-wise” clustering.

Document Topic Classification

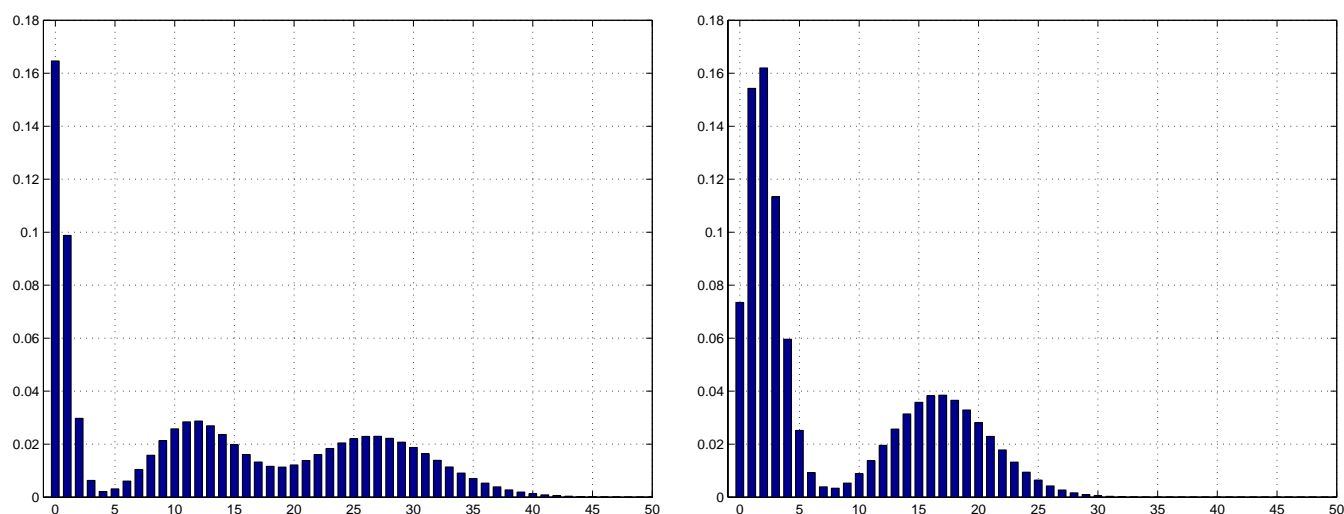
- Classify documents into different topics.
- Five document classes from the Newsgroup data set collected by Lang (1995):
 1. *comp.graphics*
 2. *rec.sport.baseball*
 3. *sci.med*
 4. *sci.space*
 5. *talk.politics.guns*
- Classification is based on word counts.
 - Examples: bit: 2, graphic: 3, sun: 2.
- Each document is represented by a vector of word counts. Every dimension corresponds to a particular word.
- Each class contains about 1000 documents. Roughly half of them are randomly selected as training data, and the others testing.
- Pre-processing: for each document class, the 1000 words with the largest total counts in the training data are used as variables.
- The dimension of the word vector is $p = 3455$, $p > n$.

Mixture of Poisson Distribution

- The Poisson distribution is uni-modal.

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} .$$

- Example mixtures of Poisson distributions:

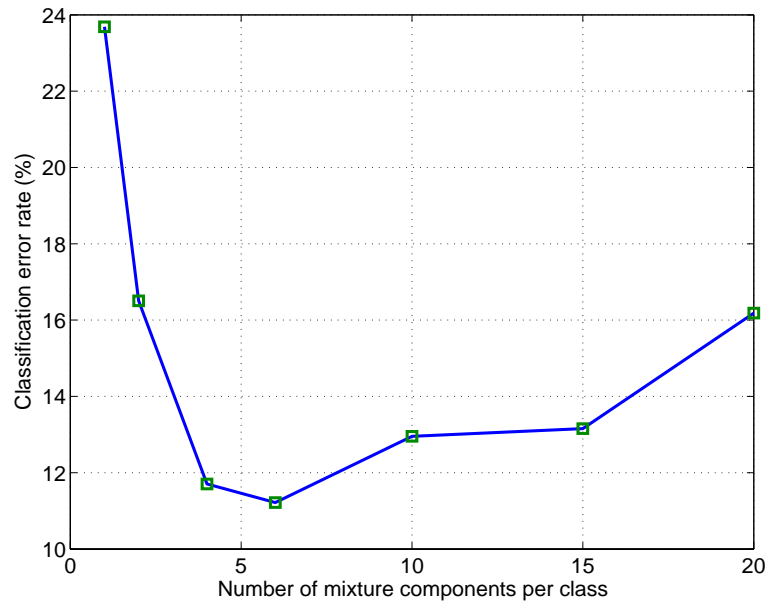


- Mixture of multivariate independent Poisson distributions with variable clustering:

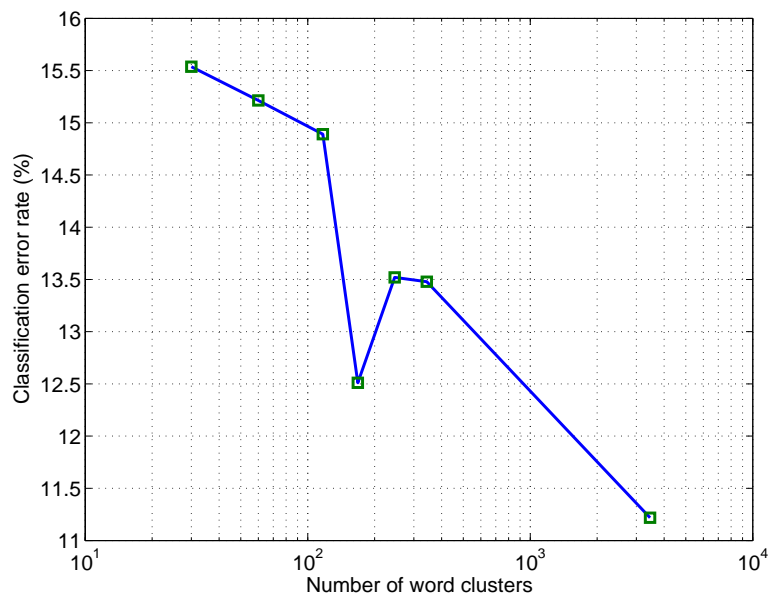
$$P(X = x, Y = k) = \sum_{m=1}^M a_m q_m(k) \prod_{j=1}^p \frac{\lambda_{m,c(j)}^{x_j}}{x_j!} \cdot e^{-\lambda_{m,c(j)}}$$

Results

- Classification error rates achieved without variable clustering. *#components per class* = 1 ~ 20.



- $L = 30 \sim 3455$, *#components per class* = 6.



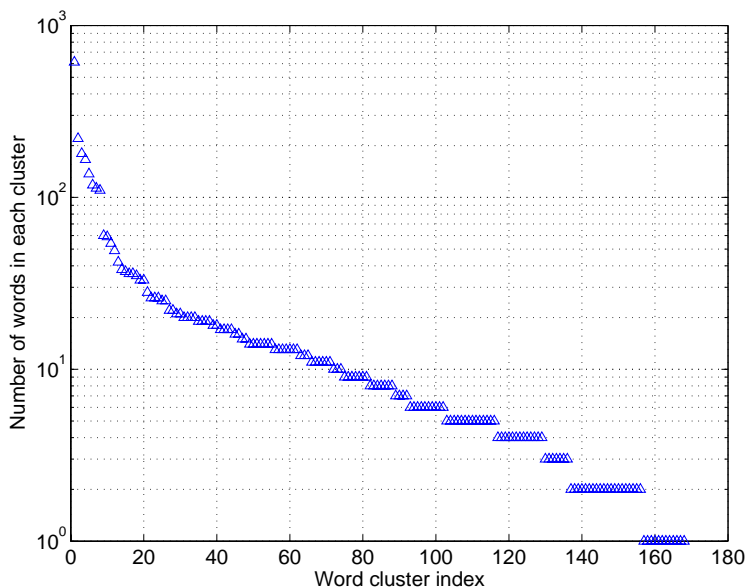
- Confusion table for $M = 30$, without word clustering, $p = 3455$. Classification error rate: 11.22%.

	graphics	baseball	sci.med	sci.space	politics.guns
graphics	463	5	9	16	3
baseball	3	459	4	2	9
sci.med	22	12	435	20	14
sci.space	27	14	28	409	18
politics.guns	11	27	17	17	434

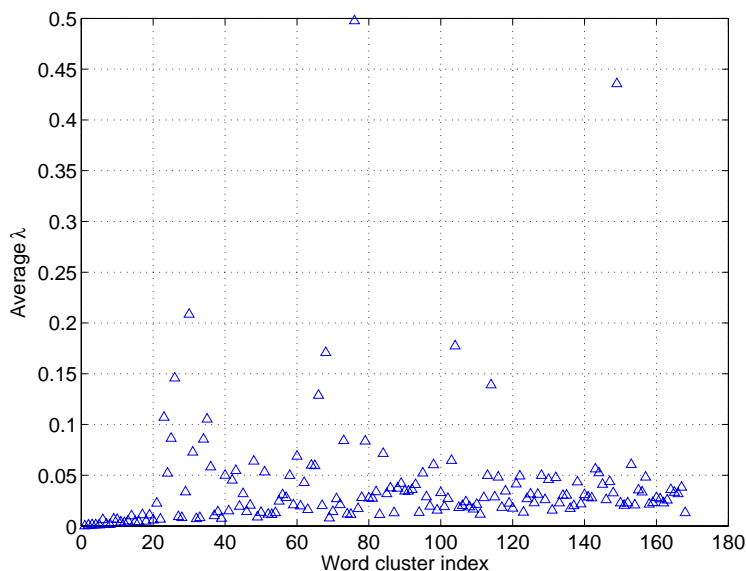
- For $M = 30$, $L = 168$. Classification error rate: 12.51%.

	graphics	baseball	sci.med	sci.space	politics.guns
graphics	458	1	12	15	10
baseball	3	446	2	5	21
sci.med	23	9	408	21	42
sci.space	24	9	21	404	38
politics.guns	4	15	18	17	452

- For $M = 30$, $L = 168$, median cluster size is 7. Highly skewed cluster sizes: the largest 10 clusters account for more than half of the 3455 words.



- The corresponding weighted average of $\lambda_{m,l}$'s for each cluster l , $\sum_{m=1}^M a_m \lambda_{m,l}$, is shown below. The largest few word clusters have very low average counts.



- If the $612 + 220 + 180 + 166 + 137 = 1315$ words in the largest five clusters are not used when classifying test samples, the error rate is only slightly increased from 12.15% to 12.99%.
- Words in all of the clusters with size 5:
 - *patient, eat, food, treatment, physician*
 - *nasa, space, earth, mission, satellit*
 - *compil, transform, enhanc, misc, lc*
 - *game, team, player, fan, pitcher*
 - *unit, period, journal, march, sale*
 - *wai, switch, describ, directli, docum*
 - *faq, resourc, tool, distribut, hardwar*
 - *approxim, aspect, north, angl, simul*
 - *recogn, wisdom, vm, significantli, breast*
 - *bought, simultan, composit, walter, mag*
 - *statu, ny, dark, eventu, phase*
 - *closer, po, paid, er, huge*
 - *necessarili, steven, ct, encourag, dougla*
 - *replac, chri, slow, nl, adob*

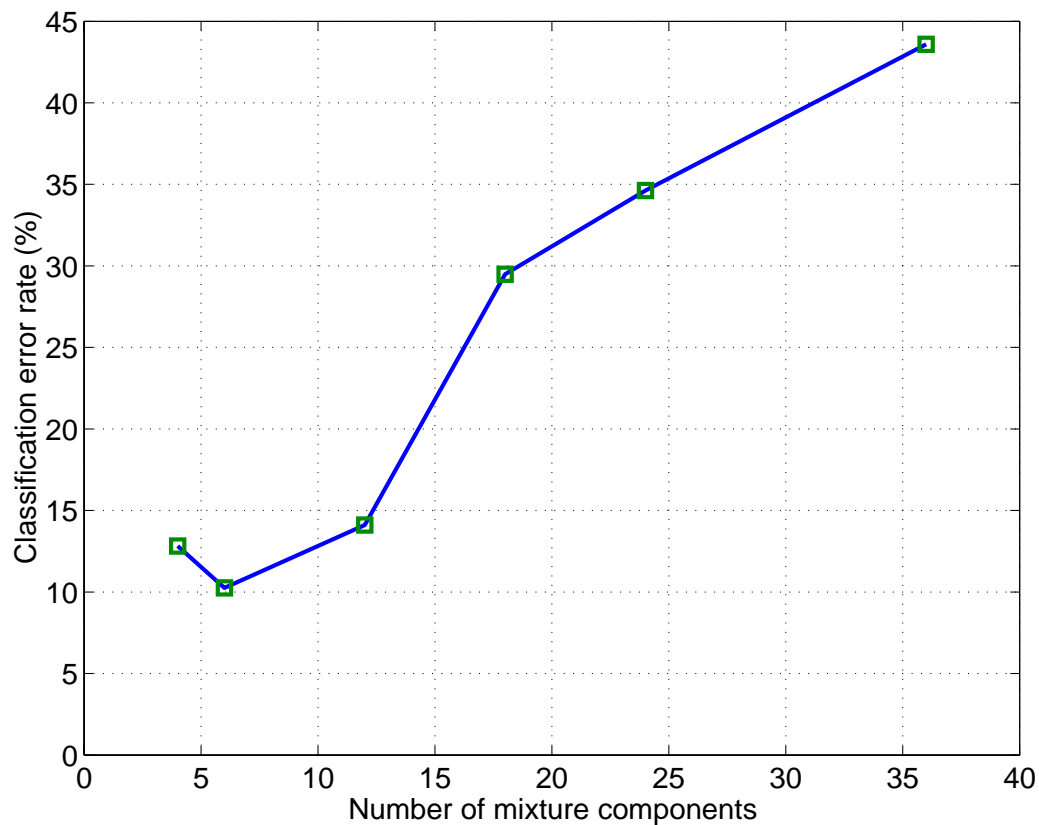
Disease Classification by Microarray Data

- The microarray data are provided at the web site:
<http://llmpp.nih.gov/lymphoma/>
- Every sample in the data set contains expression levels of 4026 genes.
- There are 96 samples divided into 9 classes.
- Four classes of 78 samples are chosen for the classification experiment.
 - DLBCL (diffuse large B-cell lymphoma): 42
 - ABB (activated blood B): 16
 - FL (follicular lymphoma): 9
 - CLL (chronic lymphocytic leukemia): 11
- Five-fold cross-validation is used to assess the accuracy of classification.
- Mixture of normal distribution with variable clustering:

$$P(X = x, Y = k) = \sum_{m=1}^M a_m q_m(k) \prod_{j=1}^p \frac{1}{\sqrt{2\pi\sigma_{m,c(j)}^2}} \exp\left(\frac{-(x_j - \mu_{m,c(j)})^2}{2\sigma_{m,c(j)}^2}\right)$$

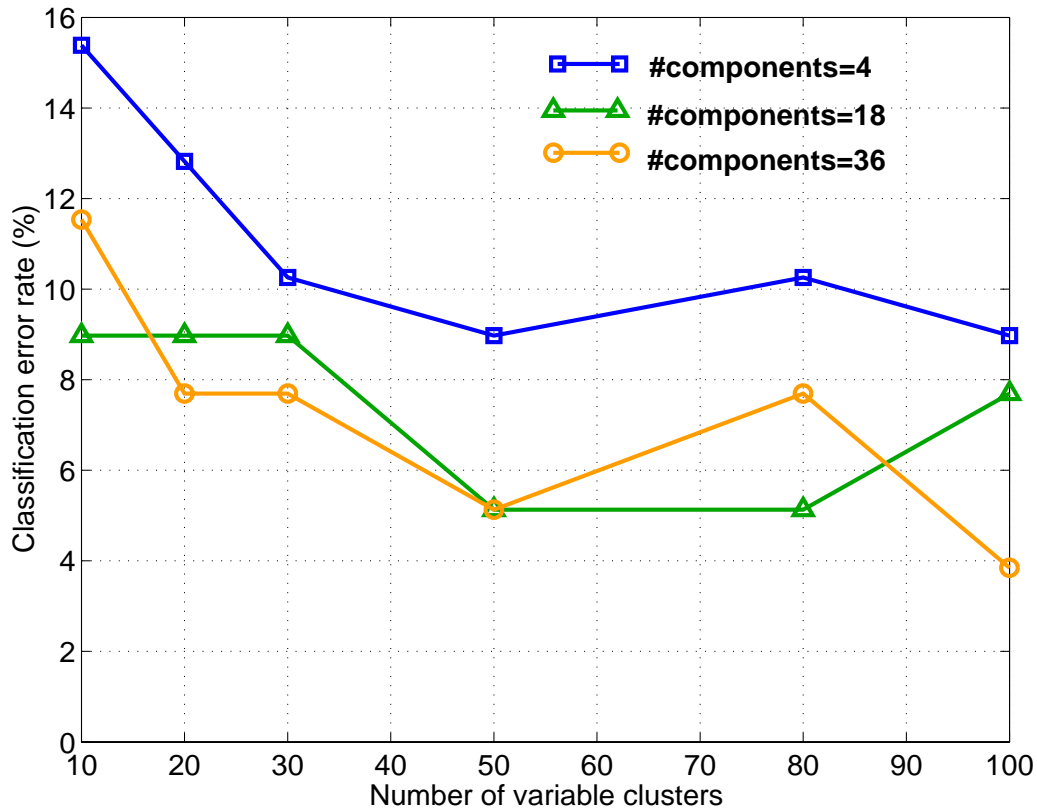
Results

- Classification error rates achieved without variable clustering. $M = 4 \sim 36$.



- Minimum error rate 10.26% is achieved at $M = 6$.
- Due to the small sample size, classification performance degrades rapidly when M increases.

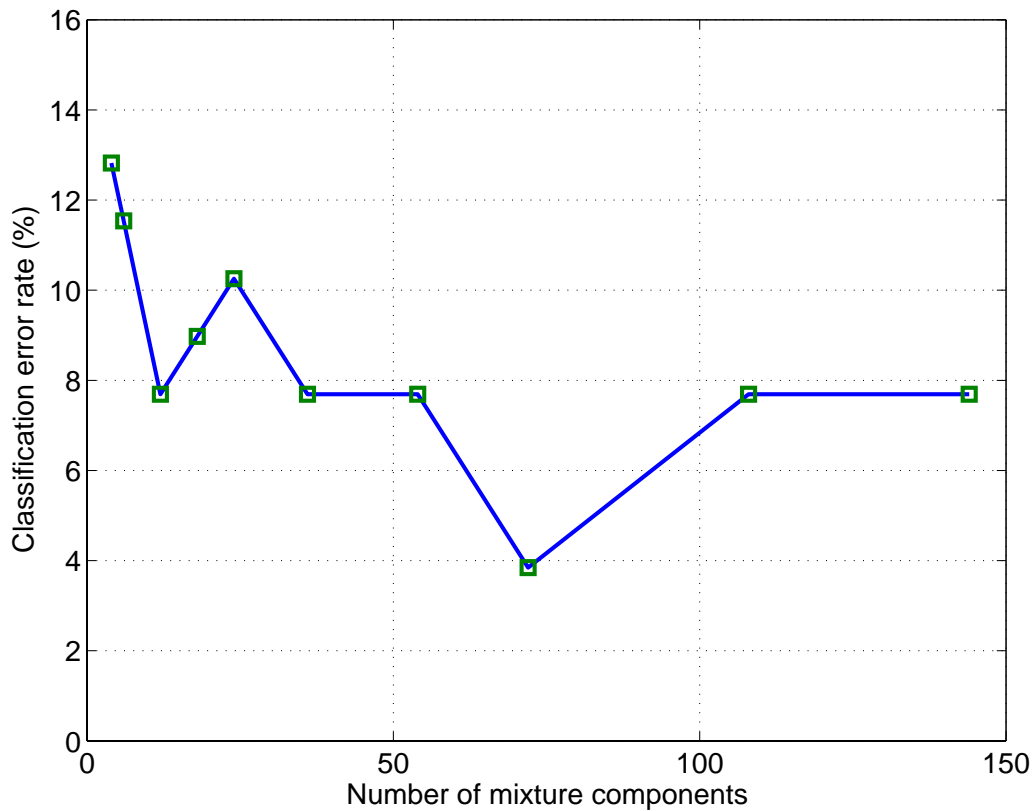
- Classification error rates achieved with gene clustering. $L = 10 \sim 100$, $M = 4, 18, 36$.



- Gene clustering improves classification.

Error rate (%)	$M = 4$	$M = 6$	$M = 12$	$M = 18$	$M = 36$
No clustering	12.82	10.26	14.10	29.49	43.59
$L = 50$	8.97	10.26	7.69	5.13	5.13
$L = 100$	8.97	8.97	6.41	7.69	3.85

- Variable clustering allows us to have more mixture components than the sample size.
- The number of parameters in the model is small due to clustering along variables.
- Fix $L = 20$ (20 gene clusters). $M = 4 \sim 144$.



- When $M \geq 36$, the classification error rate remains below 8%.

Conclusions

- A two-way mixture model approach is developed to classify high dimensional data.
 - This model implies dimension reduction.
 - Attributes are clustered in a way to preserve information about the class of a sample.
- Applications of both discrete and continuous models have been studied.
- Future work:
 - Can the two-way mixture approach be extended to achieve dimension reduction under more general settings?