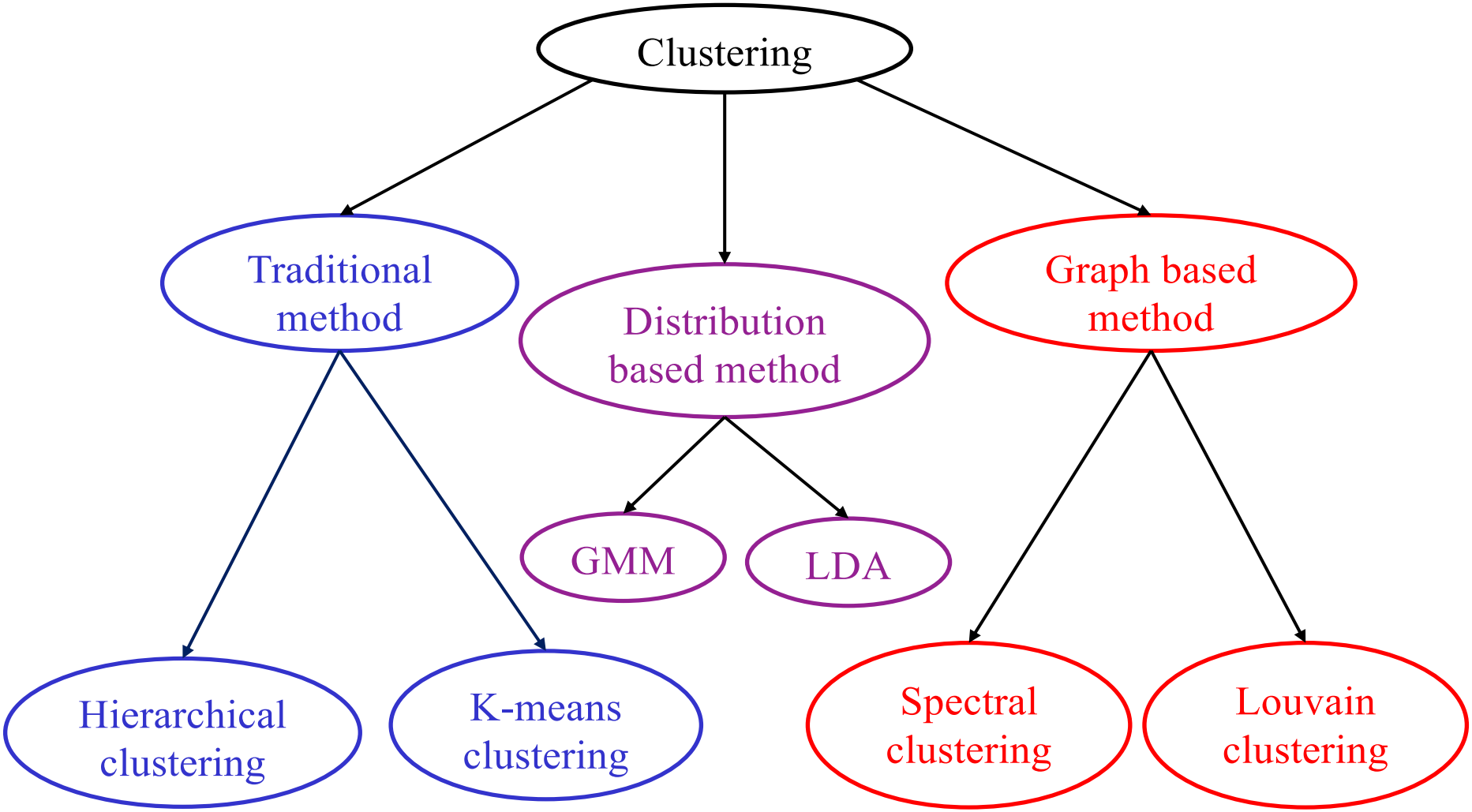
Clustering Methods:
From k-means to Gaussian Mixture Model and Louvain Algorithm

Maxwell Lee

High-dimension Data Analysis Group
Laboratory of Cancer Biology and Genetics
Center for Cancer Research
National Cancer Institute

October 5, 2020

Outline of Clustering Methods



GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation

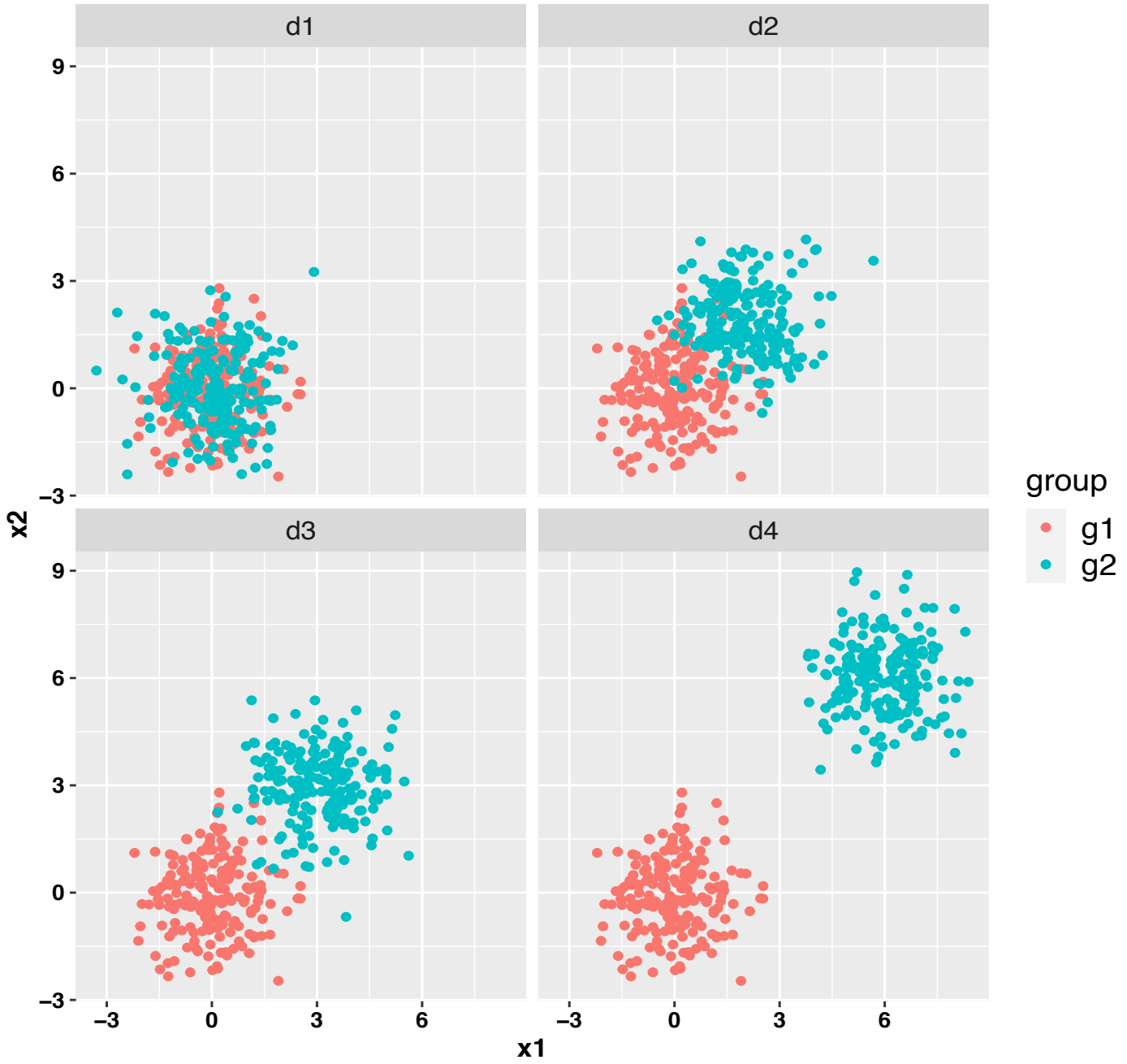
Contributed by Emily Tai

Increased Separation Between Clusters Is Related to Increased Distance Between the Groups

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

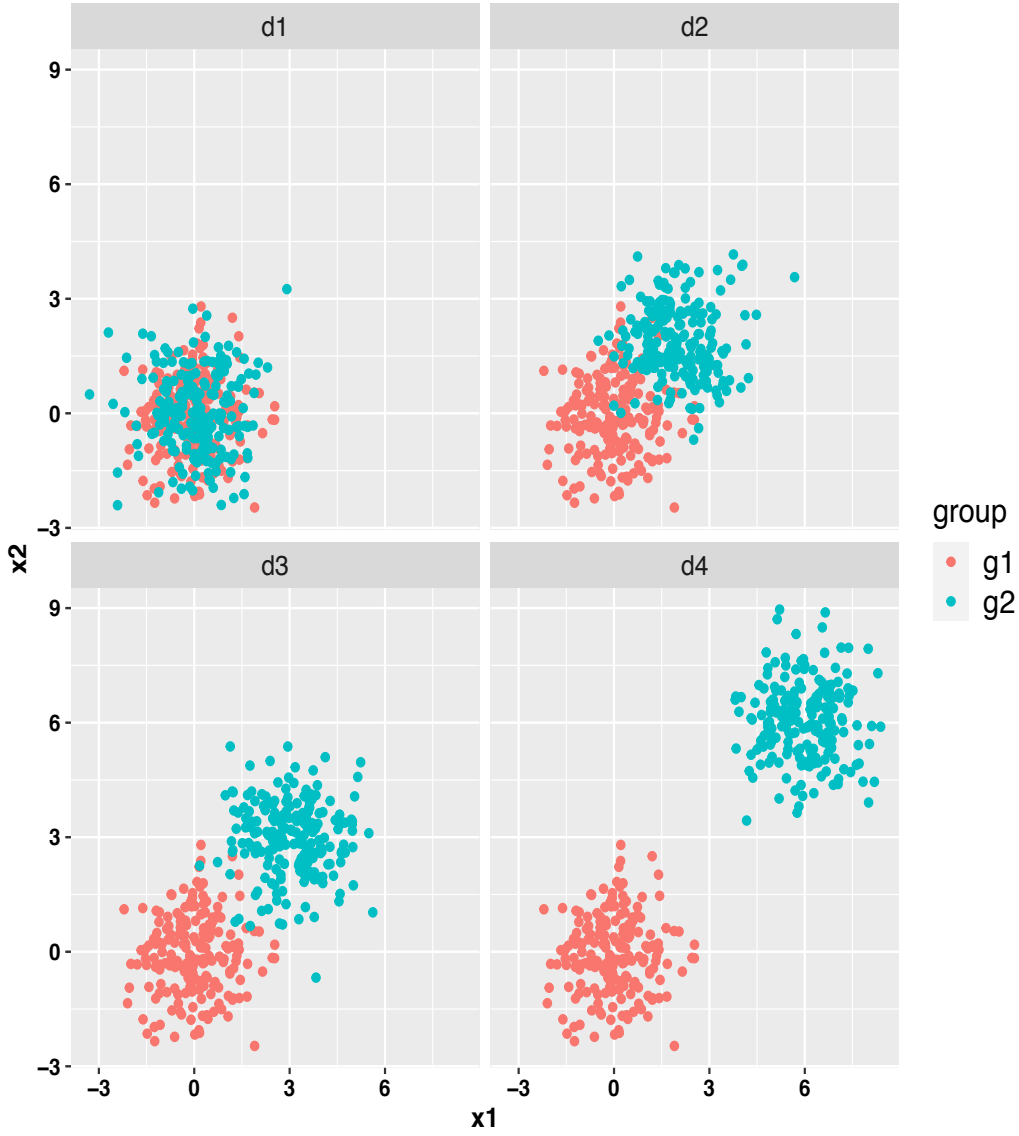
$\rho = 0$

	Group1		Group2	
	μ_1	μ_2	μ_1	μ_2
d1	0	0	0	0
d2	0	0	2	2
d3	0	0	3	3
d4	0	0	6	6

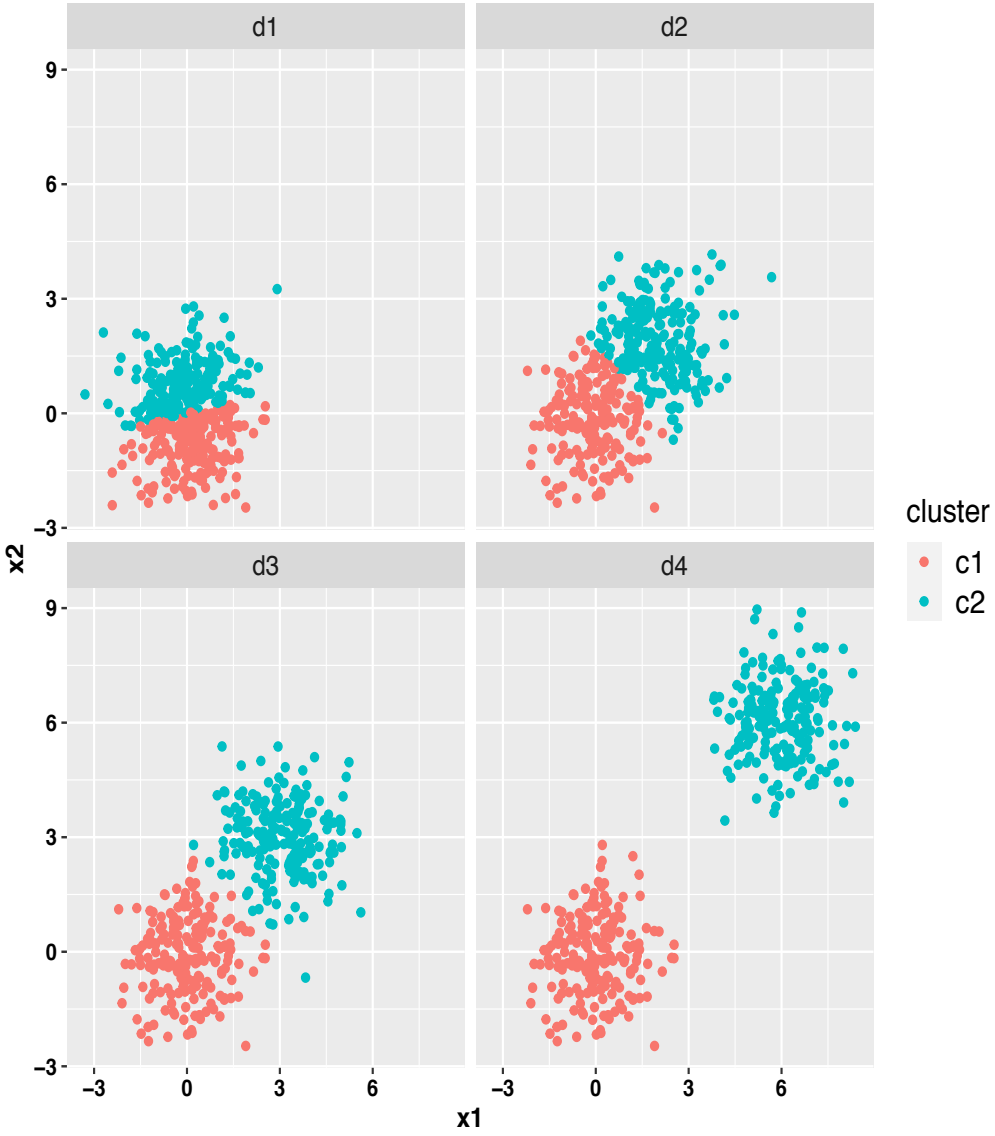


K-means Clustering

color by group



color by k-means cluster



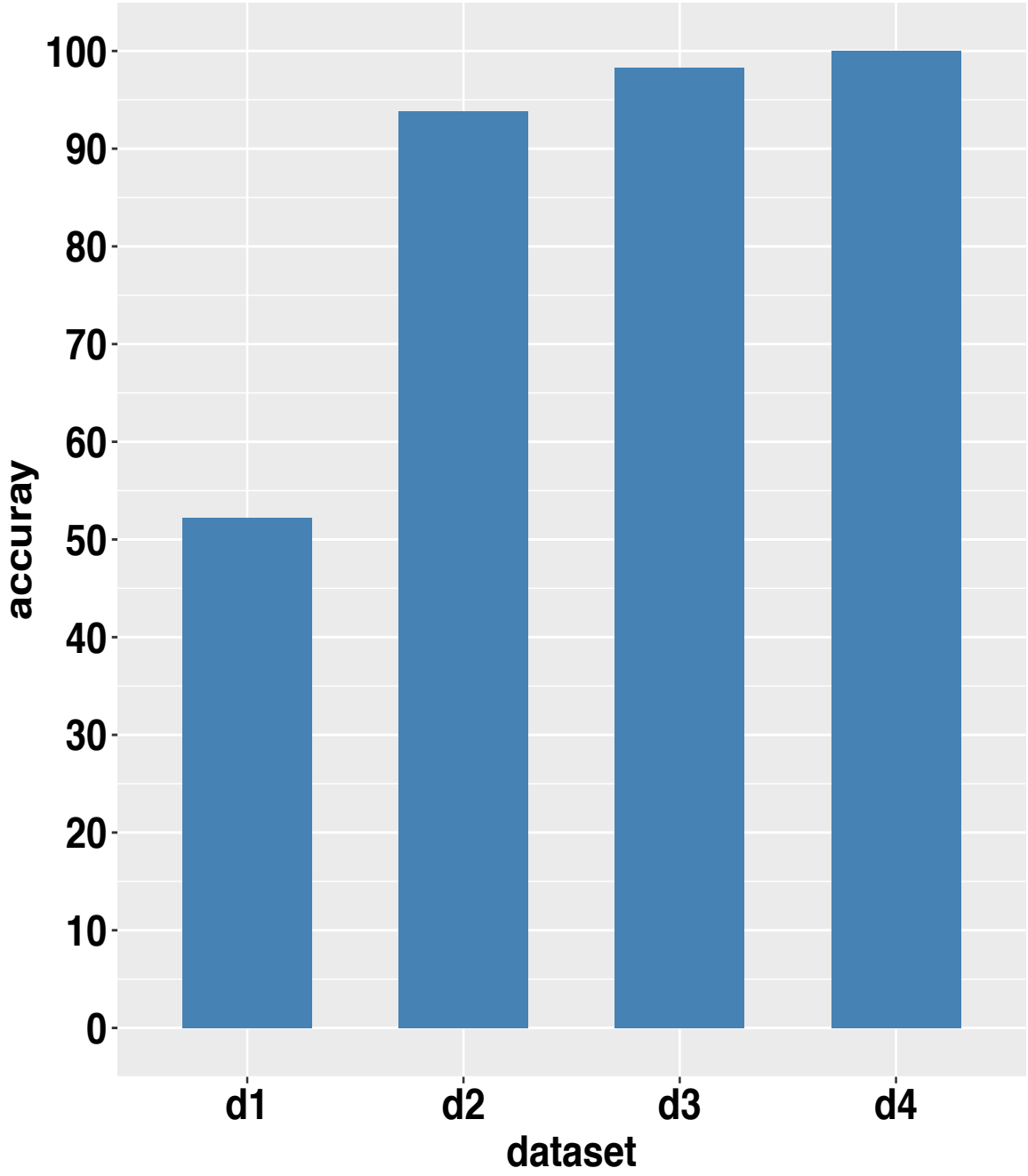
Accuracy of k-means Clustering

Confusion matrix
Column: actual category
Row: assigned category

	g1	g2
c1	183	8
c2	17	192

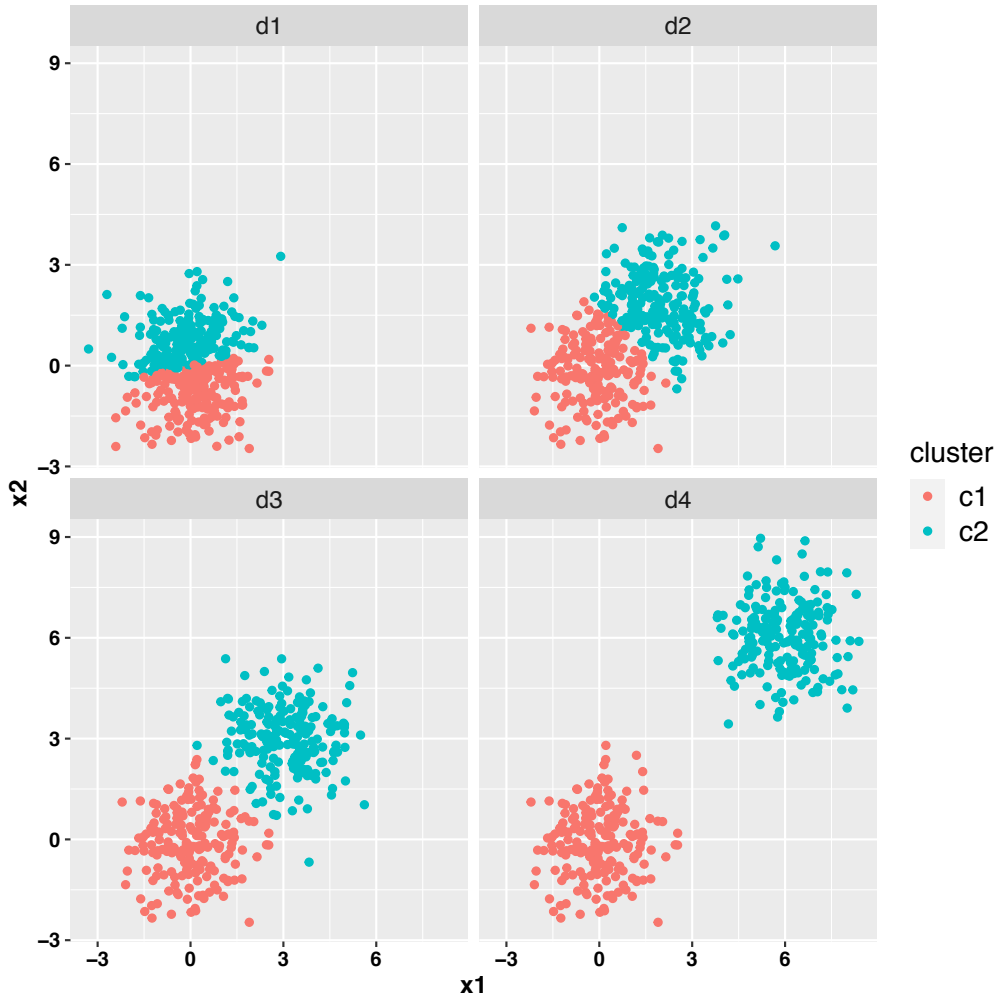
accuracy of dataset d2
Match: diagonal elements (red)
Mismatch: off diagonal elements (green)

accuracy=(183+192)/400
=93.75%



K-means Clustering Uses Euclidean Distance

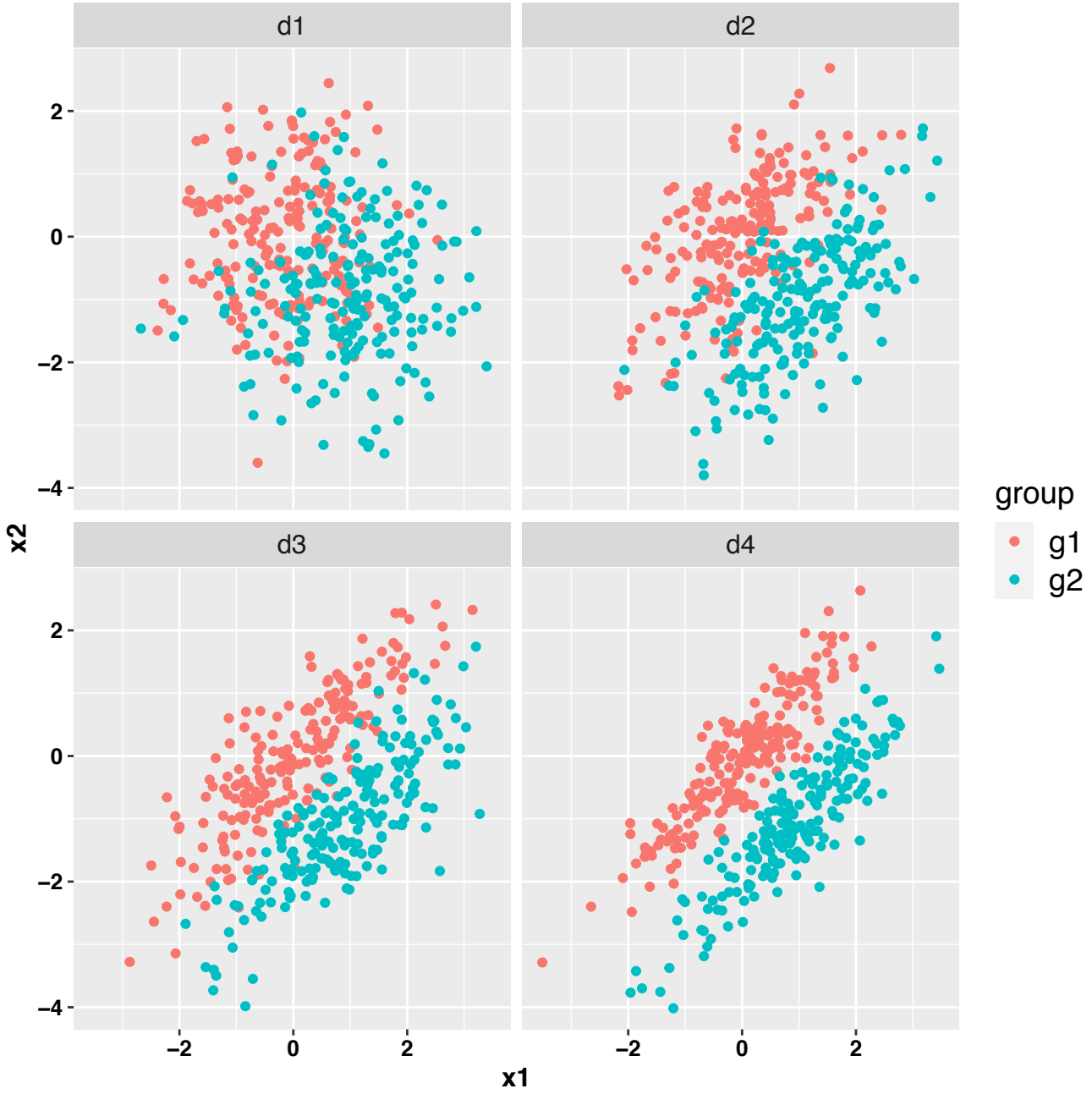
An implicit assumption:
shape of data is sphere (correlation=0)



Effect of Covariance Structure on Clustering

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

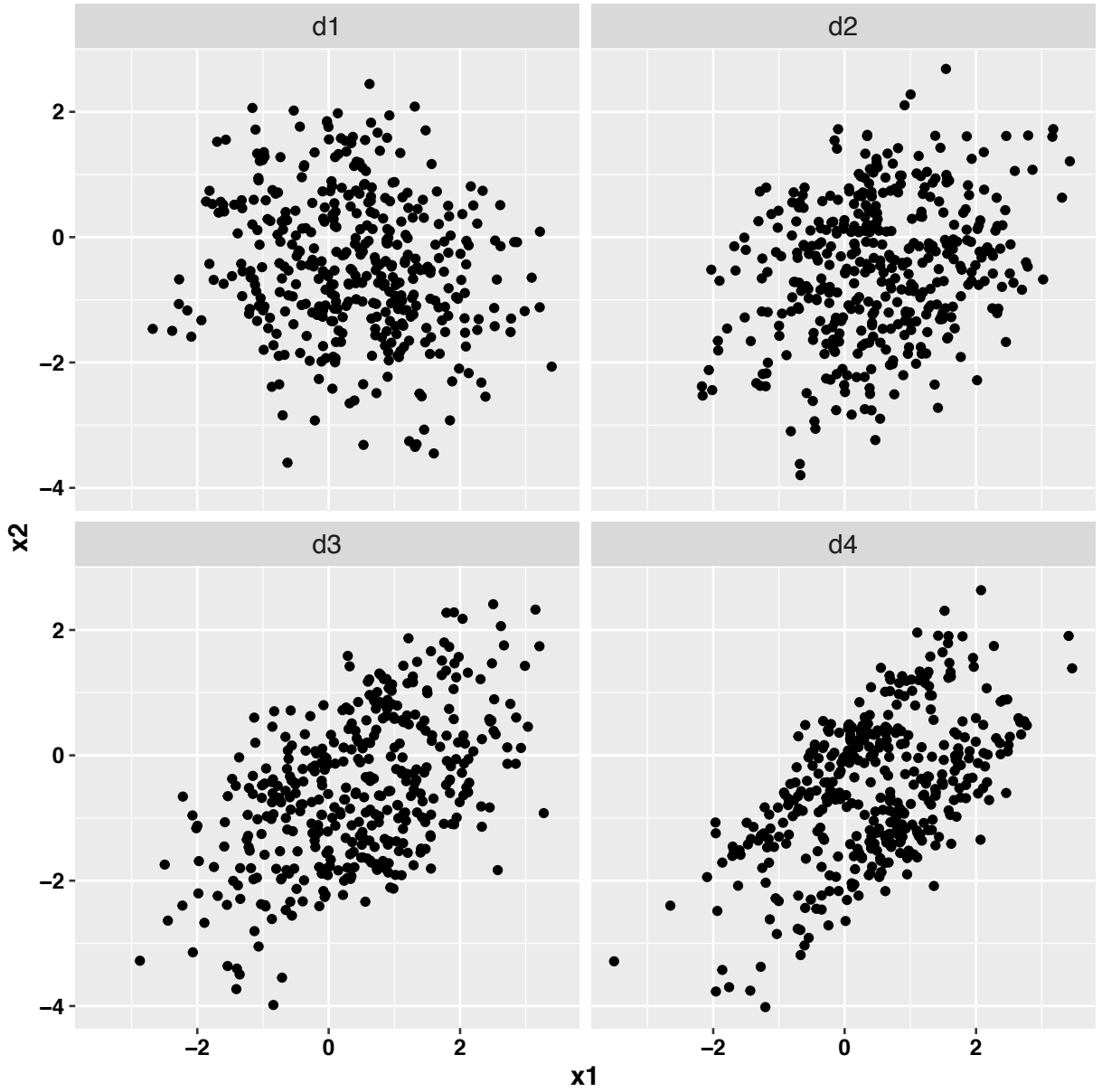
	Group1			Group2		
	μ_1	μ_2	ρ_1	μ_1	μ_2	ρ_2
d1	0	0	0	1	-1	0
d2	0	0	0.7	1	-1	0.7
d3	0	0	0.8	1	-1	0.8
d4	0	0	0.9	1	-1	0.9



Effect of Covariance Structure on Clustering

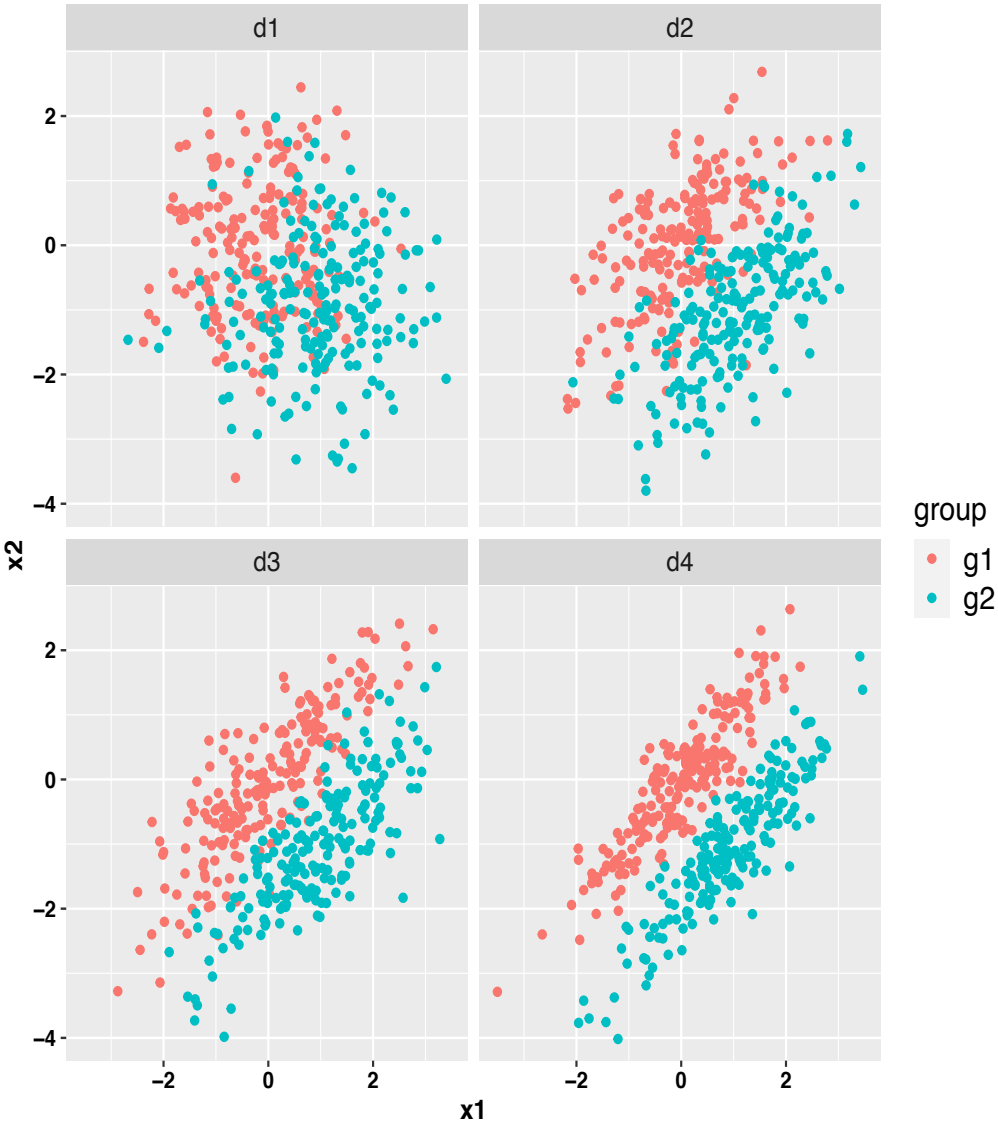
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

	Group1			Group2		
	μ_1	μ_2	ρ_1	μ_1	μ_2	ρ_2
d1	0	0	0	1	-1	0
d2	0	0	0.7	1	-1	0.7
d3	0	0	0.8	1	-1	0.8
d4	0	0	0.9	1	-1	0.9

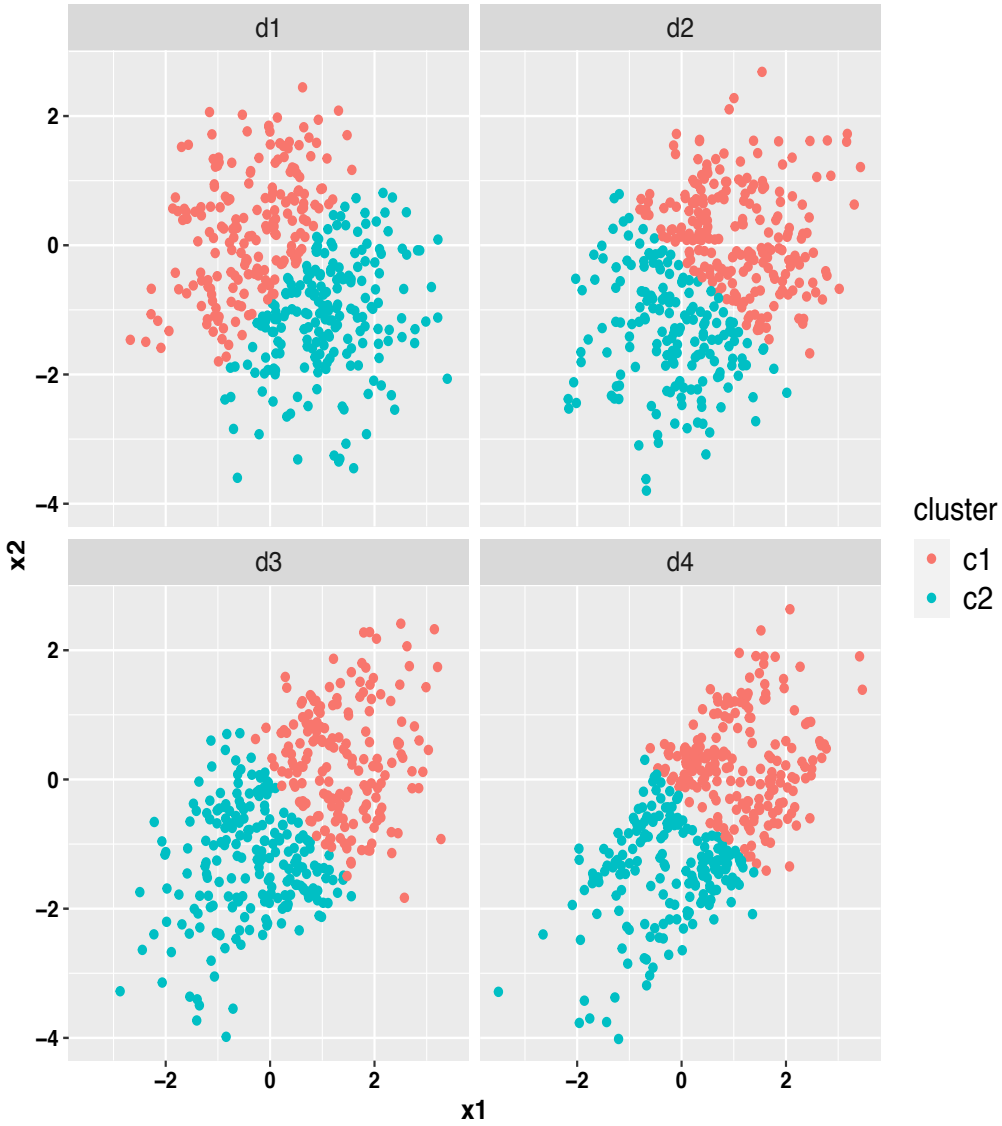


Effect of Covariance Structure on k-means Clustering

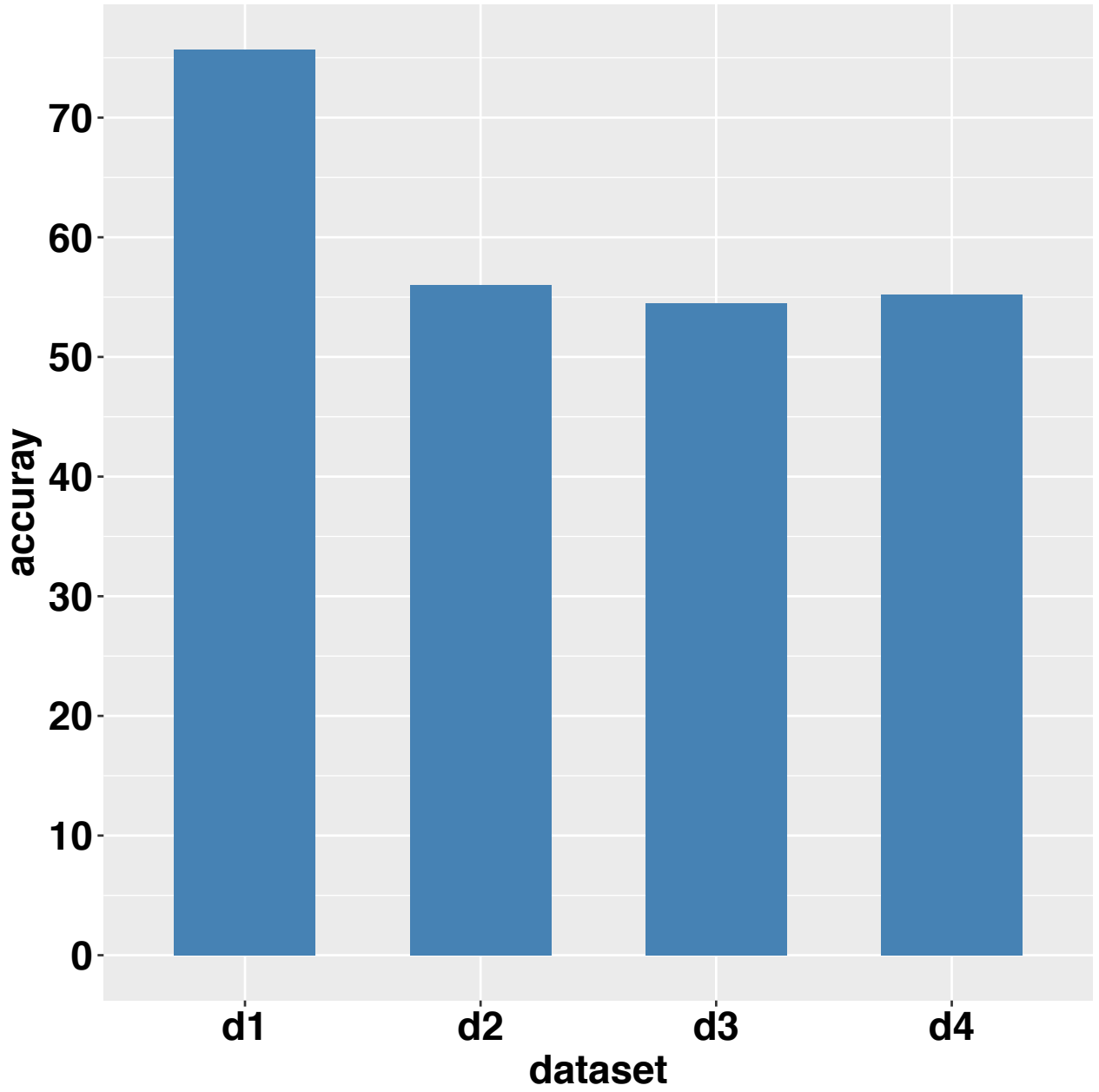
Label by group



Label by k-means cluster

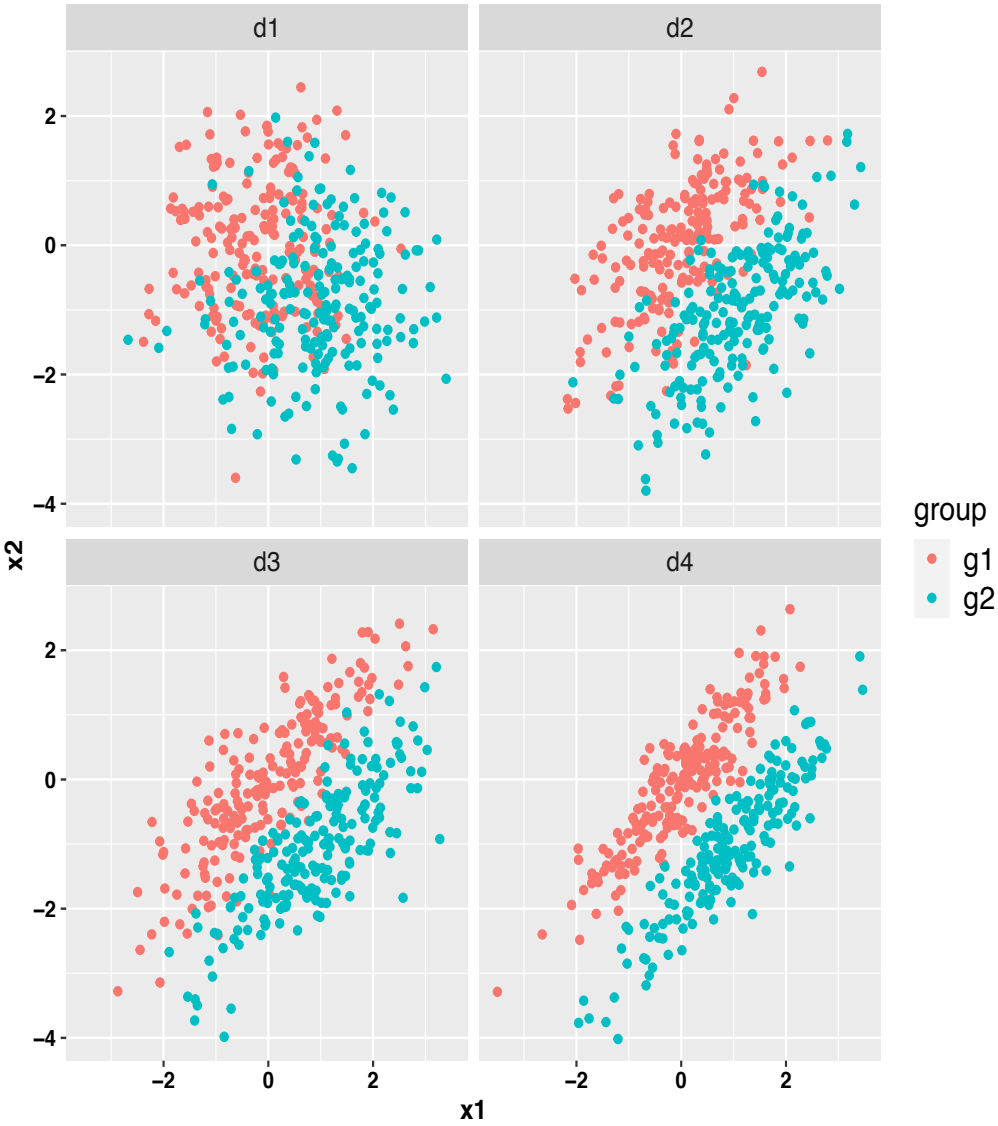


Accuracy of k-means Clustering Decreases as Covariance Increases

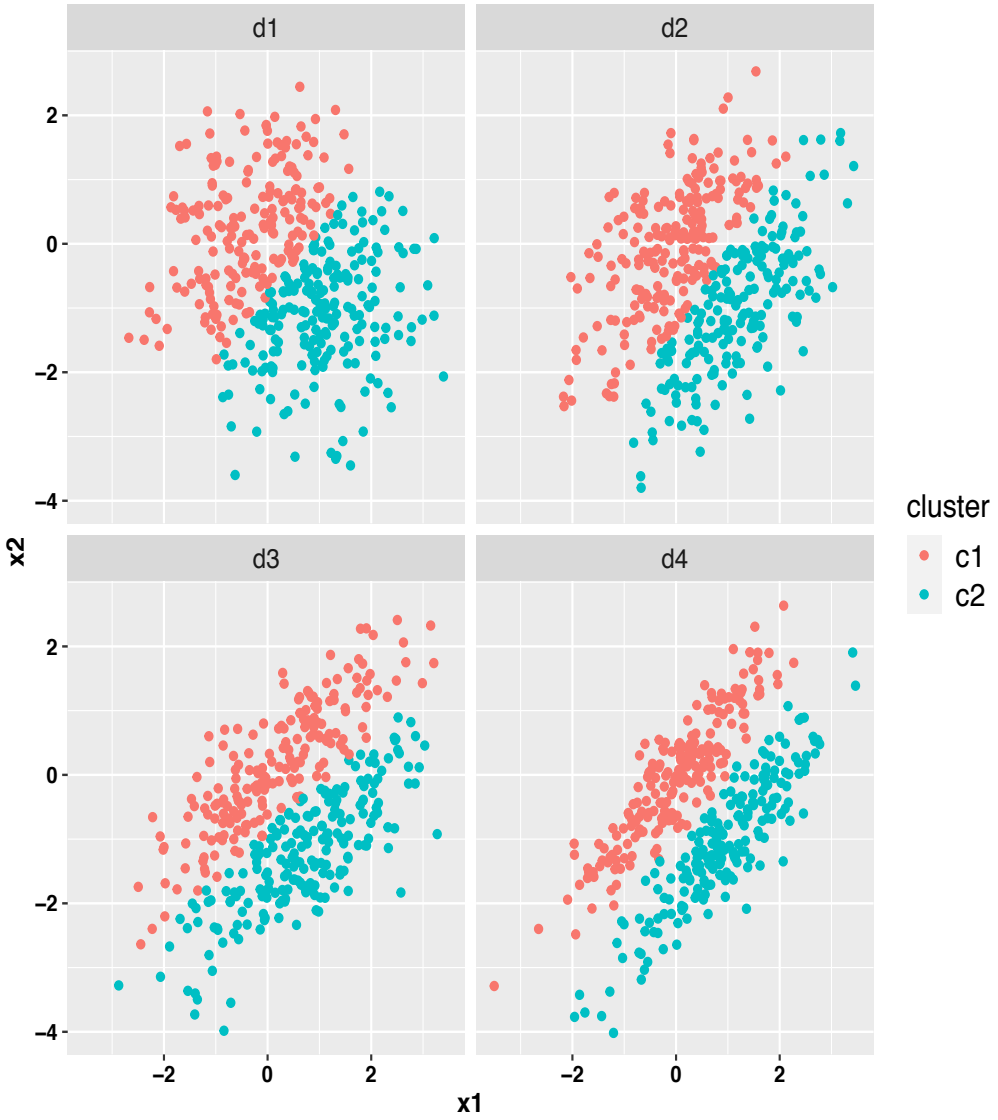


Effect of Covariance Structure on GMM Clustering

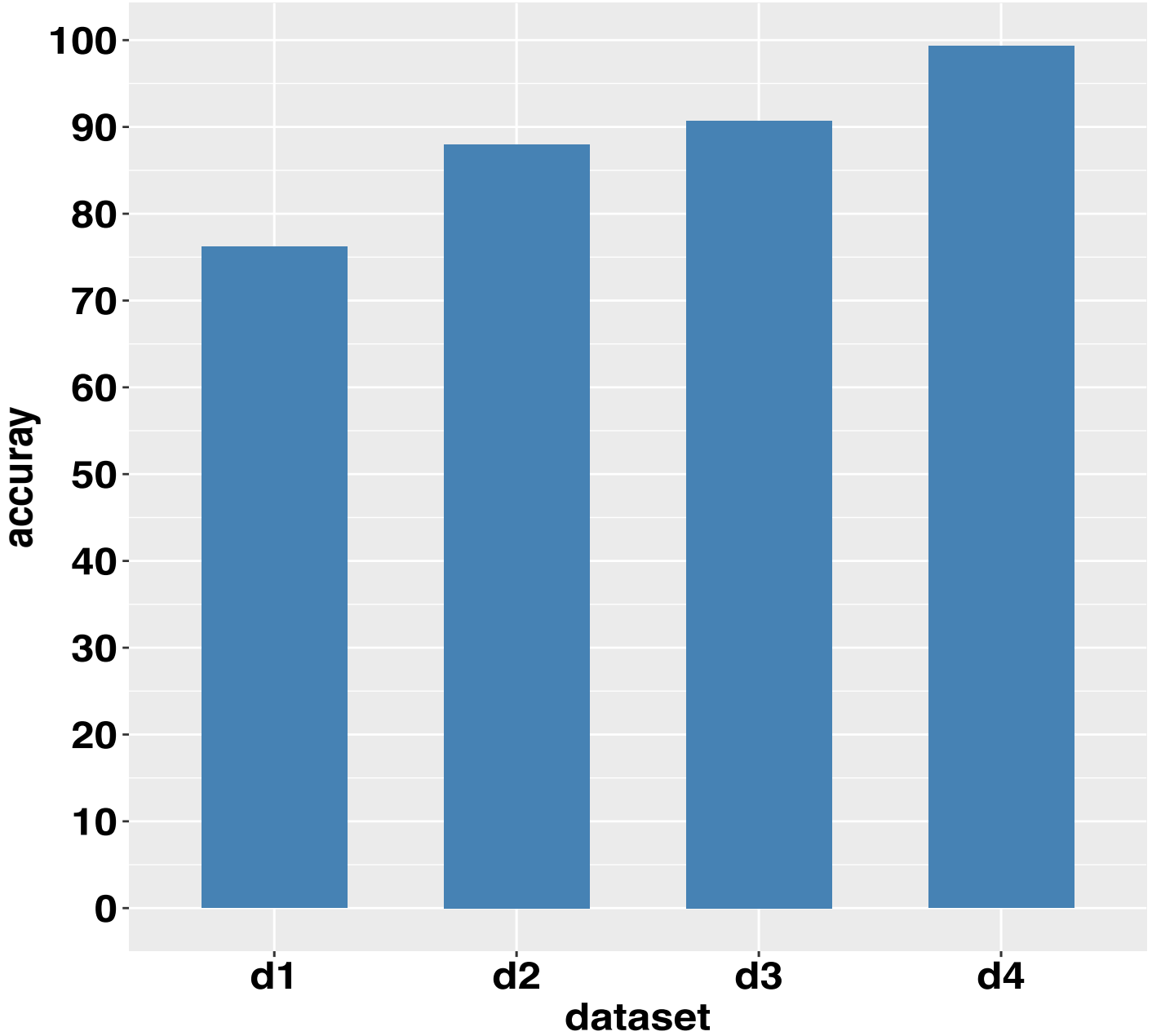
Label by group



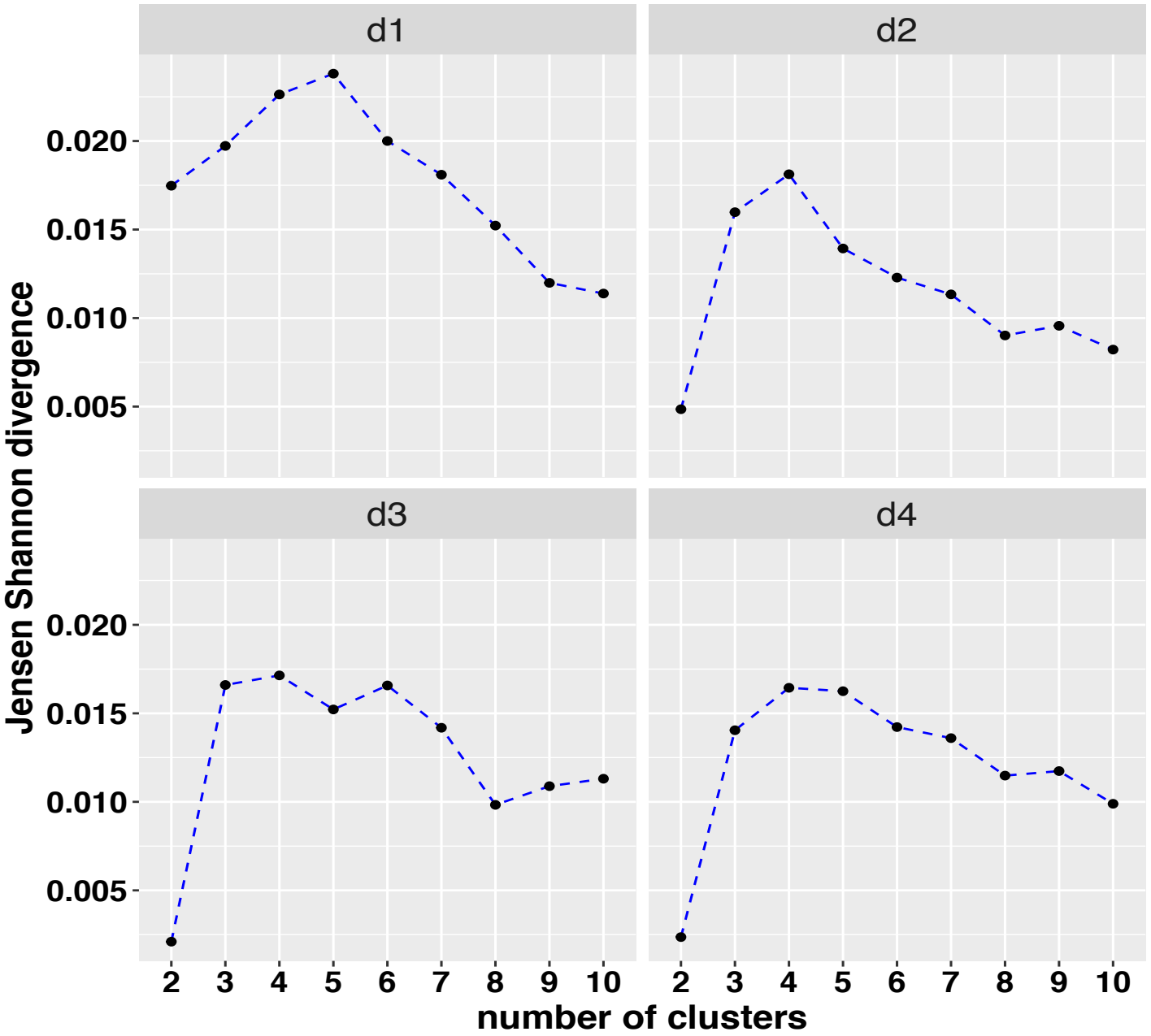
Label by GMM cluster



Accuracy of GMM Clustering Increases as Covariance Increases



Choose the Number of Clusters with Jensen-Shannon Divergence



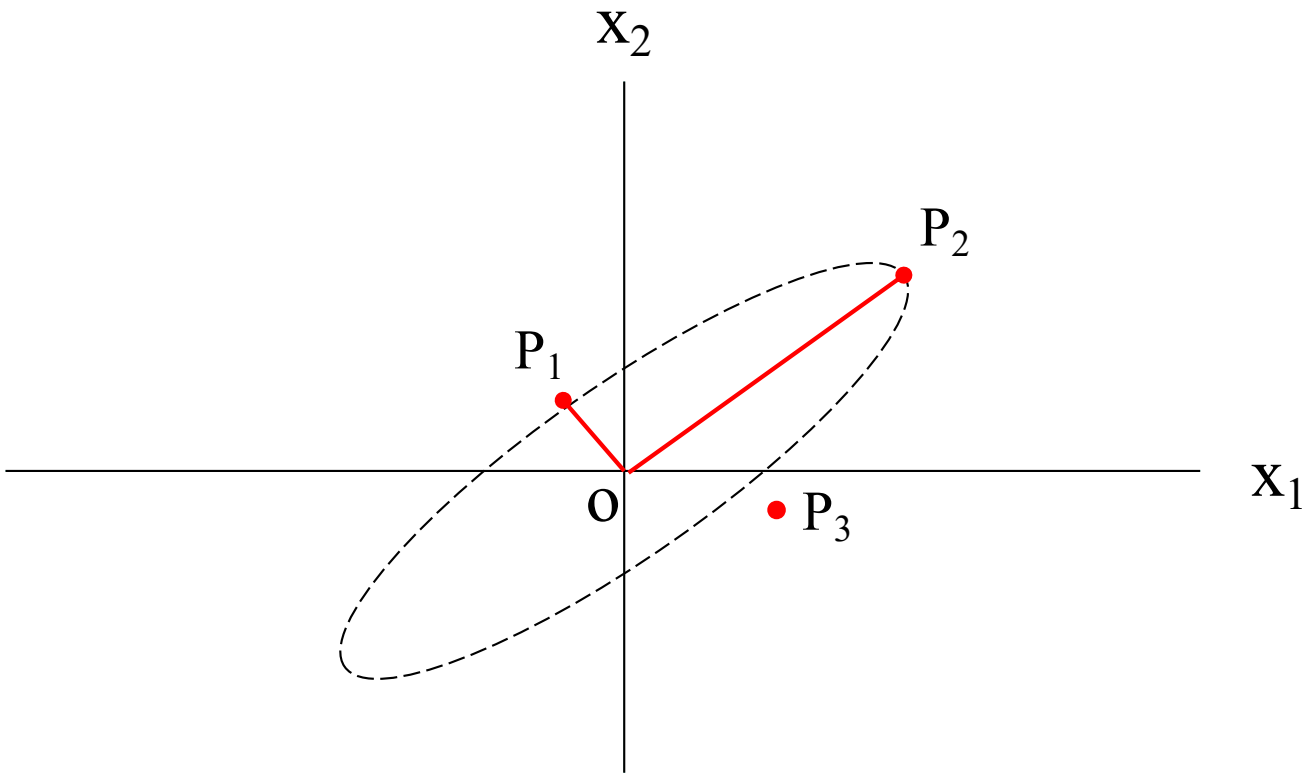
Jensen-Shannon Divergence

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M)$$

where $M = \frac{1}{2}(P + Q)$

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

Euclidean Distance vs. Mahalanobis Distance



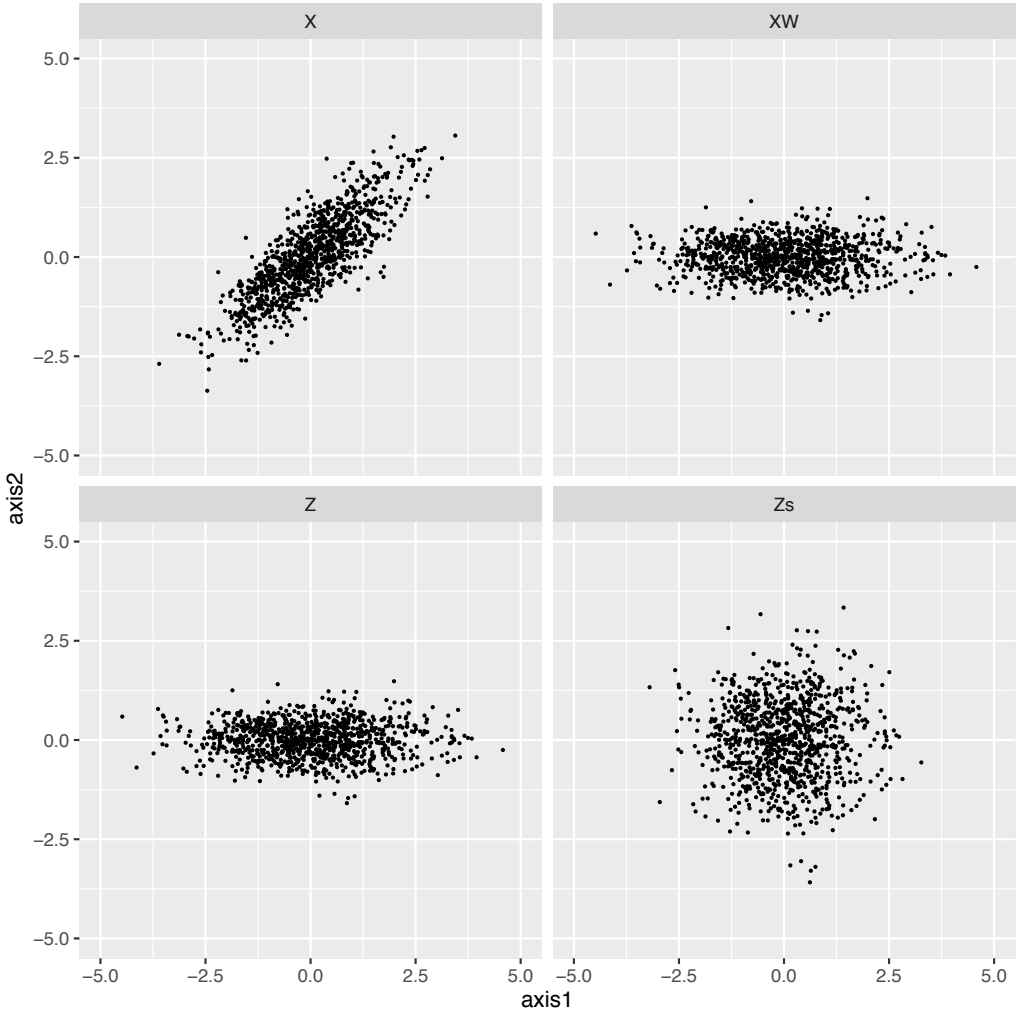
Euclidean distance: $P_1 < P_3 < P_2$

Probability: $p_1 = p_2 > p_3$

Mahalanobis distance is a statistical distance related to probability

Prasanta Chandra Mahalanobis in 1936

Multivariate Gaussian Distribution

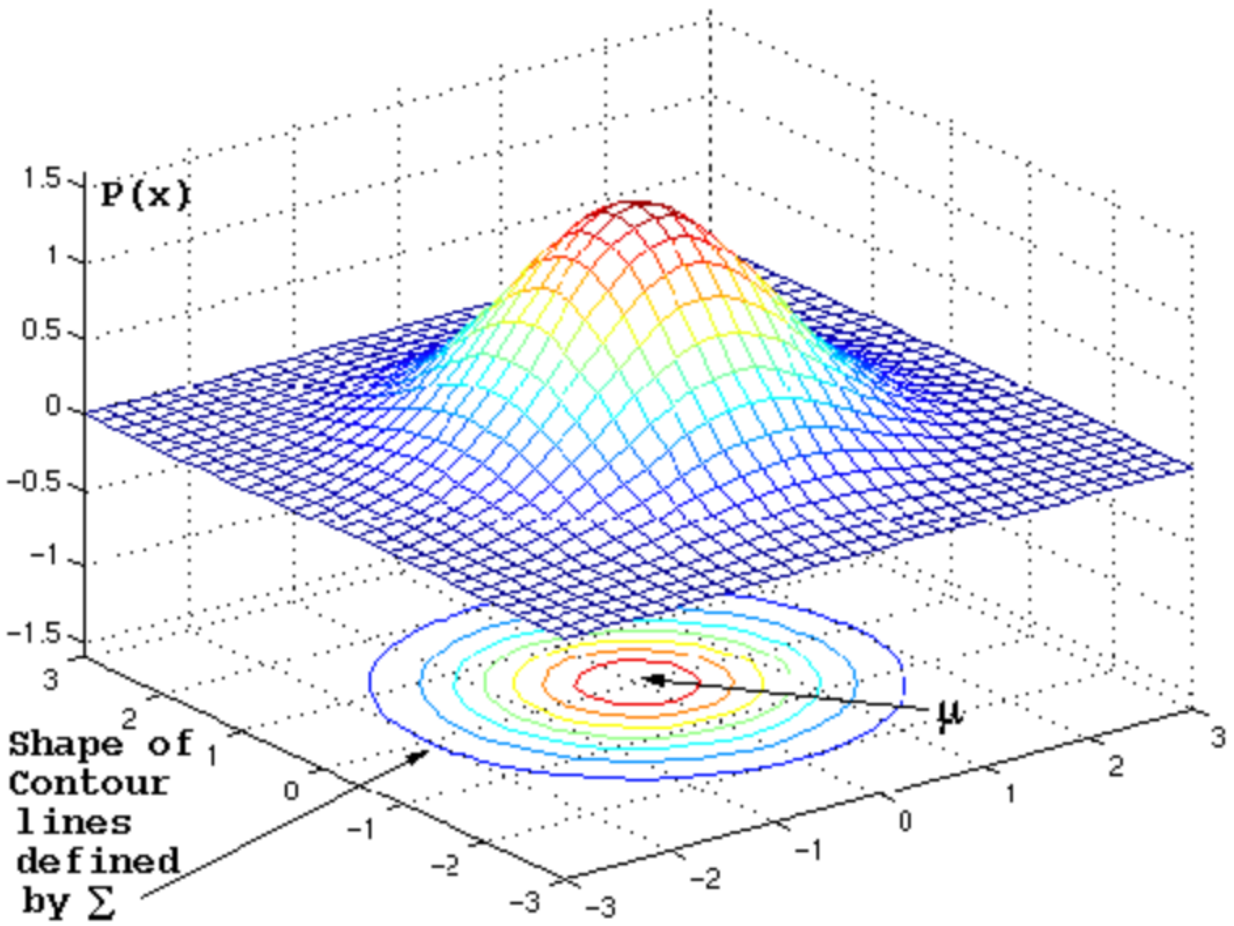


$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

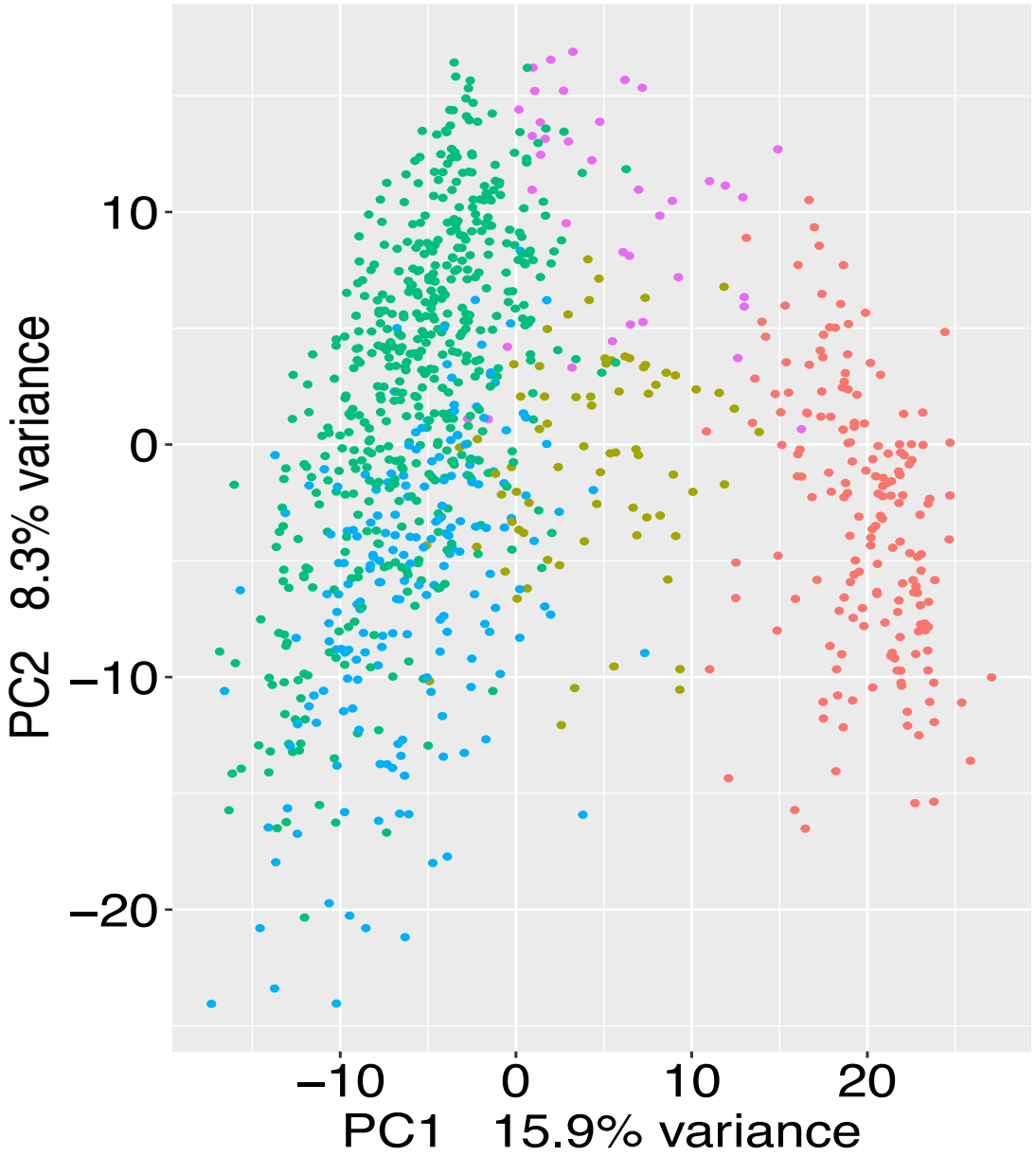
Σ: covariance matrix
Σ^{-1}: inverse of Σ
Λ: Diagonal matrix with Eigen values
W: Eigen vectors
Z: Principal Components
Z_s: Standardized Z
z: a sample from Z_s
T: Transposition
μ: mean vector

$$Z = XW$$
$$Z_s = XW\Lambda^{-1/2}$$
$$z = \Lambda^{-1/2}W^Tx$$
$$z^Tz = x^TW\Lambda^{-1/2}\Lambda^{-1/2}W^Tx$$
$$z^Tz = x^T\Sigma^{-1}x$$

Multivariate Gaussian Distribution



PCA of TCGA BRCA Samples

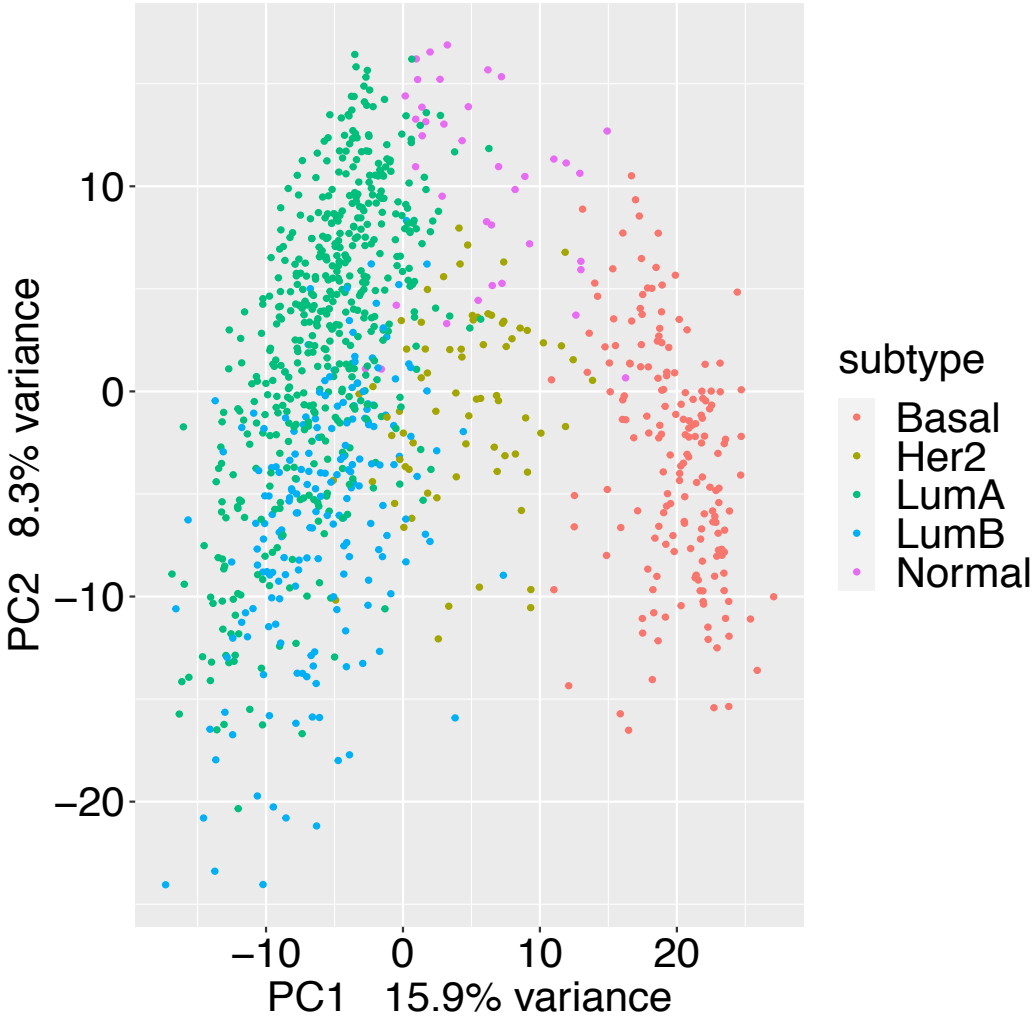


977 samples
5000 genes

x	freq
Basal	173
Her2	73
LumA	500
LumB	193
Normal	38

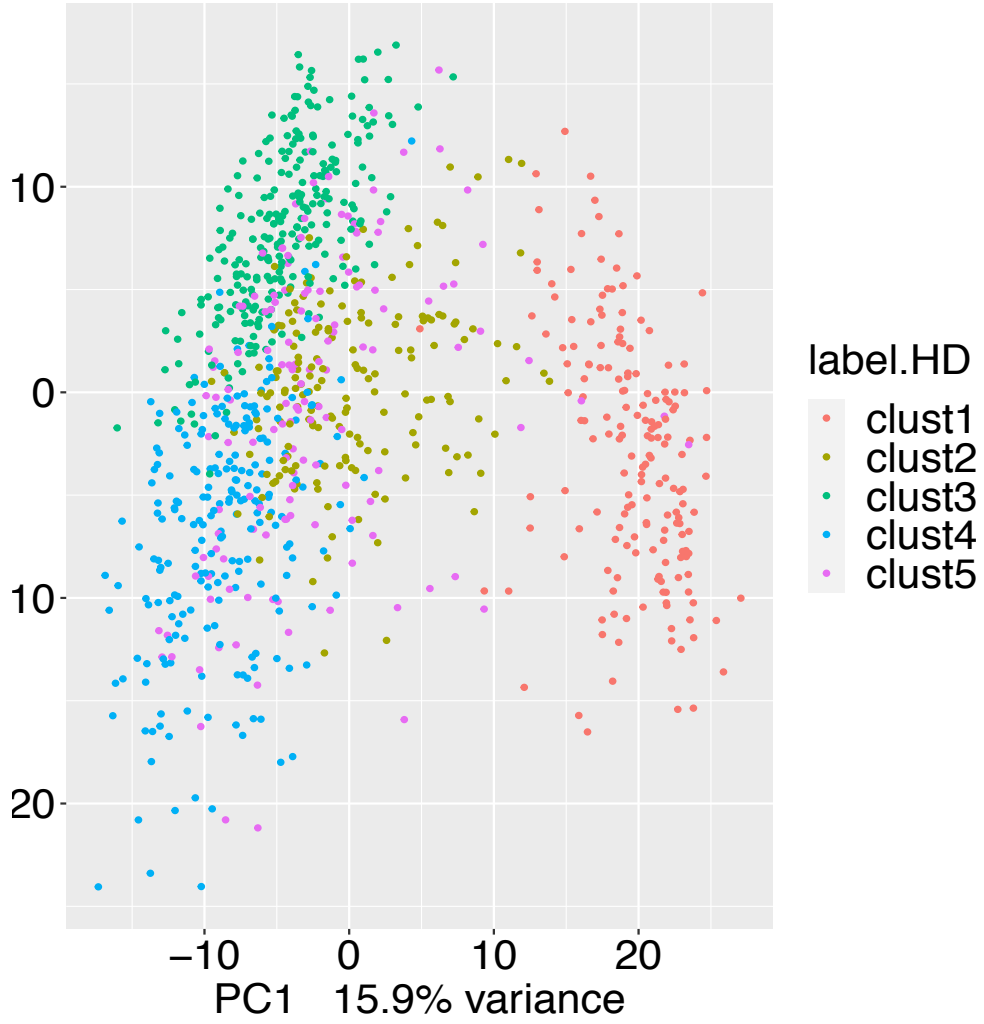
PCA: Label by Subtype vs. by GMM Cluster

Label by subtype



Label by GMM clusters in high-dimension

accuracy 60%

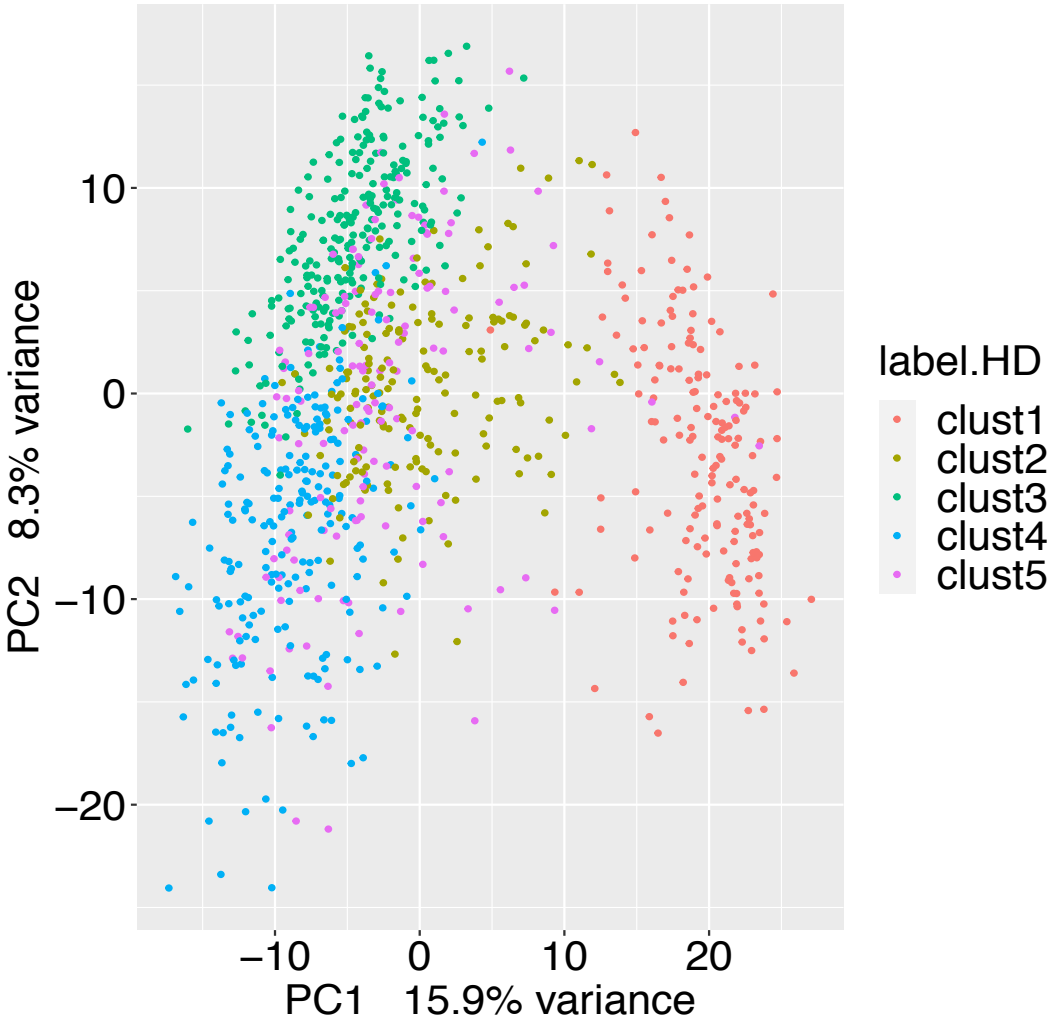


HD: high dimension, 5000 genes

PCA: Label by GMM Cluster vs. by k-means Cluster

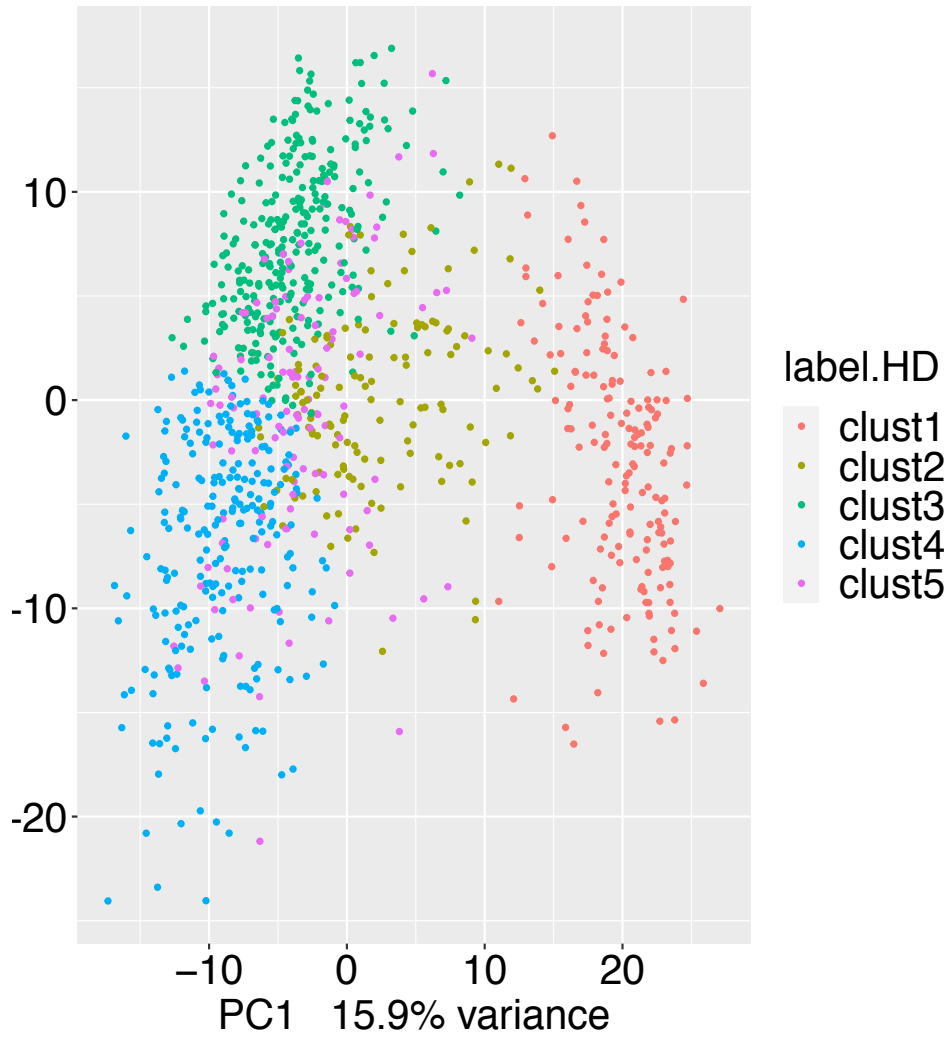
Label by GMM clusters in high-dimension

accuracy 60%



Label by k-means clusters in high-dimension

accuracy 65%



HD: high dimension, 5000 genes

Comparison Between Subtype and GMM vs. k-means Cluster (HD)

	Basal	Her2	LumA	LumB	Normal
clust1	168	1	1	0	6
clust2	2	59	**47**	**57**	8
clust3	0	0	**241**	6	15
clust4	0	3	112	106	1
clust5	3	10	99	24	8

GMM

Accuracy = (168 + 59 + 241 + 106 + 8) / 977 = 59.6%

Match
Mismatch

	Basal	Her2	LumA	LumB	Normal
clust1	169	0	0	0	6
clust2	4	69	**17**	**40**	5
clust3	0	0	**268**	11	21
clust4	0	0	125	119	0
clust5	0	4	90	23	6

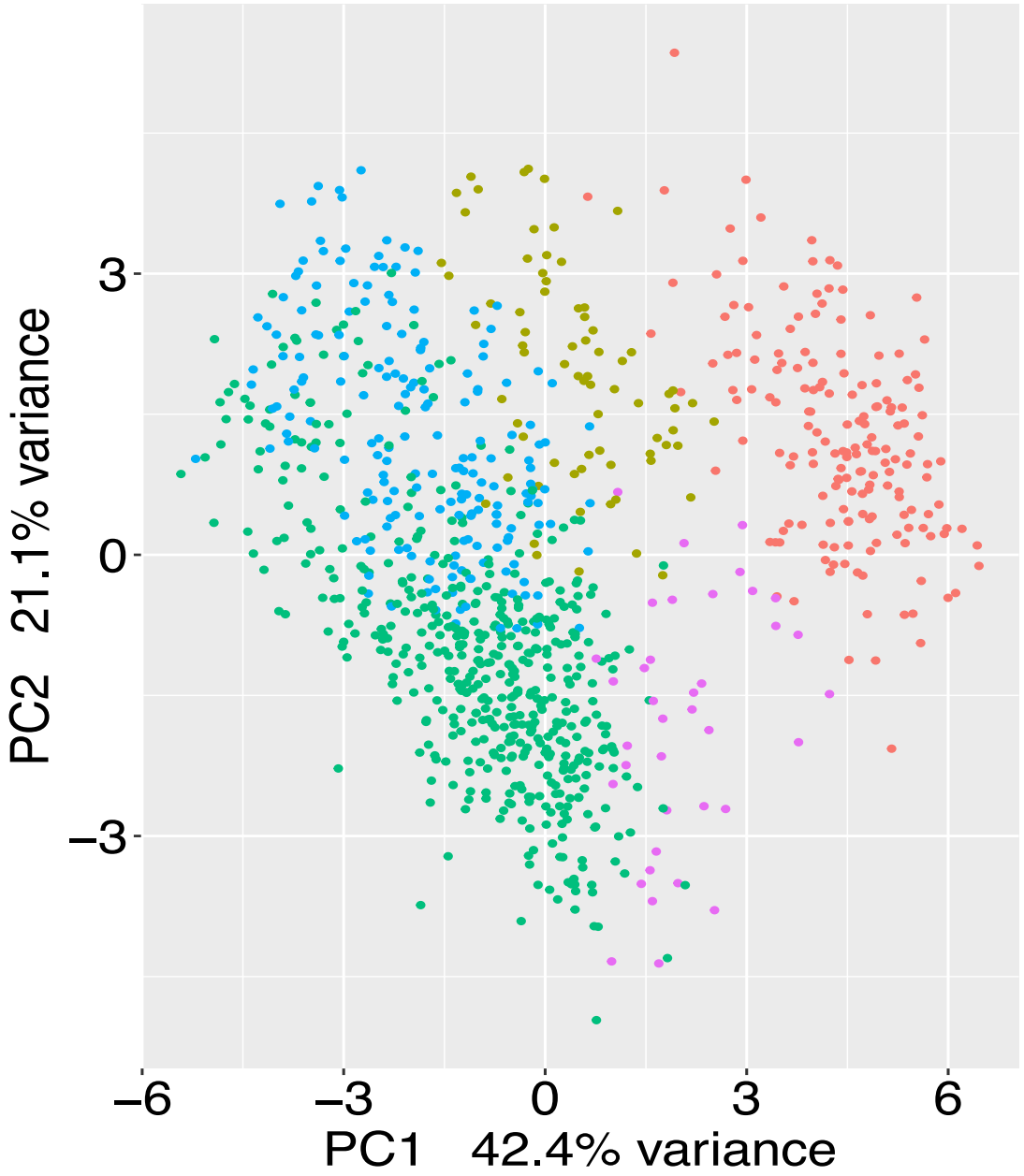
K-means

Accuracy = (169 + 69 + 268 + 119 + 6) / 977 = 64.6%

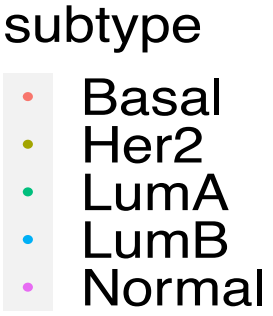
Potential Issues of GMM and k-means Clustering

1) Local maxima (MLE)
2) Incorrect data model
3) Curse of dimensionality
4) Data are not linearly separable

PCA of TCGA BRCA Samples with Pam50 Genes

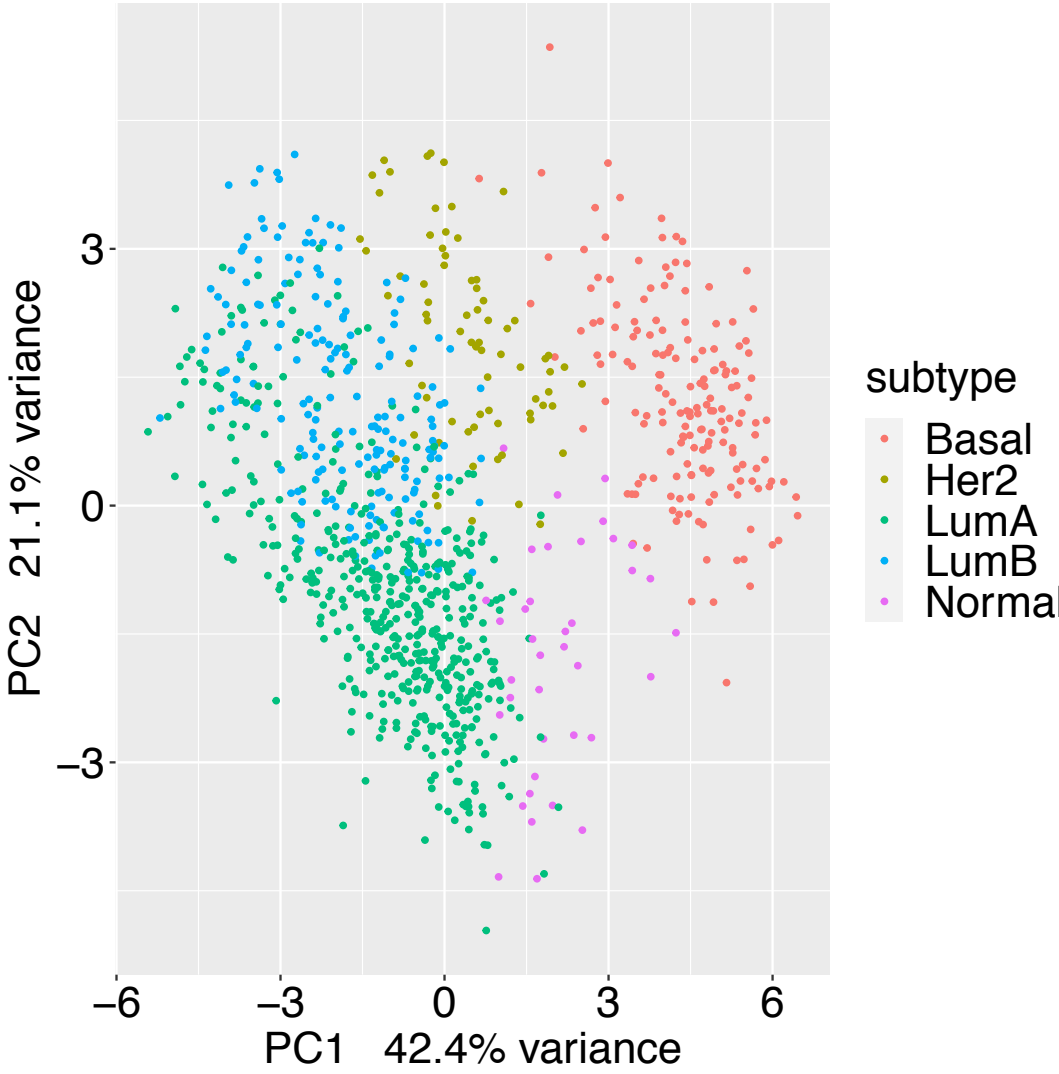


977 samples
39 genes



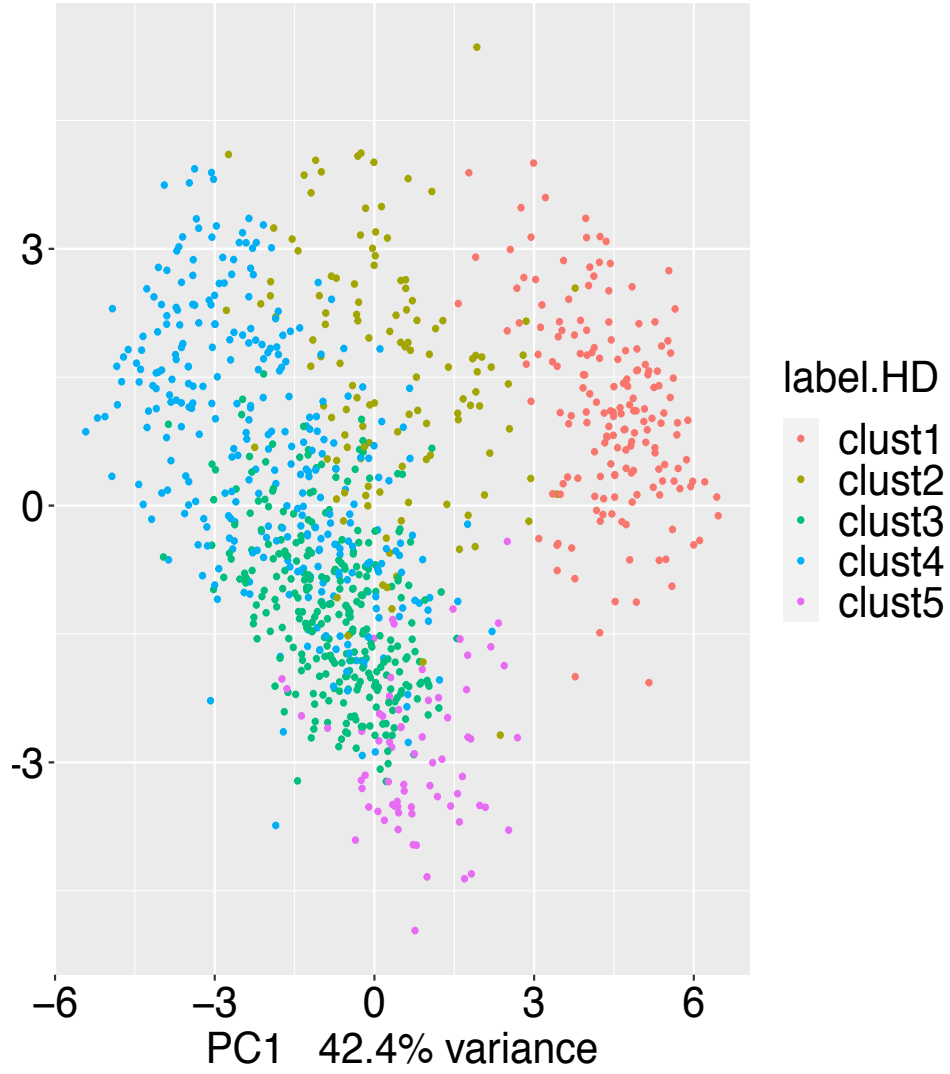
PCA with pam50: Label by Subtype vs. by GMM Clusters

Label by subtype



Label by GMM clusters in high-dimension

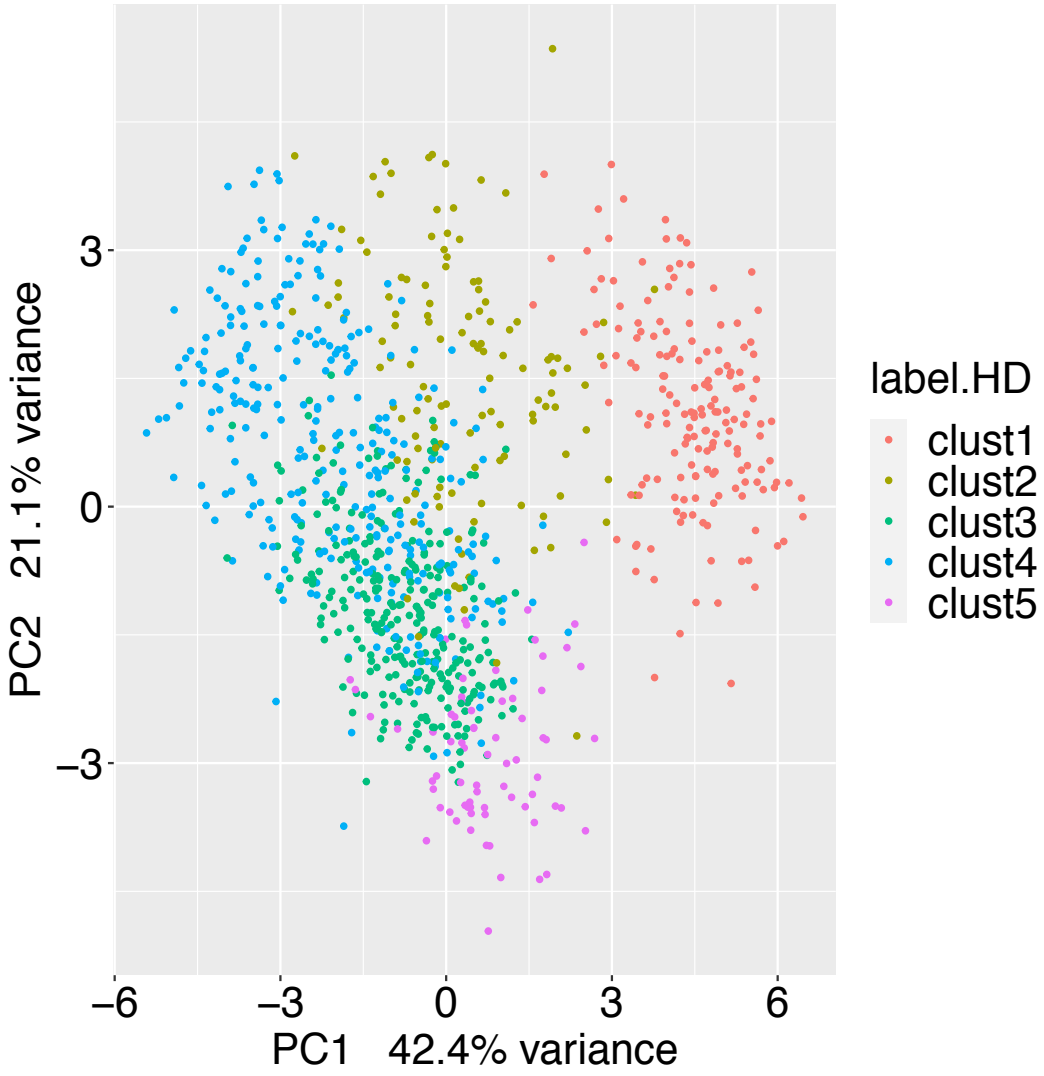
accuracy 65%



PCA with pam50: Label by GMM vs. k-means Clusters

Label by GMM clusters in high-dimension

accuracy 65%



Label by k-means clusters in high-dimension

accuracy 59%

Comparison Between Subtype and GMM vs. k-means Cluster (HD)

	Basal	Her2	LumA	LumB	Normal
clust1	165	0	0	0	6
clust2	8	70	19	22	6
clust3	0	0	**254**	40	2
clust4	0	3	175	**131**	5
clust5	0	0	**52**	0	19

GMM

Accuracy = (165 + 70 + 254 + 131 + 19) / 977 = 65.4%

Match
Mismatch

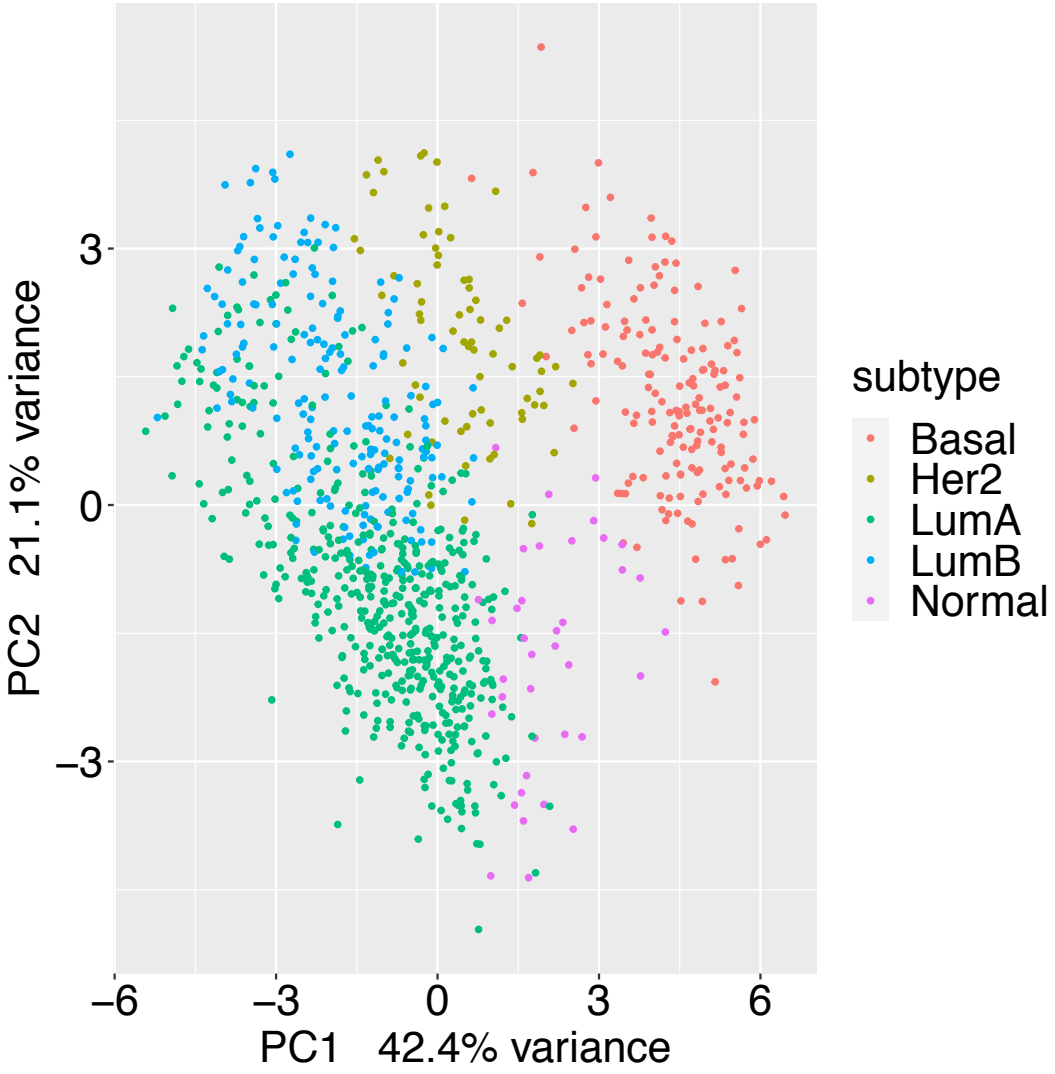
	Basal	Her2	LumA	LumB	Normal
clust1	170	0	0	0	8
clust2	2	72	11	31	4
clust3	0	0	**214**	76	0
clust4	1	1	82	**86**	0
clust5	0	0	**193**	0	26

K-means

Accuracy = (170 + 72 + 221 + 87 + 26) / 977 = 59%

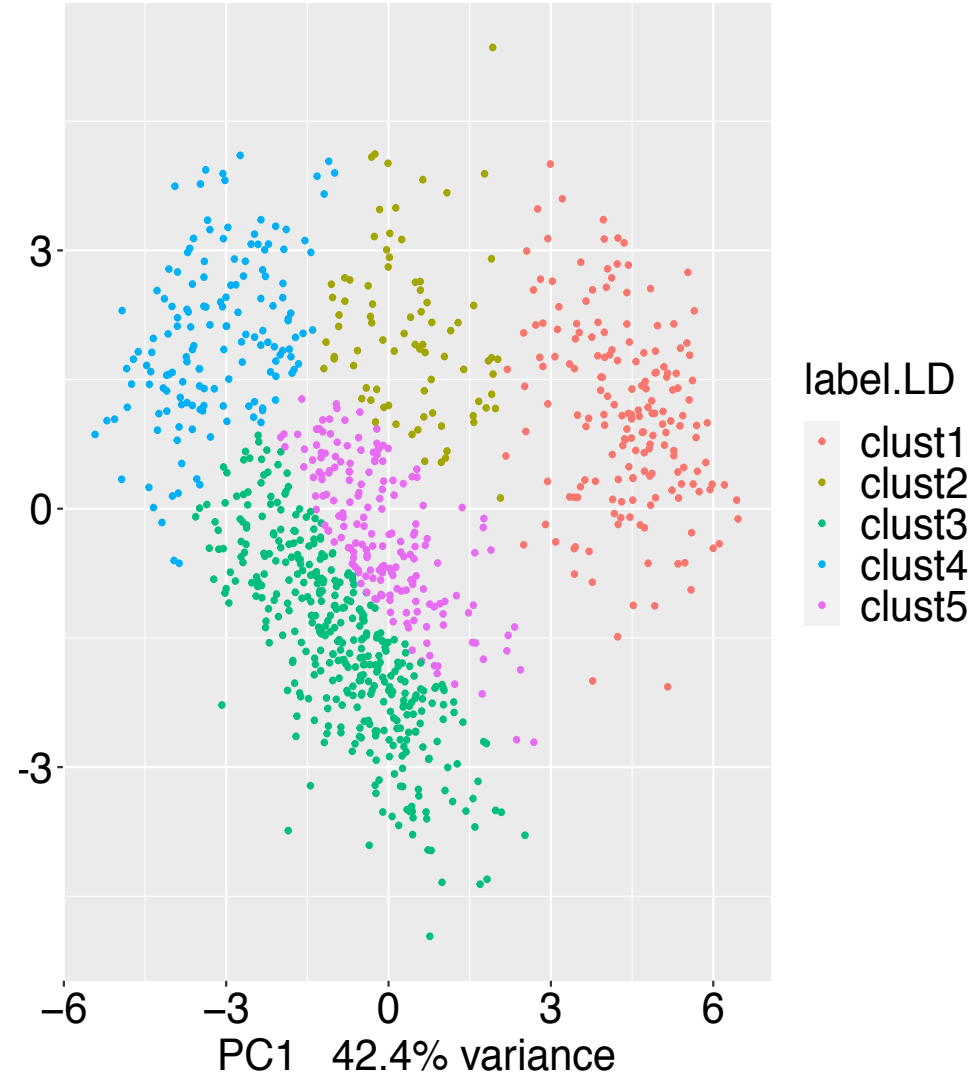
PCA with pam50: Label by Subtype vs. by GMM Clusters

Label by subtype



Label by GMM clusters with 2 PCs

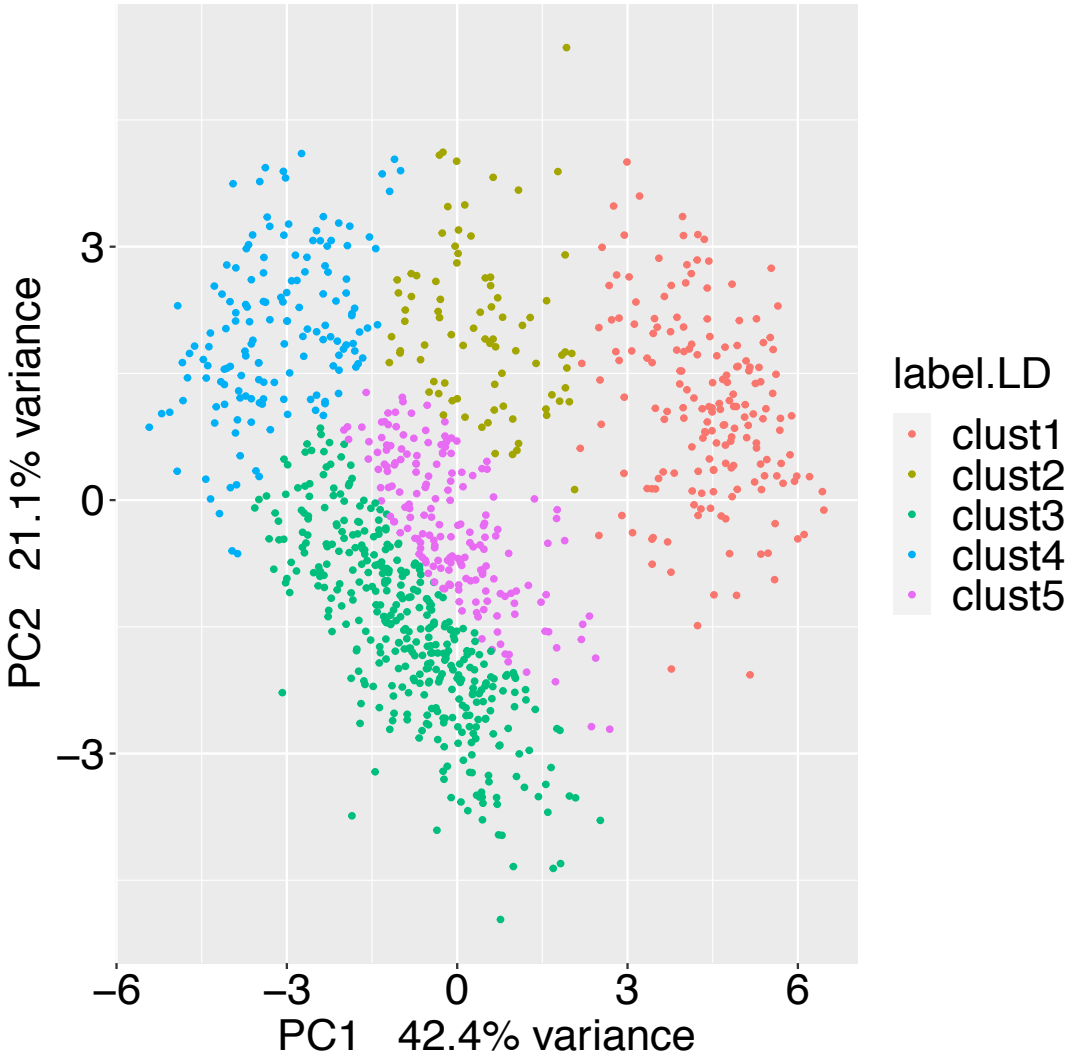
accuracy 67%



PCA with pam50: Label by GMM vs. k-means Clusters

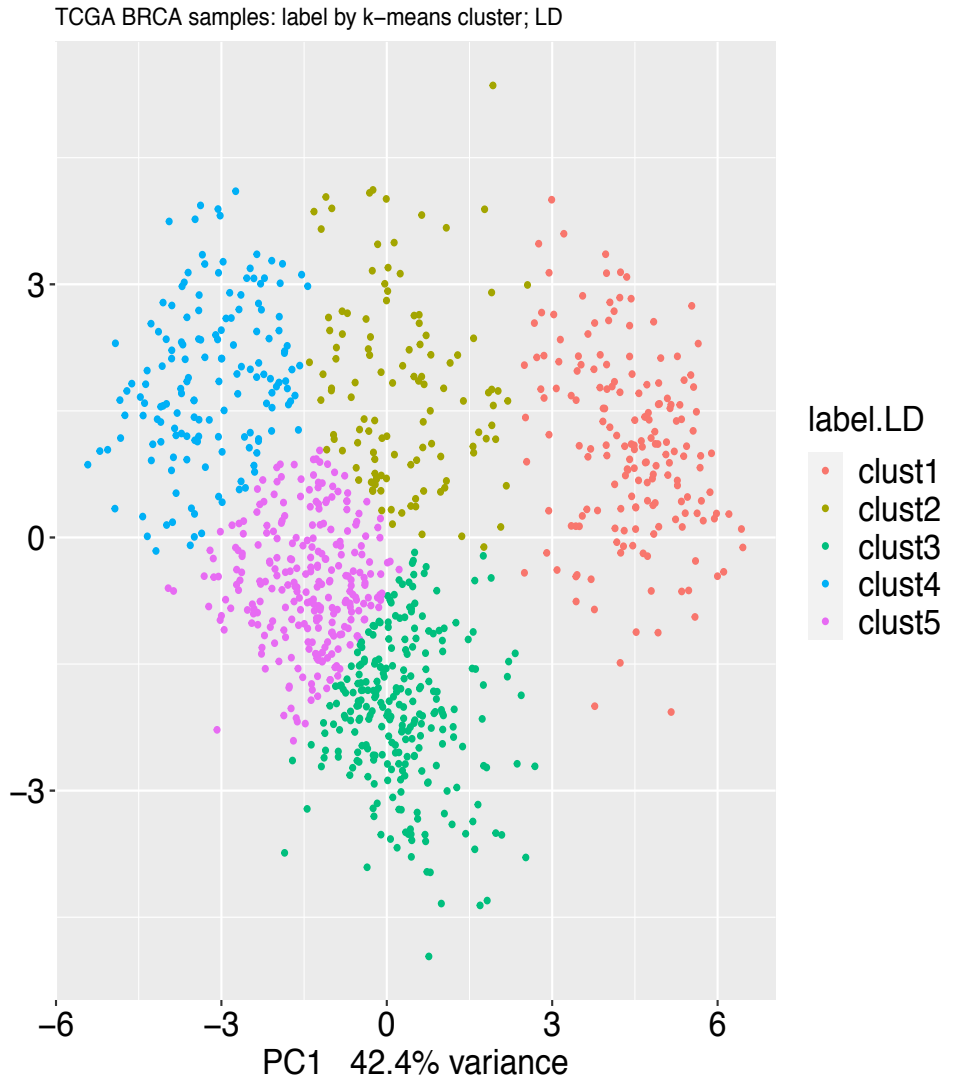
Label by GMM clusters with 2 PCs

accuracy 67%



Label by k-means clusters with 2 PCs

accuracy 56%



Comparison Between Subtype and GMM vs. k-means Cluster (LD)

	Basal	Her2	LumA	LumB	Normal
clust1	167	3	0	0	9
clust2	6	55	0	19	2
clust3	0	0	**336**	34	11
clust4	0	6	64	81	0
clust5	0	9	**100**	59	16

GMM

Accuracy = (167 + 55 + 336 + 81 + 16) / 977 = 67%

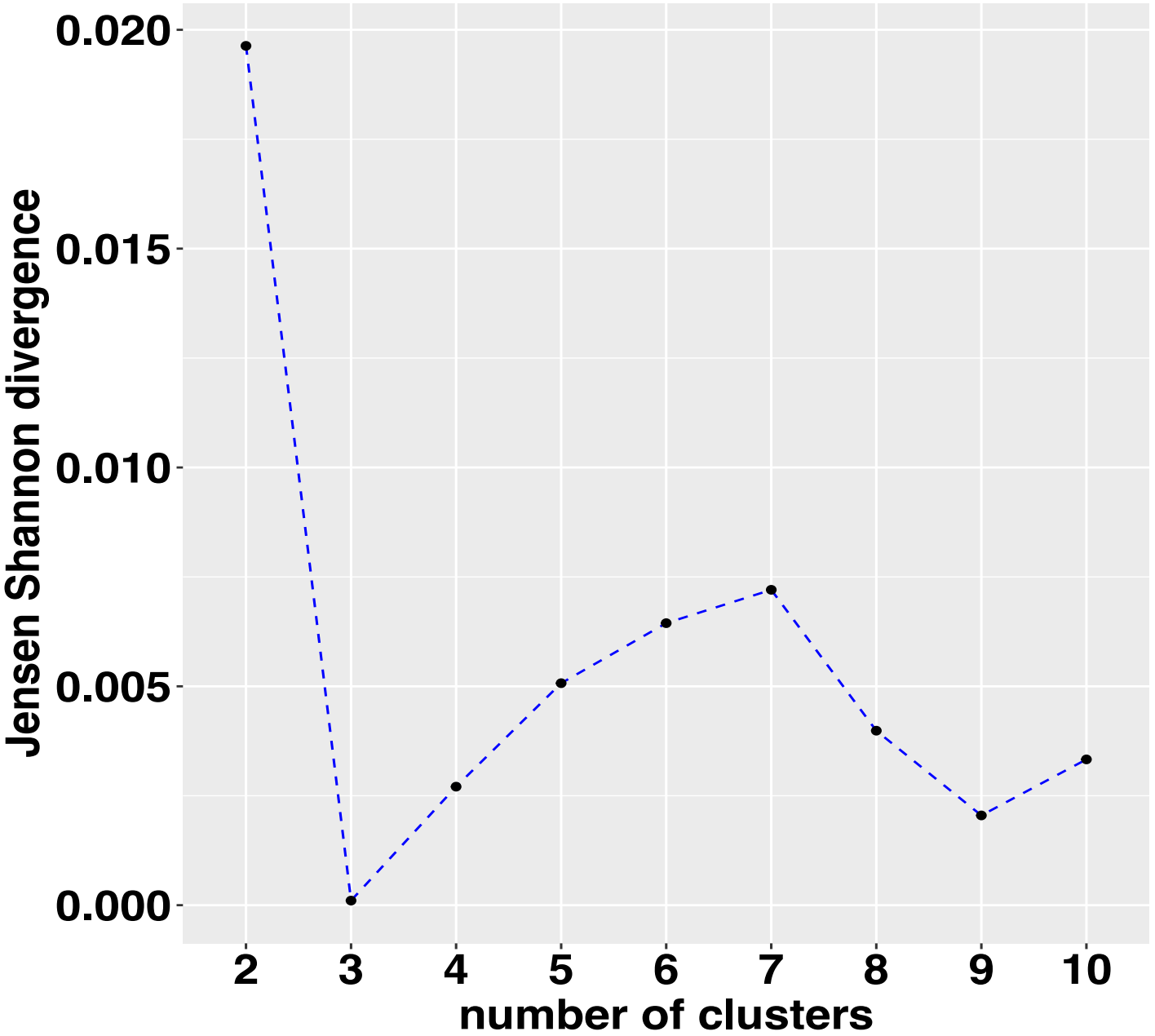
Match
Mismatch

	Basal	Her2	LumA	LumB	Normal
clust1	166	1	0	0	9
clust2	7	65	11	33	2
clust3	0	2	**227**	1	27
clust4	0	2	66	87	0
clust5	0	3	**196**	72	0

K-means

Accuracy = (166 + 65 + 227 + 87) / 977 = 55.7%

Jensen-Shannon Divergence vs. Number of Cluster (LD)



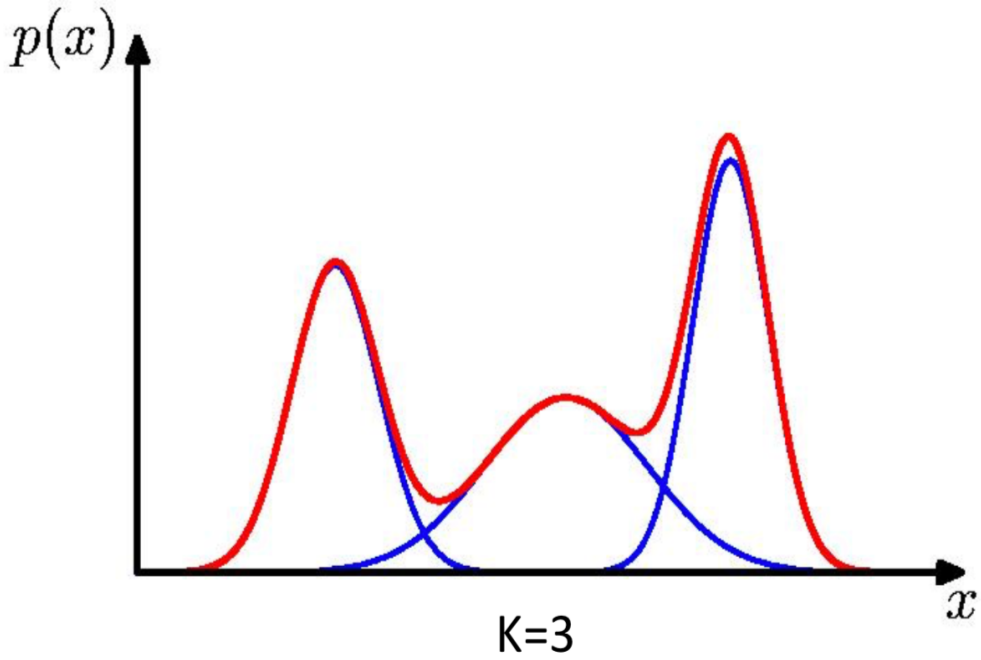
Mixture of Univariate Gaussian Distribution

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Component

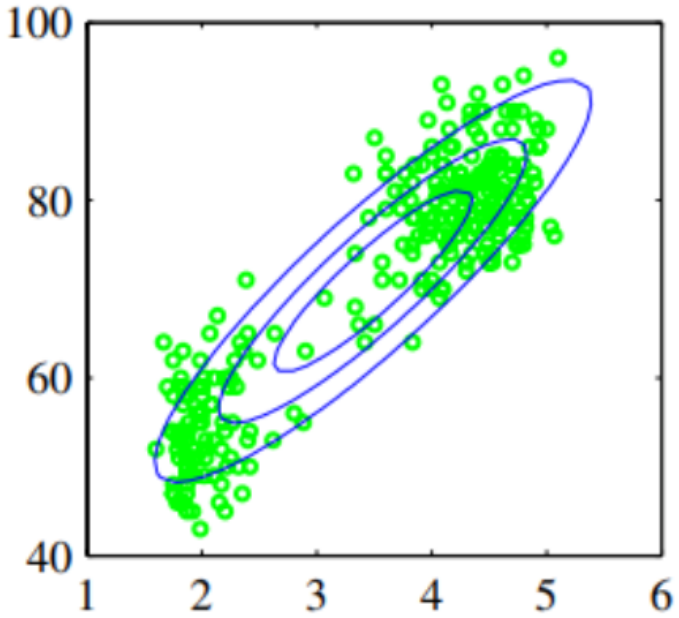
Mixing coefficient

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$



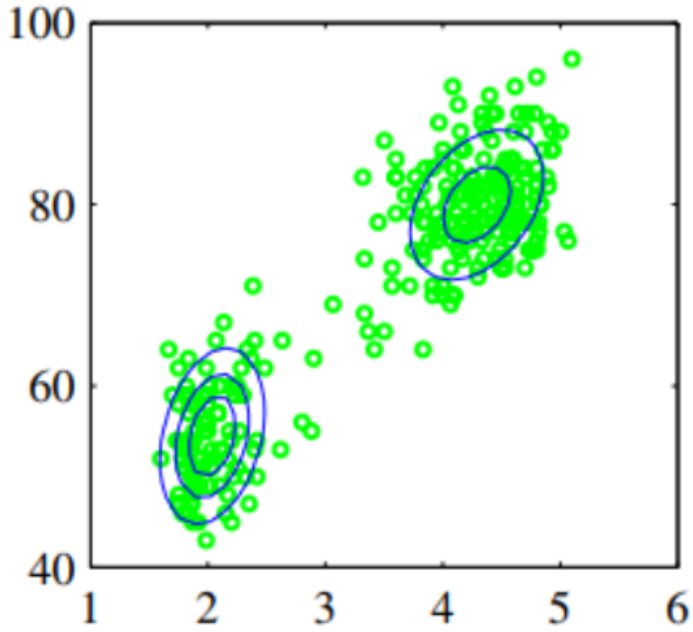
Mixture of Bivariate Gaussian Distributions

Single Gaussian



$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Mixture of two Gaussians



$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Component

Mixing coefficient

Algorithm of GMM: Maximal Likelihood Estimate

$$\prod_{j=1}^{m}\sum_{k=1}^{K}\frac{1}{(2\pi)^{m/2}\,\|\Sigma_k\|^{1/2}}\exp\left[-\frac{1}{2}\left(\mathbf{x}_j-\mu_k\right)^T\Sigma_k^{-1}\left(\mathbf{x}_j-\mu_k\right)\right]P(y=k)$$

m observations
k mixture model of Gaussian distributions
μ_k is centroid coordinate of kth cluster
Σ_k is covariance matrix of kth cluster
P(y=k) is the probability of observation y as a member of cluster k

$\left(\mathbf{x}_j-\mu_k\right)^T\Sigma_k^{-1}\left(\mathbf{x}_j-\mu_k\right)$ is Mahalanobis distance

Algorithm of GMM: Expectation Maximization (EM)

Initialization: initialize k centroids with hierarchical clustering or k-means or random points

Alternating between the following two steps until converge

E step: compute expected probability of each datapoint as a member for each class (soft assignment)

$$\mathrm{P}\left(Y_j = k \middle| x_j, \lambda_t\right) \propto p_k^{(t)} \mathrm{p}\left(x_j \middle| \mu_k^{(t)}, \Sigma_k^{(t)}\right)$$

M step: update Gaussian distribution parameters for each class

$$\mu_k^{(t+1)} = \frac{\sum_j \mathrm{P}\left(Y_j = k \middle| x_j, \lambda_t\right) x_j}{\sum_j \mathrm{P}\left(Y_j = k \middle| x_j, \lambda_t\right)} \qquad \Sigma_k^{(t+1)} = \frac{\sum_j \mathrm{P}\left(Y_j = k \middle| x_j, \lambda_t\right) \left[x_j - \mu_k^{(t+1)}\right]\left[x_j - \mu_k^{(t+1)}\right]^T}{\sum_j \mathrm{P}\left(Y_j = k \middle| x_j, \lambda_t\right)}$$

$$\lambda_t = \{ \mu_1^{(t)}, \mu_2^{(t)} \dots \mu_K^{(t)}, \Sigma_1^{(t)}, \Sigma_2^{(t)} \dots \Sigma_K^{(t)}, p_1^{(t)}, p_2^{(t)} \dots p_K^{(t)} \}$$

Comparison Between GMM and k-means Clustering

Initialization: initialize k centroids with hierarchical clustering or k-means

Alternating between the following two steps until converge

E step: compute expected probability of each datapoint as a member for each class

hard assignment

$$\mathrm{P}\left(Y_j = k \middle| x_j, \lambda_t\right) \propto p_k^{(t)} \mathrm{p}\left(x_j \middle| \mu_k^{(t)}, \Sigma_k^{(t)}\right)$$

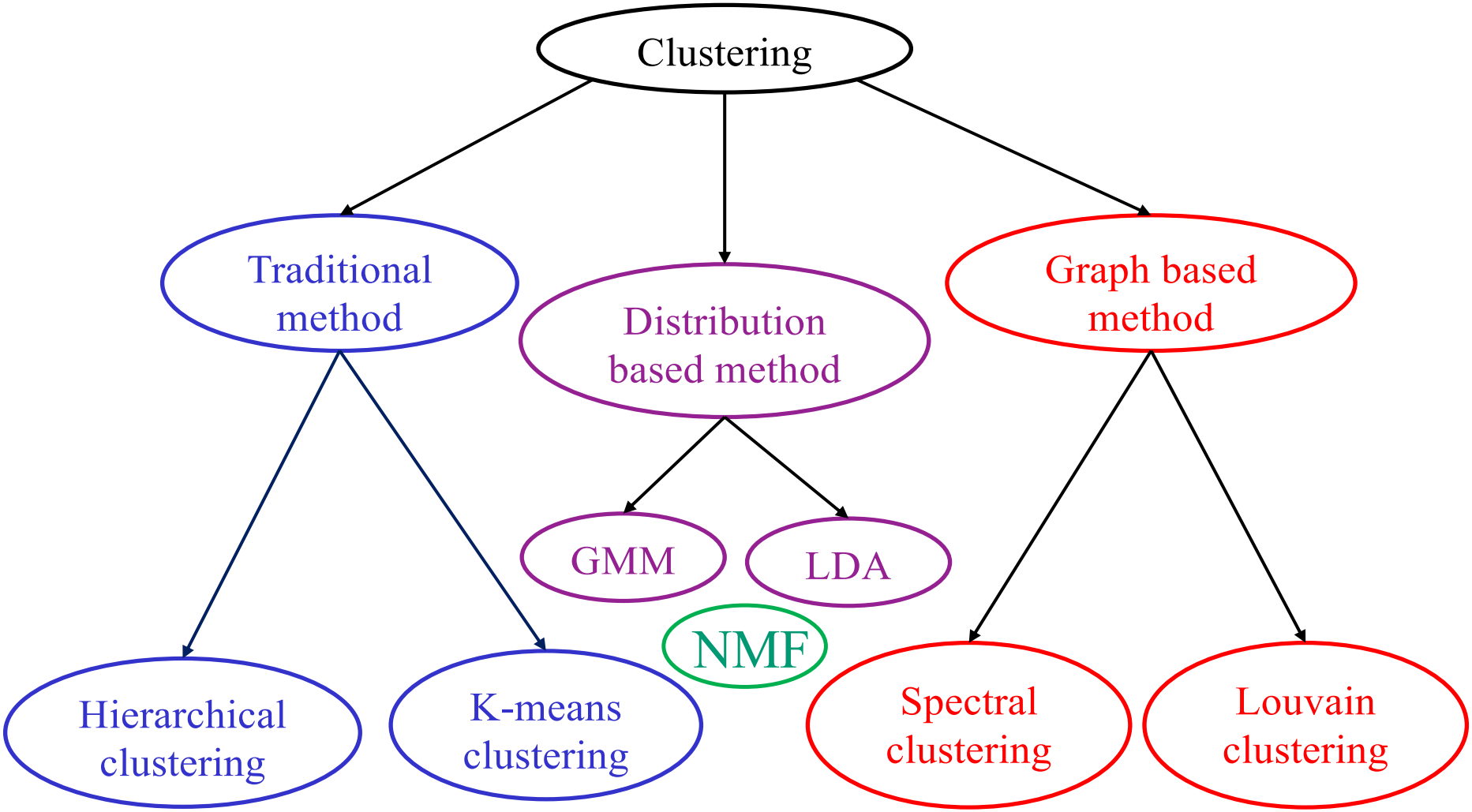
M step: update Gaussian distribution parameters for each class

$$\mu_k^{(t+1)} = \frac{\sum_j \mathrm{P}\left(Y_j = k \middle| x_j, \lambda_t\right) x_j}{\sum_j \mathrm{P}\left(Y_j = k \middle| x_j, \lambda_t\right)}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_j \mathrm{P}\left(Y_j = k \middle| x_j, \lambda_t\right) \left[x_j - \mu_k^{(t+1)}\right]\left[x_j - \mu_k^{(t+1)}\right]^T}{\sum_j \mathrm{P}\left(Y_j = k \middle| x_j, \lambda_t\right)}$$

$$\lambda_t = \{ \mu_1^{(t)}, \mu_2^{(t)} \dots \mu_K^{(t)}, \Sigma_1^{(t)}, \Sigma_2^{(t)} \dots \Sigma_K^{(t)}, p_1^{(t)}, p_2^{(t)} \dots p_K^{(t)} \}$$

Outline of Clustering Methods



GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization

Contributed by Emily Tai