Dimension Reduction Methods:
From PCA To TSNE And UMAP

Maxwell Lee

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

April 16, 2020

Outline Of The Talk

1) **Linear dimension reduction methods**
PCA, MDS, and SVD

2) **Nonlinear dimension reduction methods**
Isomap, LLE, Laplacian Eigenmap, TSNE, and UMAP

3) **Canonical correlation and Trajectory analysis**
Data integration and reversed graph embedding

Data Matrix (Table)

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ . & & & . \\ . & & & . \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

$$X_{np}$$

n observations and p variables

Multivariate Linear Regression Model

y is response variable or
dependent variable

$x_1 \ldots x_p$ are independent variables

$$
\left[\begin{array}{c|cccc}
y_1 & x_{11} & x_{12} & \ldots & x_{1p} \\
y_2 & x_{21} & x_{22} & \ldots & x_{2p} \\
. & & & & . \\
. & & & & . \\
y_n & x_{n1} & x_{n2} & \ldots & x_{np}
\end{array}\right]
$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_p + \varepsilon$$

$$y = X\beta + \varepsilon$$

Application Of Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y	x	application
Tumor size	Gene expression	correlation
Gene expression	Treatment vs control	t-test
Treatment response	Gene expression	Classification (glm)

Unsupervised Analysis

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ . & & & . \\ . & & & . \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

- We do not have data for response variable y or sample label
- We are more interested in intrinsic relationship among samples

Supervised And Unsupervised Statistical Learning

supervised

continuous variable

Regression model
linear regression, polynomial regression, glm

categorical variable

Classification
Logistic regression, LDA, random forest, gbm

unsupervised

Clustering analysis
hierarchical clustering, k-means

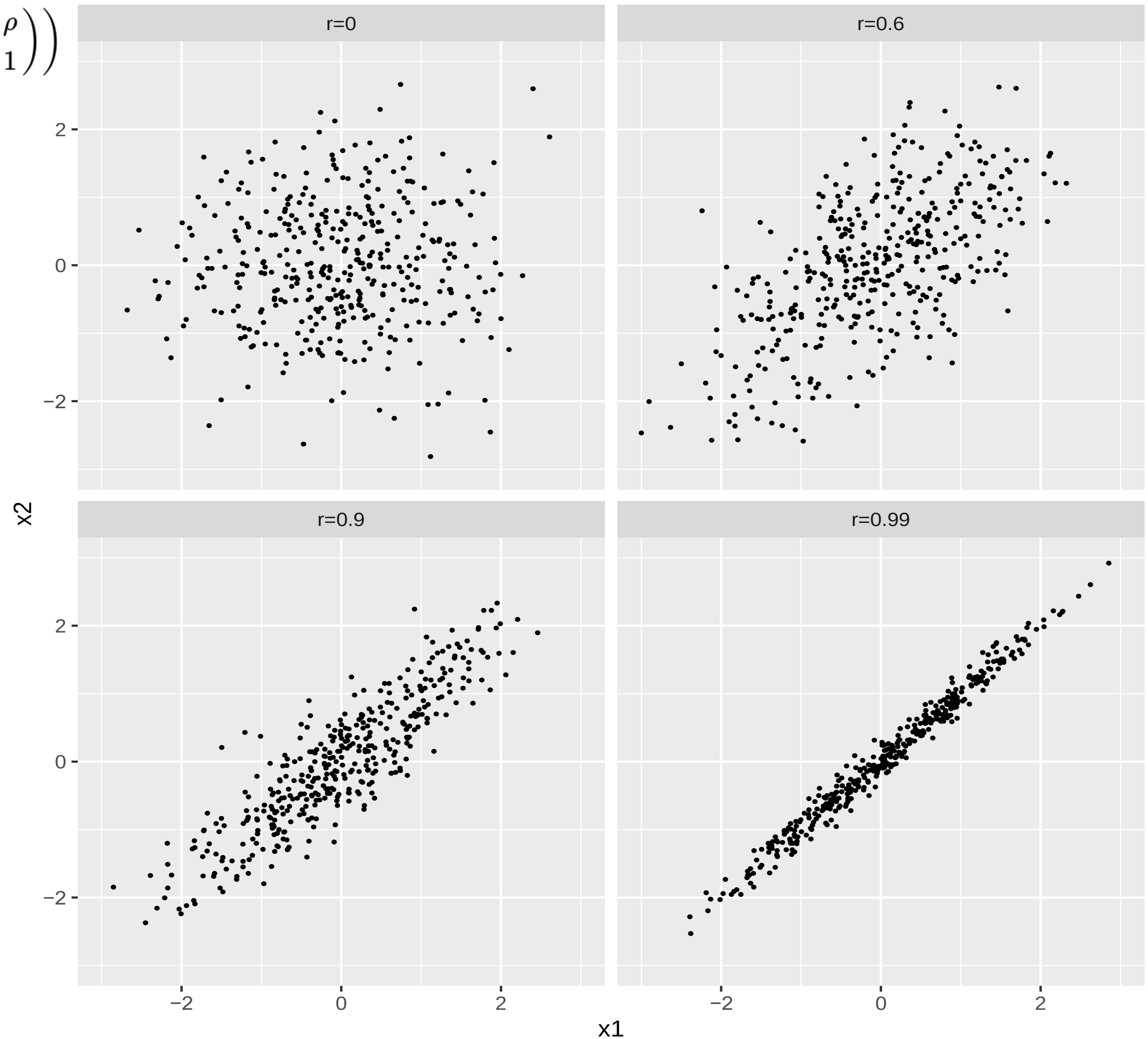
Dimension reduction
PCA, MDS, t-SNE

Principal Component Analysis (PCA)

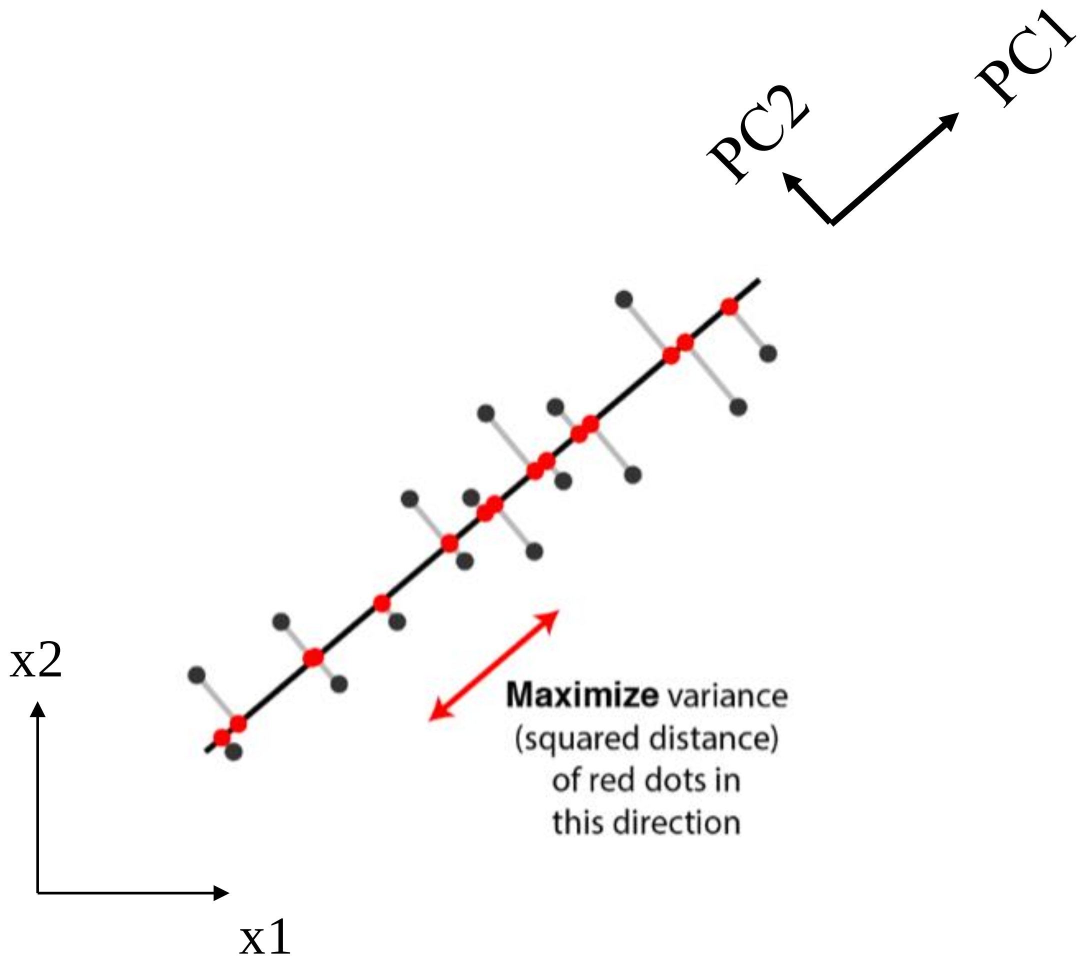
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

r = ρ

n = 400

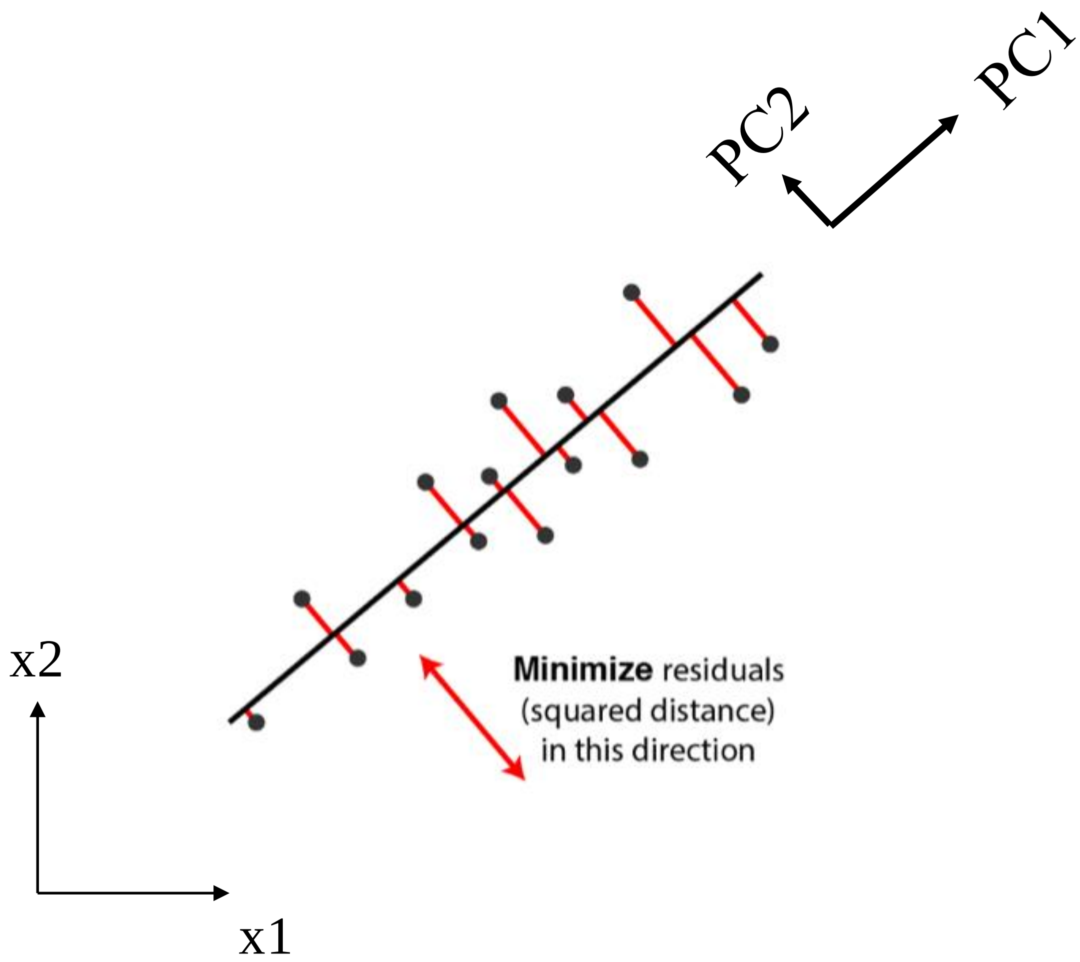


Principal Component Analysis (PCA)



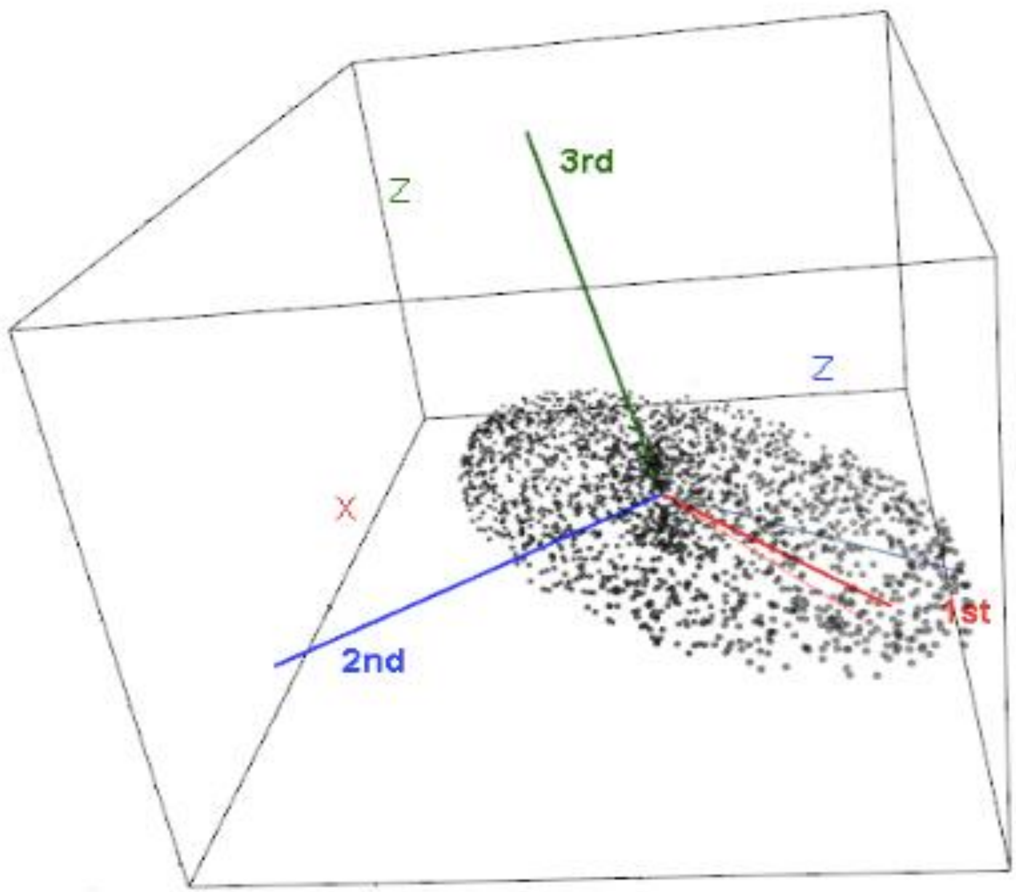
Karl Pearson 1901; Harold Hotelling 1933-1936

Principal Component Analysis (PCA)



Karl Pearson 1901

Principal Component Analysis (PCA)

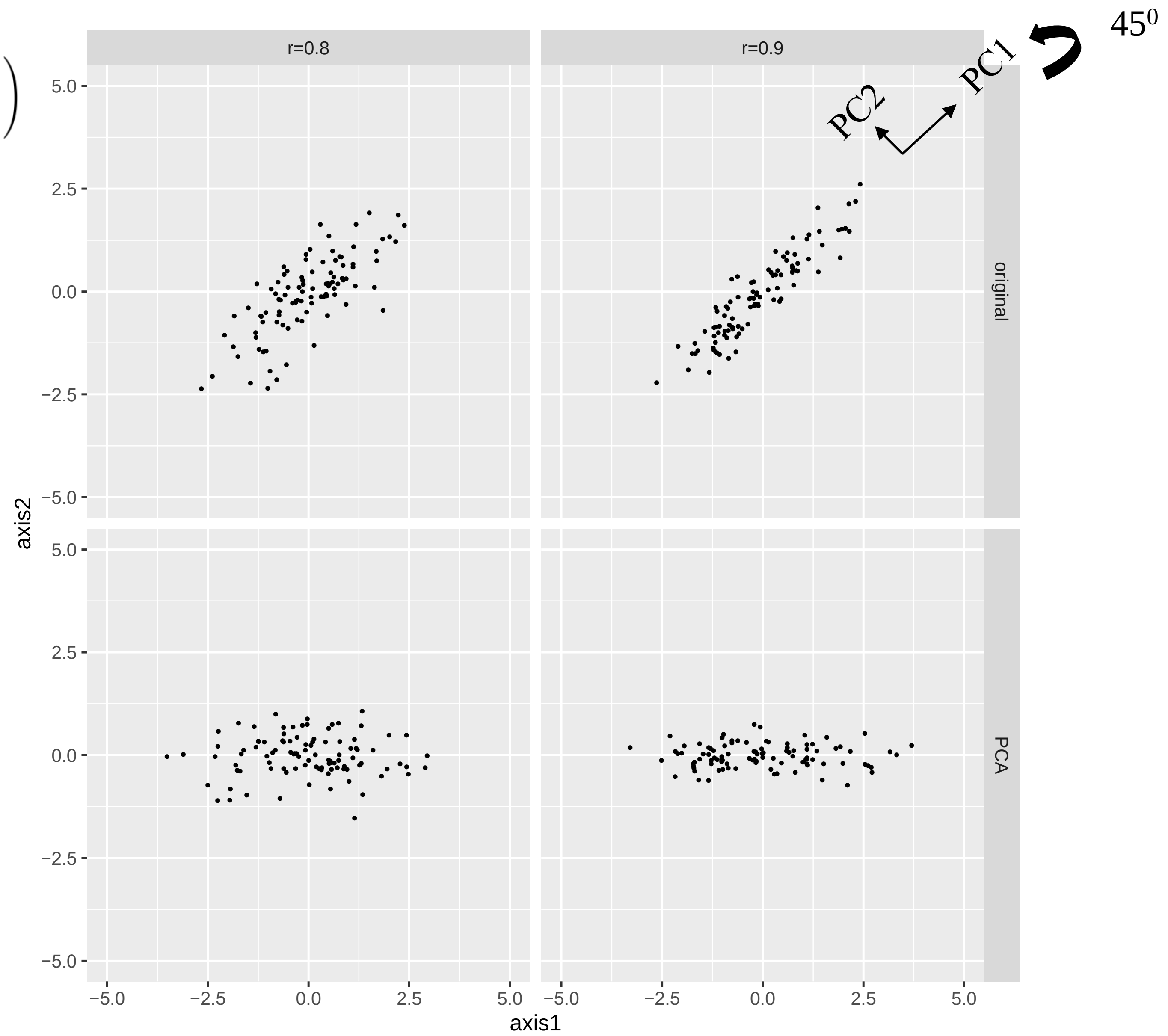


Geometric View Of PCA: Rotation Of Coordinates

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

$r = \rho$

$n = 100$



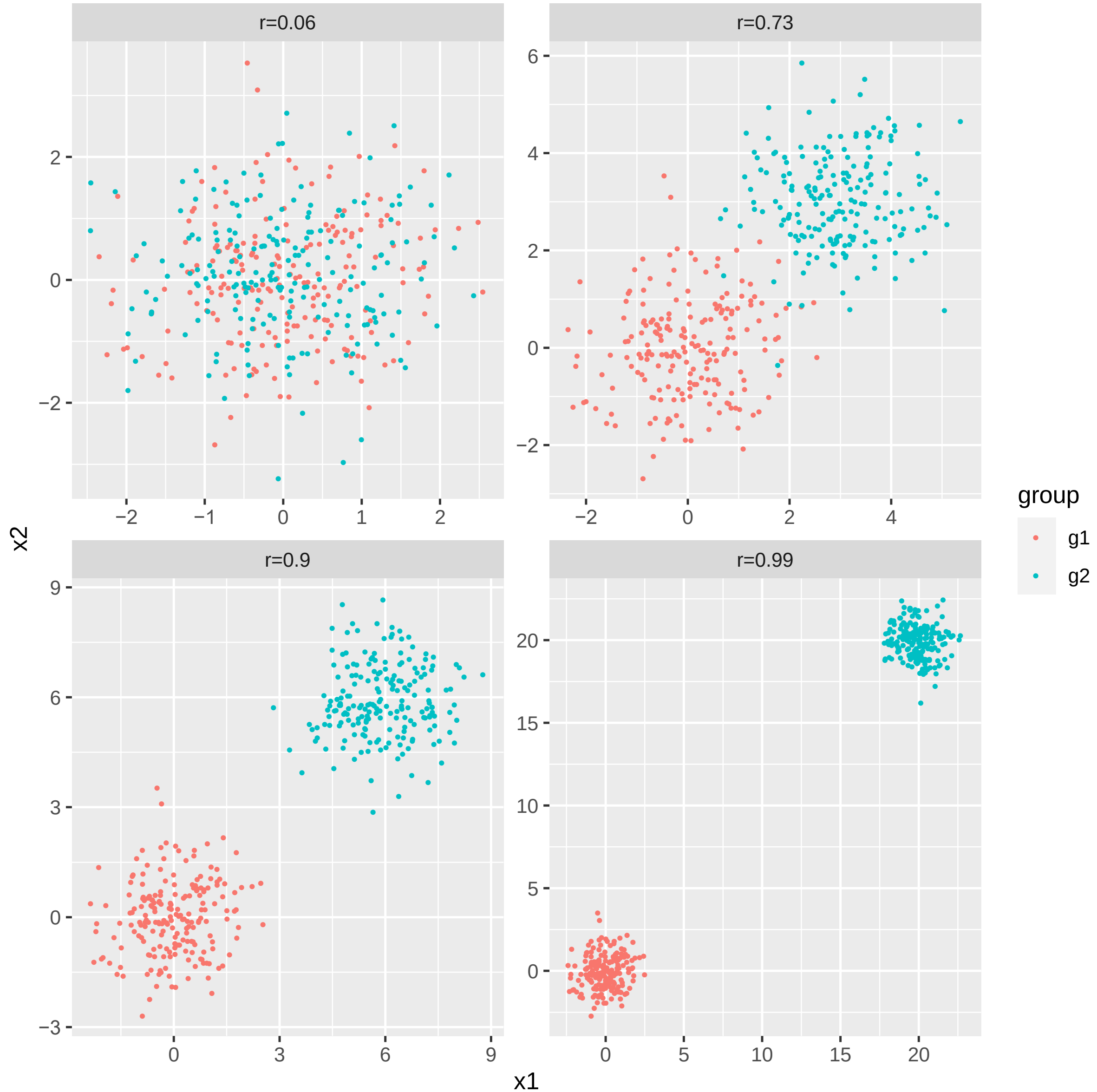
PCA: Samples With Two Groups

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

$\rho = 0$

group1
$\mu_1 = (0, 0, 0, 0)$
$\mu_2 = (0, 0, 0, 0)$

group2
$\mu_1 = (0, 3, 6, 20)$
$\mu_2 = (0, 3, 6, 20)$



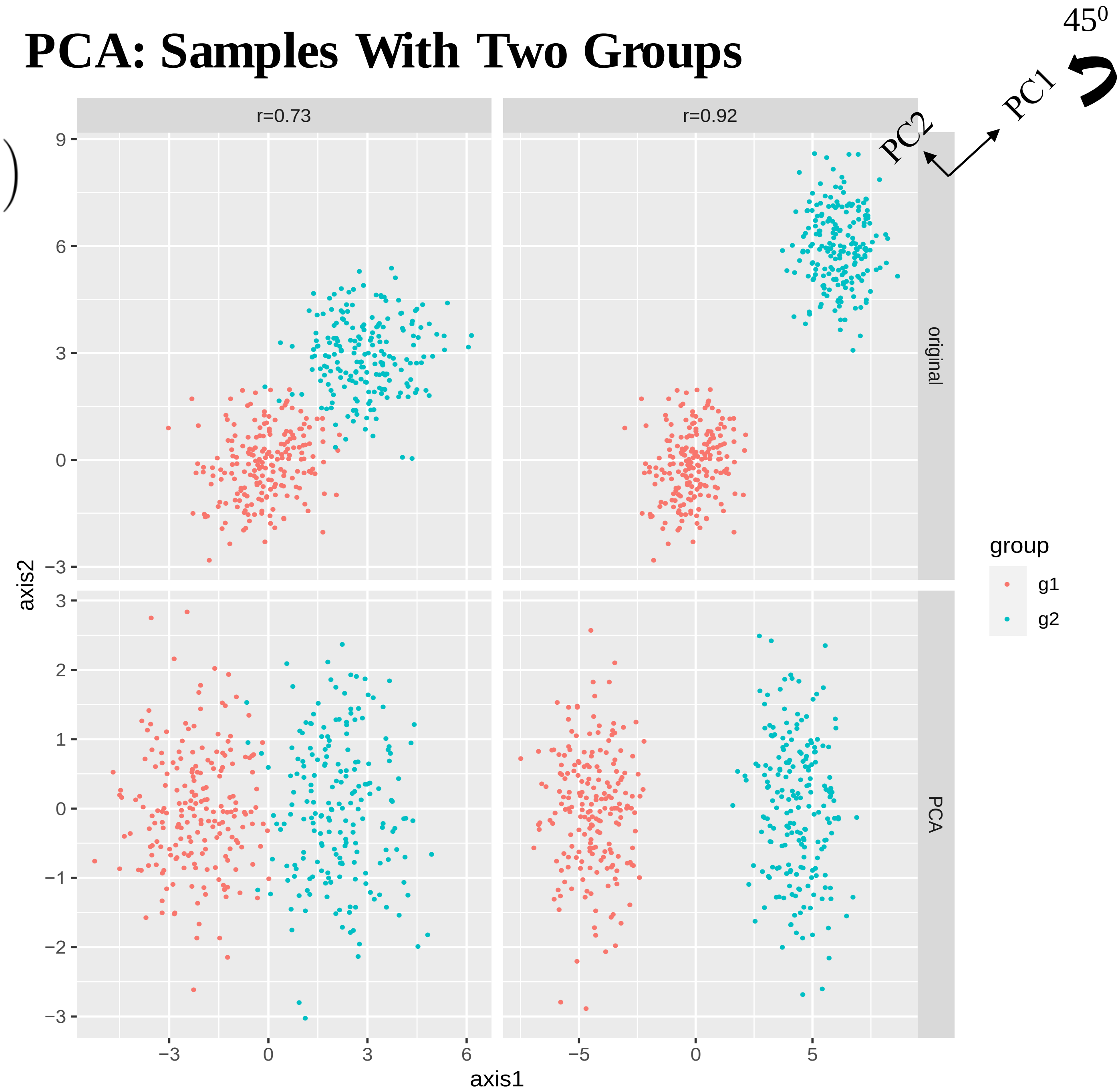
PCA: Samples With Two Groups

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

$\rho = 0$

group1
$\mu_1 = (0, 0)$
$\mu_2 = (0, 0)$

group2
$\mu_1 = (3, 6)$
$\mu_2 = (3, 6)$



PCA: Samples With Three Groups

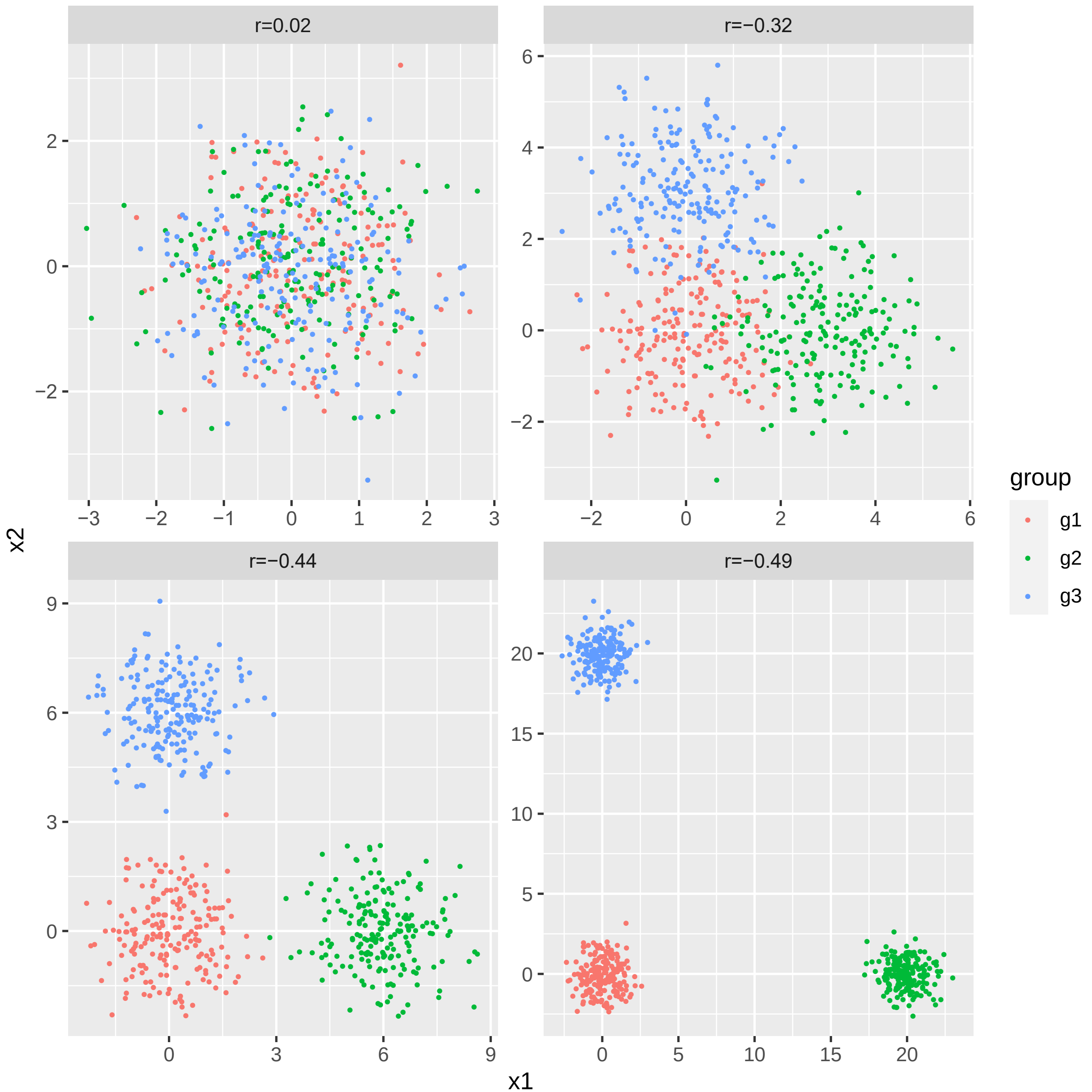
$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$\sigma_{ii} = 1$
$\sigma_{ij} = 0$

group1
$\mu_1 = (0, 0, 0, 0)$
$\mu_2 = (0, 0, 0, 0)$

group2
$\mu_1 = (0, 3, 6, 20)$
$\mu_2 = (0, 0, 0, 0)$

group3
$\mu_1 = (0, 0, 0, 0)$
$\mu_2 = (0, 3, 6, 20)$



PCA: Samples With Three Groups

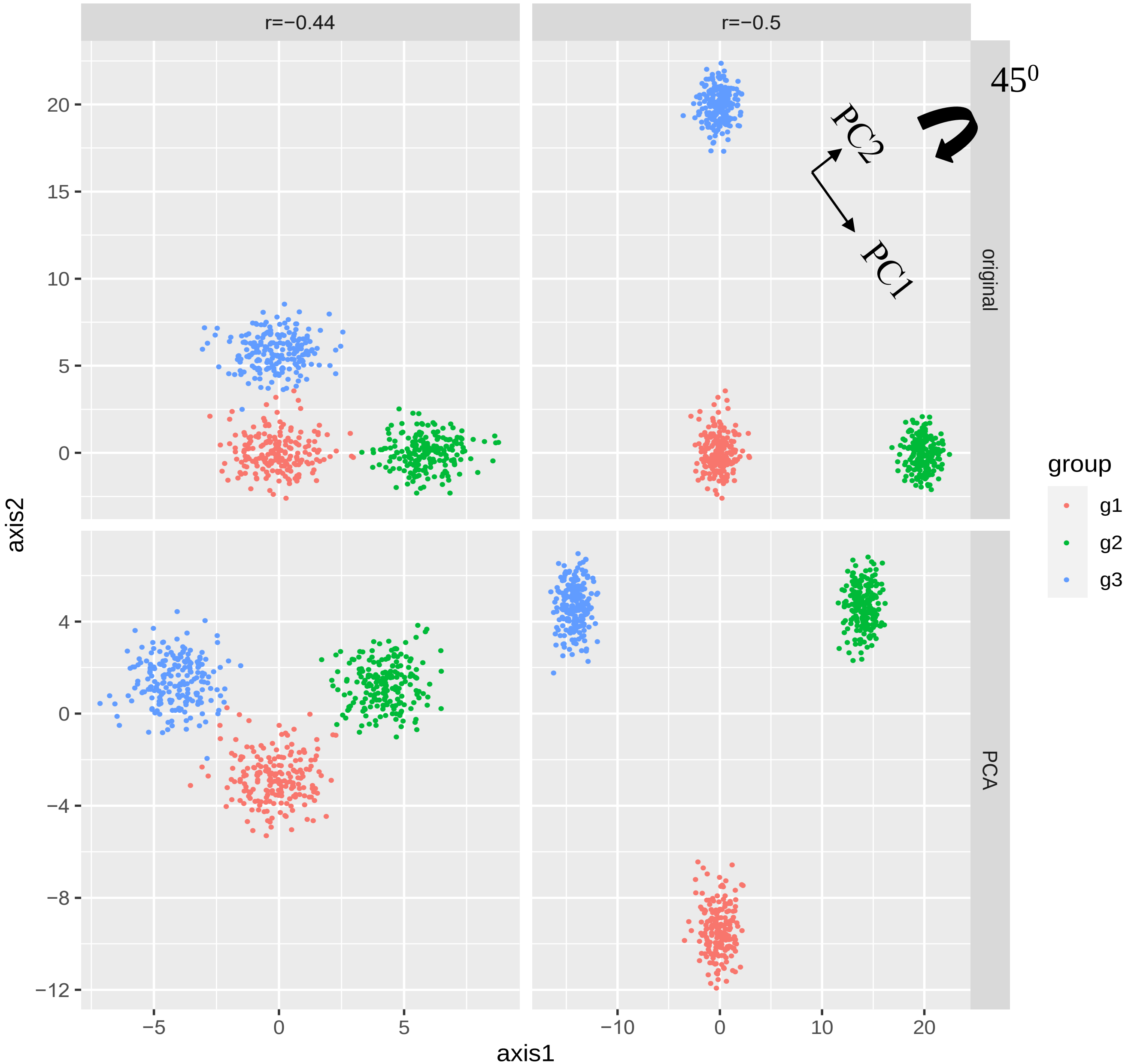
$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$\sigma_{ii} = 1$
$\sigma_{ij} = 0$

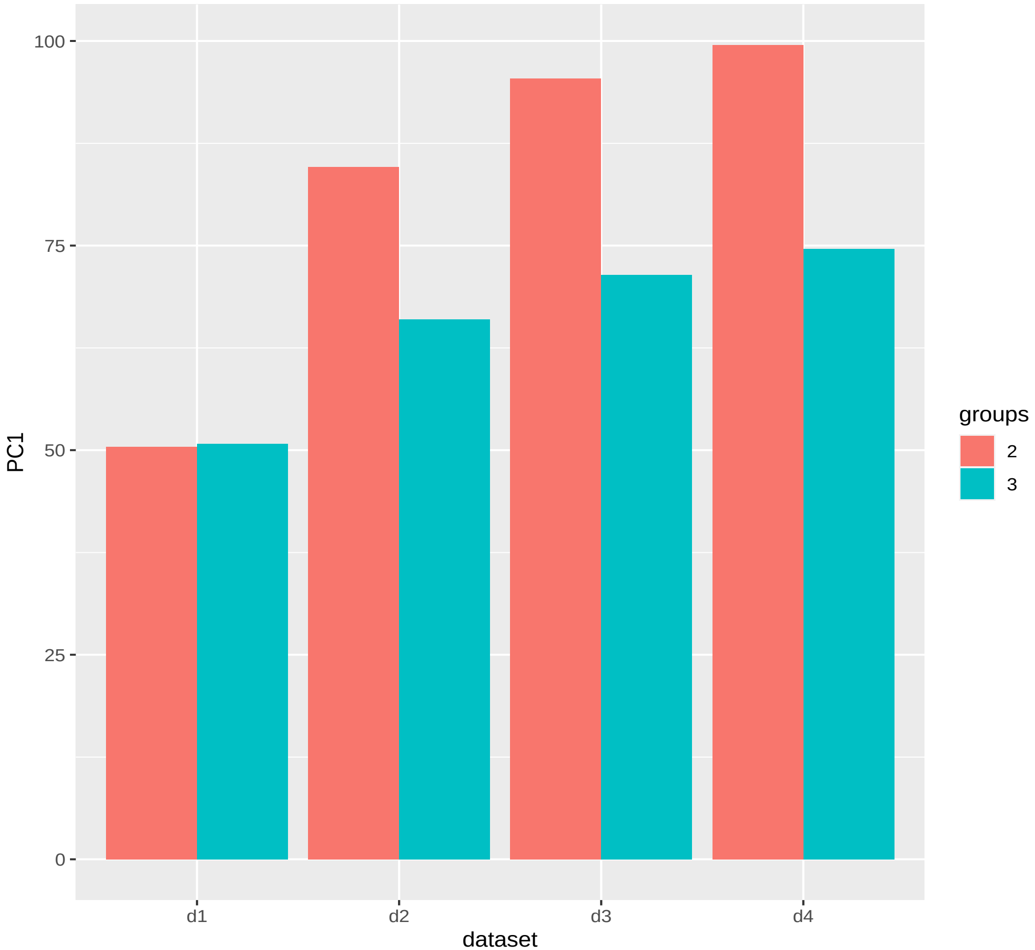
group1
$\mu_1 = (0, 0)$
$\mu_2 = (0, 0)$

group2
$\mu_1 = (6, 20)$
$\mu_2 = (0, 0)$

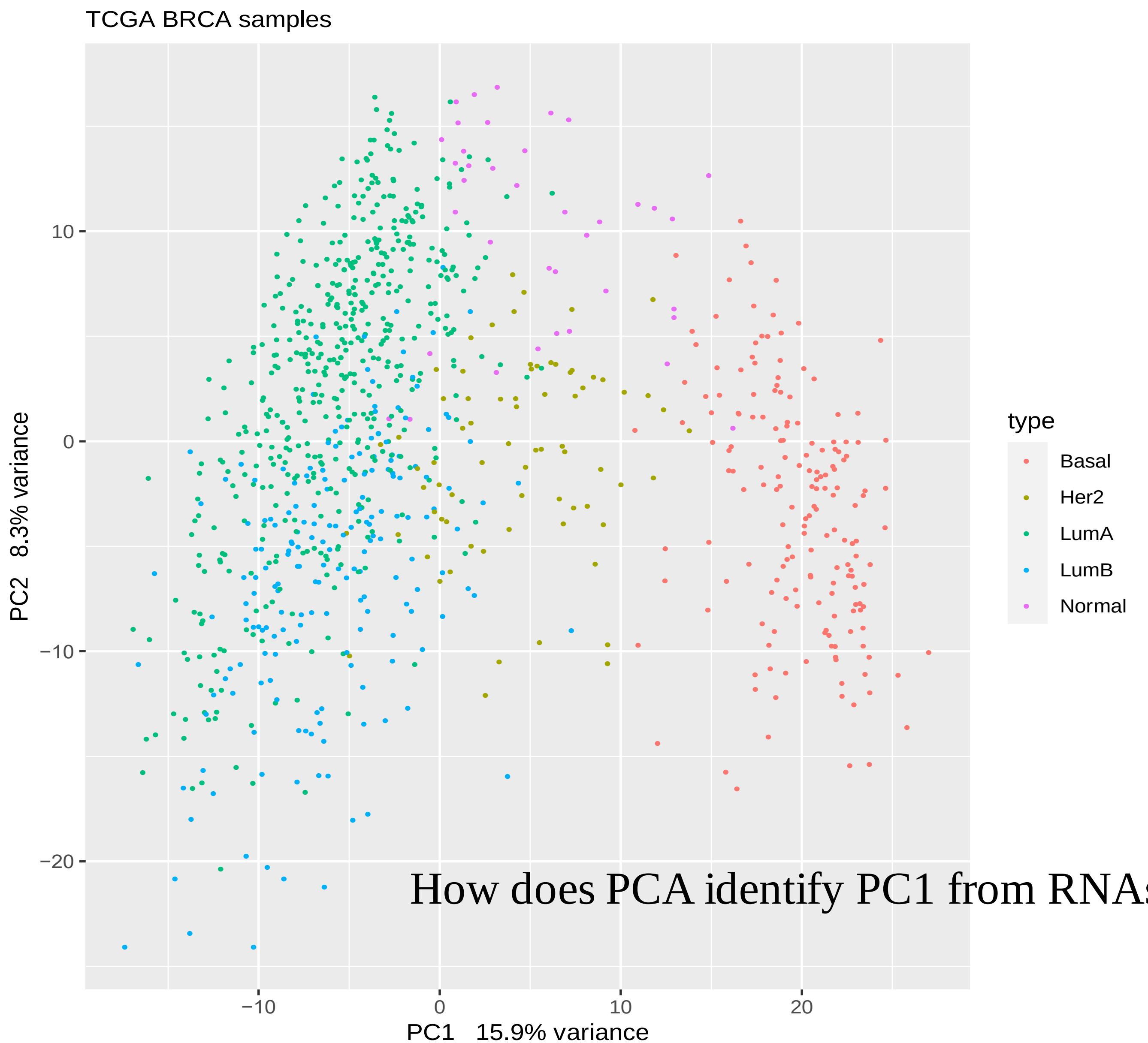
group3
$\mu_1 = (0, 0)$
$\mu_2 = (6, 20)$



Variance Accounted For By PC1



PCA Analysis Of TCGA Breast Cancer Data



How does PCA identify PC1 from RNAseq data?

Algorithm of PCA

$$z_1 = Xw_1$$

$$z_2 = Xw_2$$

$$z_3 = Xw_3$$

$$Z = XW$$

$$\text{var}(Z) = (XW)^T XW$$

$$\text{var}(Z) = W^T X^T XW = W^T SW$$

Choose w to maximize $w^T Sw$

subject to $W^T W = I$

Algorithm of PCA

Choose w to maximize $w^T S w$
subject to $W^T W = I$

$$L(w, \lambda) = w^T S w - \lambda(w^T w - 1)$$

$$\frac{\partial L}{\partial \mathbf{w}} = 2Sw - 2\lambda w$$

$$Sw = \lambda w$$

w is the eigenvector and λ is eigenvalue

Properties Of Eigen Values And Eigen Vectors

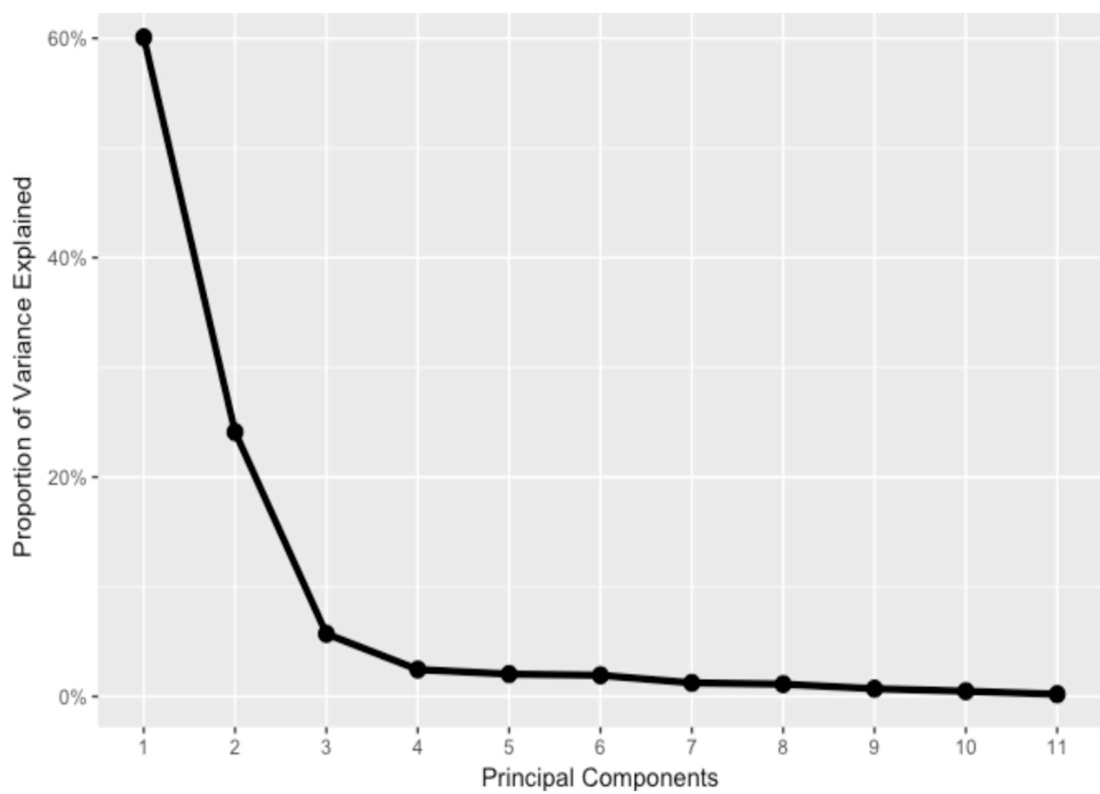
Covariance matrix S

- There are p pairs of Eigen values and Eigen vectors
- Eigen values are ranked from the largest to smallest
- For covariance matrix, all eigen values are nonnegative

Variance of PCs Are Eigen Value And Are Additive

$$\begin{aligned} \mathrm{var}(z) &= w^T S w \\ &= w^T \lambda w \\ &= \lambda \end{aligned}$$

$$\mathrm{var}(Z) = \lambda_1 + \lambda_2 \dots + \lambda_p$$



Singular Value Decomposition (SVD)

$$Z = XW$$

$$Z_s = XWD^{-1/2}$$

$$Z_s D^{1/2} W^T = X$$

$$X = Z_s D^{1/2} W^T$$

$$X = U \Sigma V^T$$

Right and Left Eigen Vectors Of SVD

$$X = U\Sigma V^T$$

$$\begin{aligned} X^TX &= V\Sigma U^T U\Sigma V^T \\ &= V\Sigma^2 V^T \end{aligned}$$

$$\begin{aligned} XX^T &= U\Sigma V^T V\Sigma U^T \\ &= U\Sigma^2 U^T \end{aligned}$$

Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS)

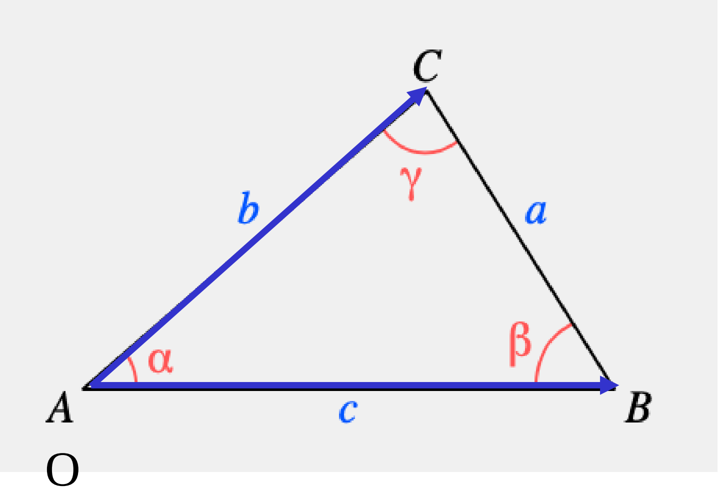
Pairwise Distance Matrix

	Athens	Berlin	Dublin	London	Madrid	Paris	Rome	Warsaw
Athens	0	1119	1777	1486	1475	1303	646	1013
Berlin	1119	0	817	577	1159	545	736	327
Dublin	1777	817	0	291	906	489	1182	1135
London	1486	577	291	0	783	213	897	904
Madrid	1475	1159	906	783	0	652	856	1483
Paris	1303	545	489	213	652	0	694	859
Rome	646	736	1182	897	856	694	0	839
Warsaw	1013	327	1135	904	1483	859	839	0

Multidimensional Scaling (MDS)

Law of cosine

$a^2 = b^2 + c^2 – 2bc \cos(\alpha)$



O

$2bc \cos(\alpha) = b^2 + c^2 – a^2$

$bc \cos(\alpha) = -½(a^2 - b^2 + c^2)$

$\mathbf{b} \bullet \mathbf{c} = bc \cos(\alpha)$

$\mathbf{b} \bullet \mathbf{c} = -1/2(a^2 – b^2 – c^2)$

Multidimensional Scaling (MDS)

$$\mathbf{b} \bullet \mathbf{c} = -1/2(a^2 - b^2 - c^2)$$

$$\begin{matrix} K_{11} & K_{12} & \dots & K_{1n} \\ K_{21} & K_{22} & \dots & K_{2n} \\ . & & & . \\ . & & & . \\ K_{n1} & K_{n2} & \dots & K_{nn} \end{matrix}$$

$$K = U\Lambda U^T$$

$$Z = U\Lambda^{1/2}$$

MDS And PCA Are Equivalent

$$Z = U\Lambda^{1/2}$$

$$X = U\Sigma V^T$$

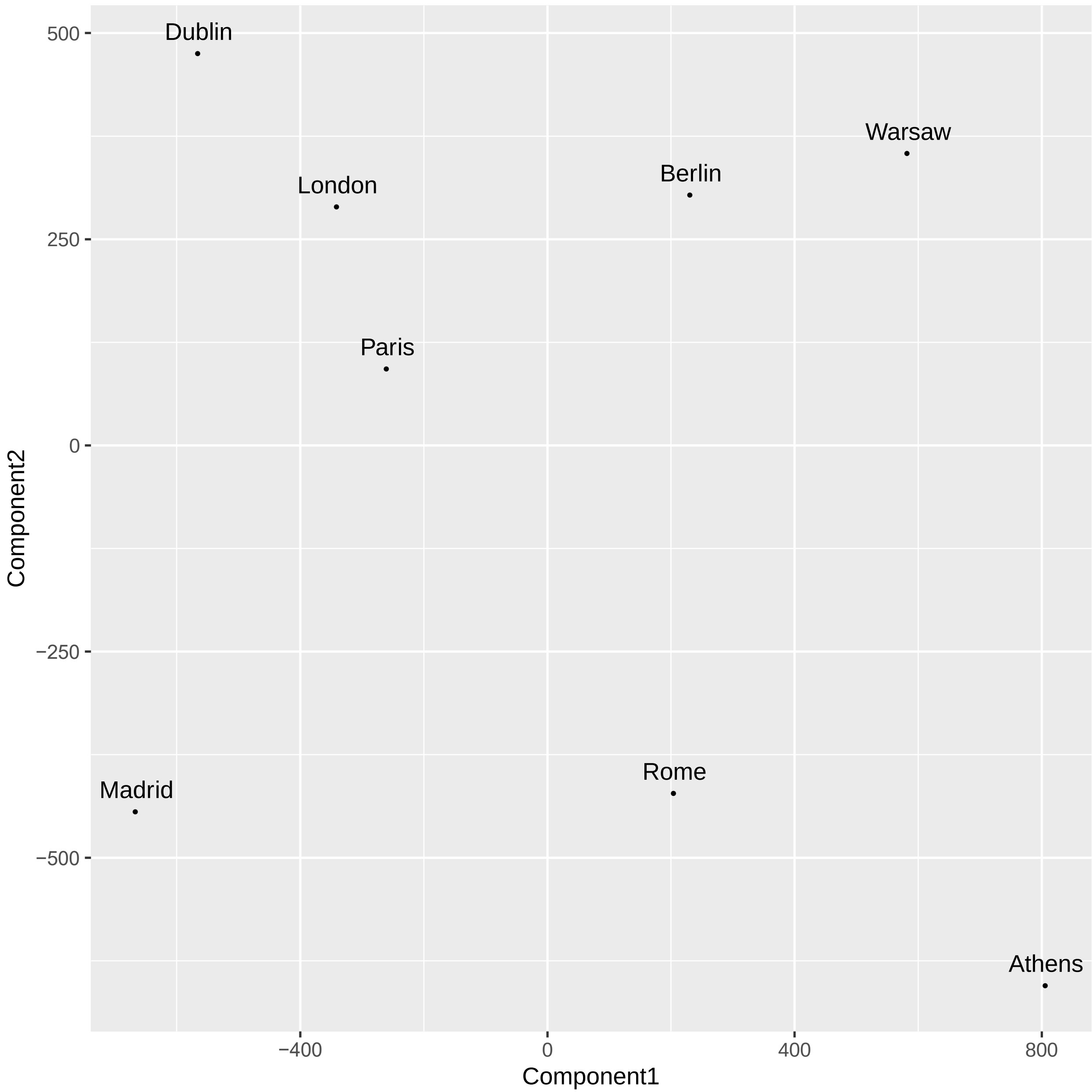
$$X^T X = V\Sigma^2 V^T$$

$$XX^T = U\Sigma^2 U^T$$

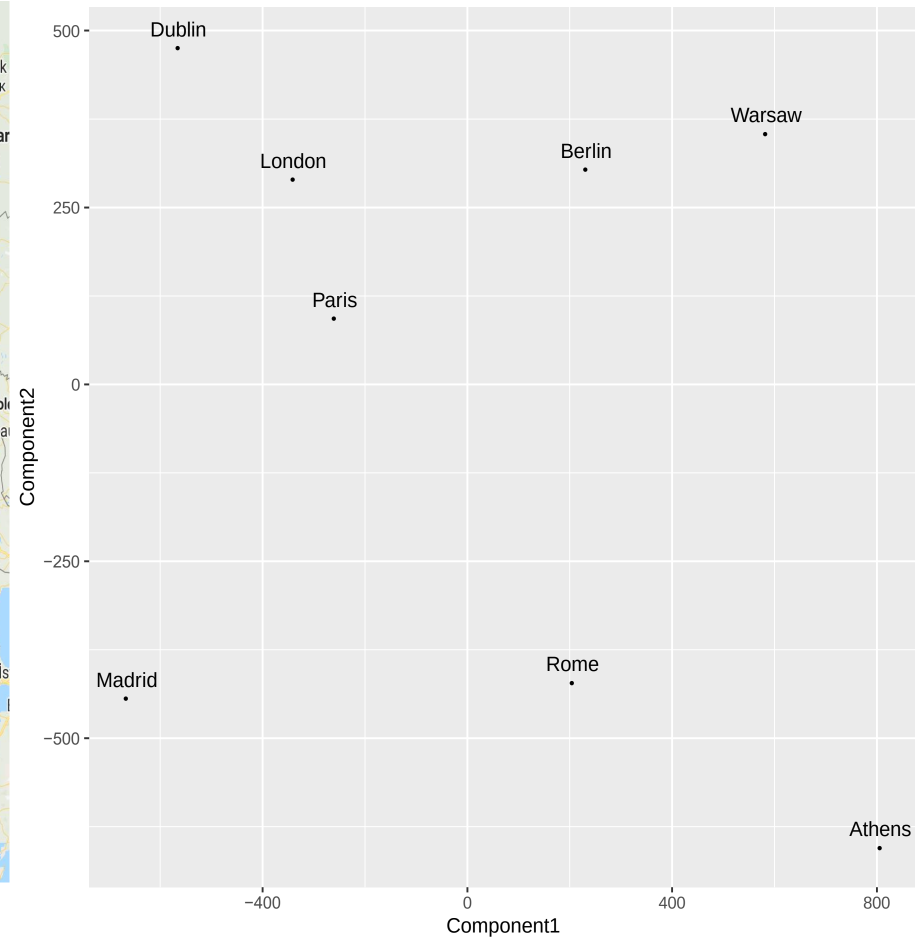
$$XV = U\Sigma$$

$$Z = U\Lambda^{1/2}$$

Multidimensional Scaling (MDS)



Multidimensional Scaling (MDS)



Outline Of The Talk

1) **Linear dimension reduction methods**
   PCA, MDS, and SVD

2) **Nonlinear dimension reduction methods**
   Isomap, LLE, Laplacian Eigenmap, TSNE, and UMAP

3) **Canonical correlation and Trajectory analysis**
   Data integration and reversed graph embedding