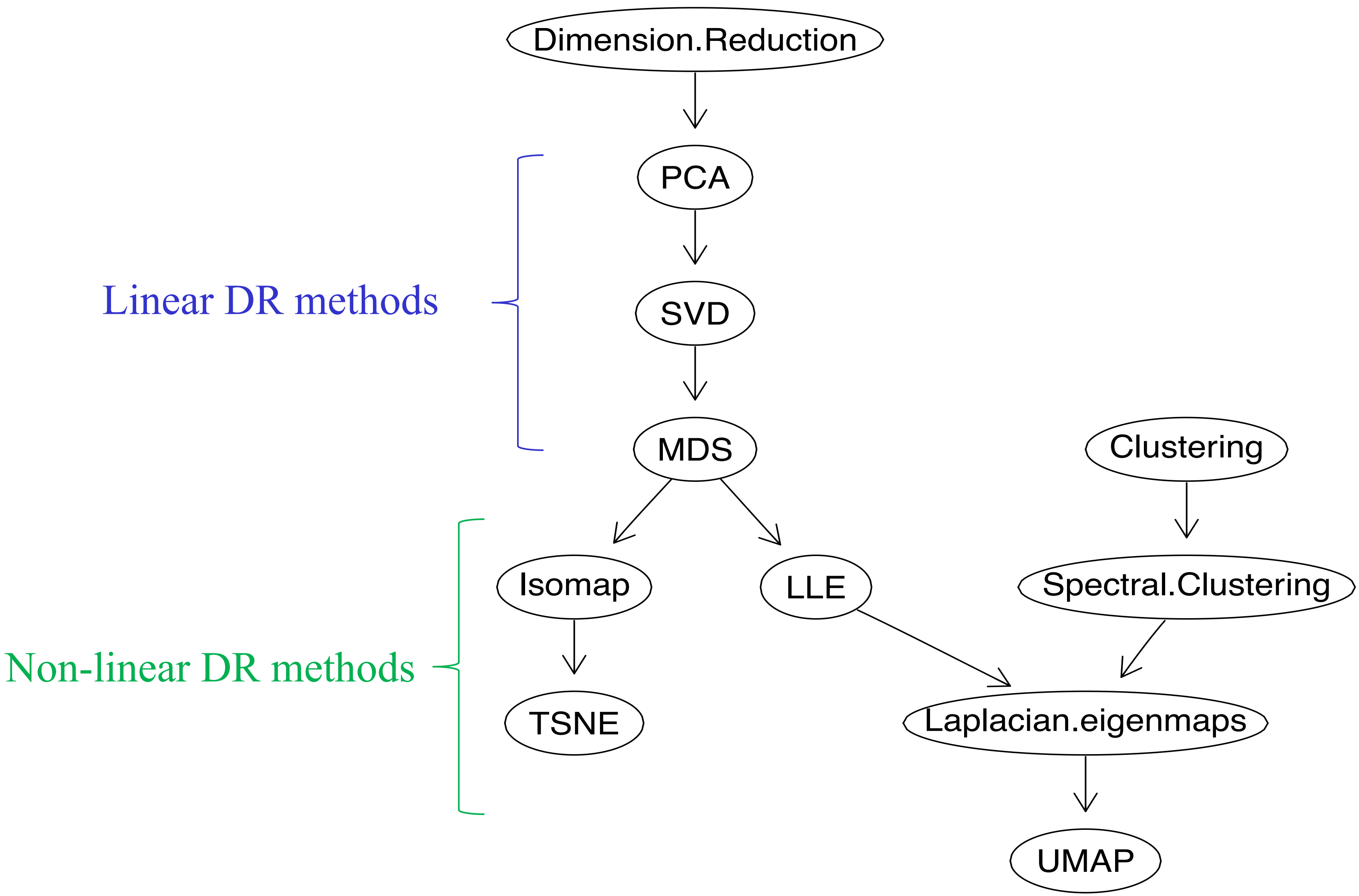
Dimension Reduction Methods:
From PCA to TSNE and UMAP

Maxwell Lee

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

May 22, 2020

Outline for Dimension Reduction Methods



Data Matrix (Table)

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

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

$$\begin{bmatrix} y_1 & x_{11} & x_{12} & \ldots & x_{1p} \\ y_2 & x_{21} & x_{22} & \ldots & x_{2p} \\ . & & & & . \\ . & & & & . \\ y_n & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

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

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

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

Unsupervised Statistical Learning

unsupervised
- Clustering analysis
  hierarchical clustering, k-means
  GMM, spectral clustering
- Dimension reduction
  PCA, MDS, TSNE, UMAP

GMM: Gaussian Mixture Model
PCA: Principal Component Analysis
MDS: Multidimensional scaling
TSNE: T-distributed Stochastic Neighbor Embedding
UMAP: Uniform Manifold Approximation and Projection

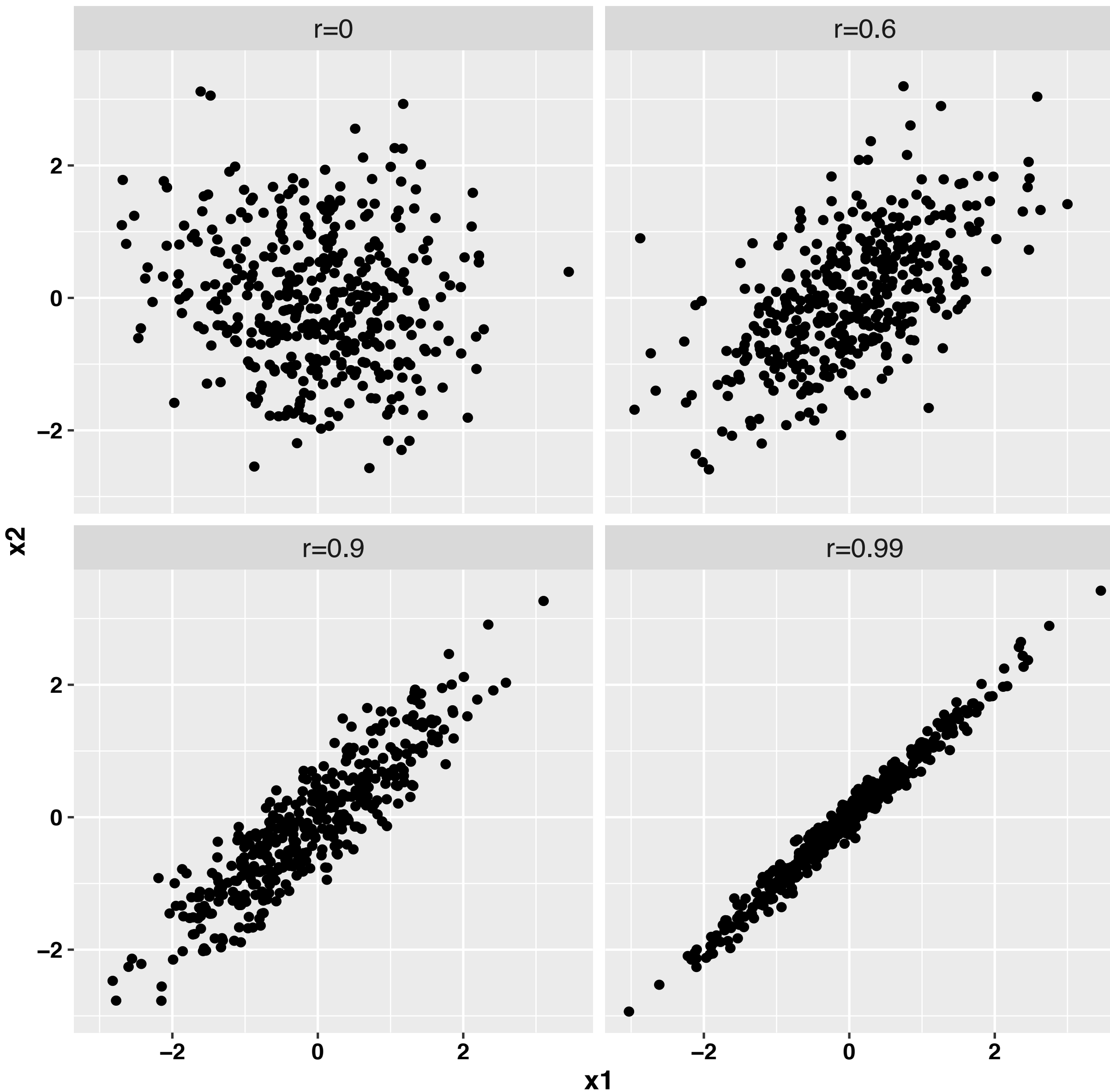
The Presence of Correlation Between Variables Is the Reason Why We Can Reduce Dimension by PCA

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

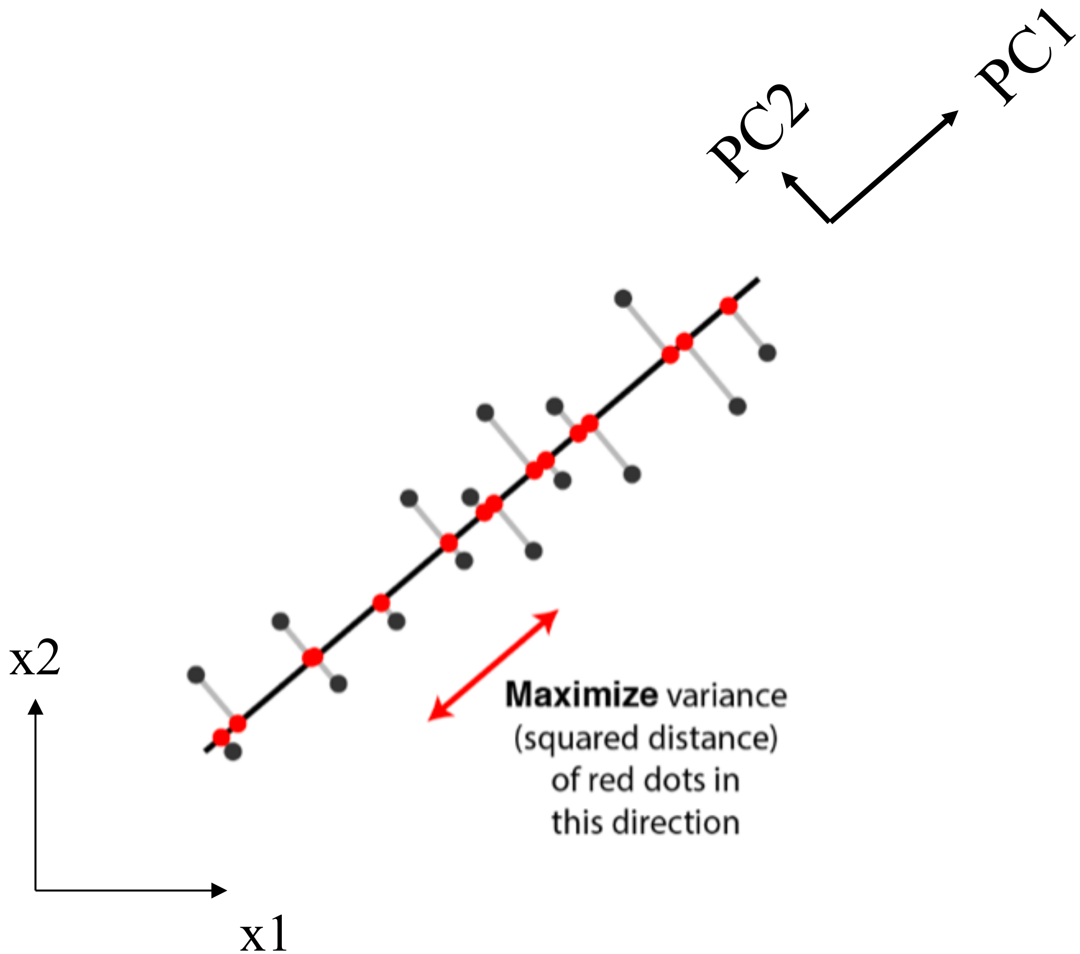
r = ρ

	μ_1	μ_2	σ_1	σ_2	ρ
d1	0	0	1	1	0
d2	0	0	1	1	0.6
d3	0	0	1	1	0.9
d4	0	0	1	1	0.99

n = 400

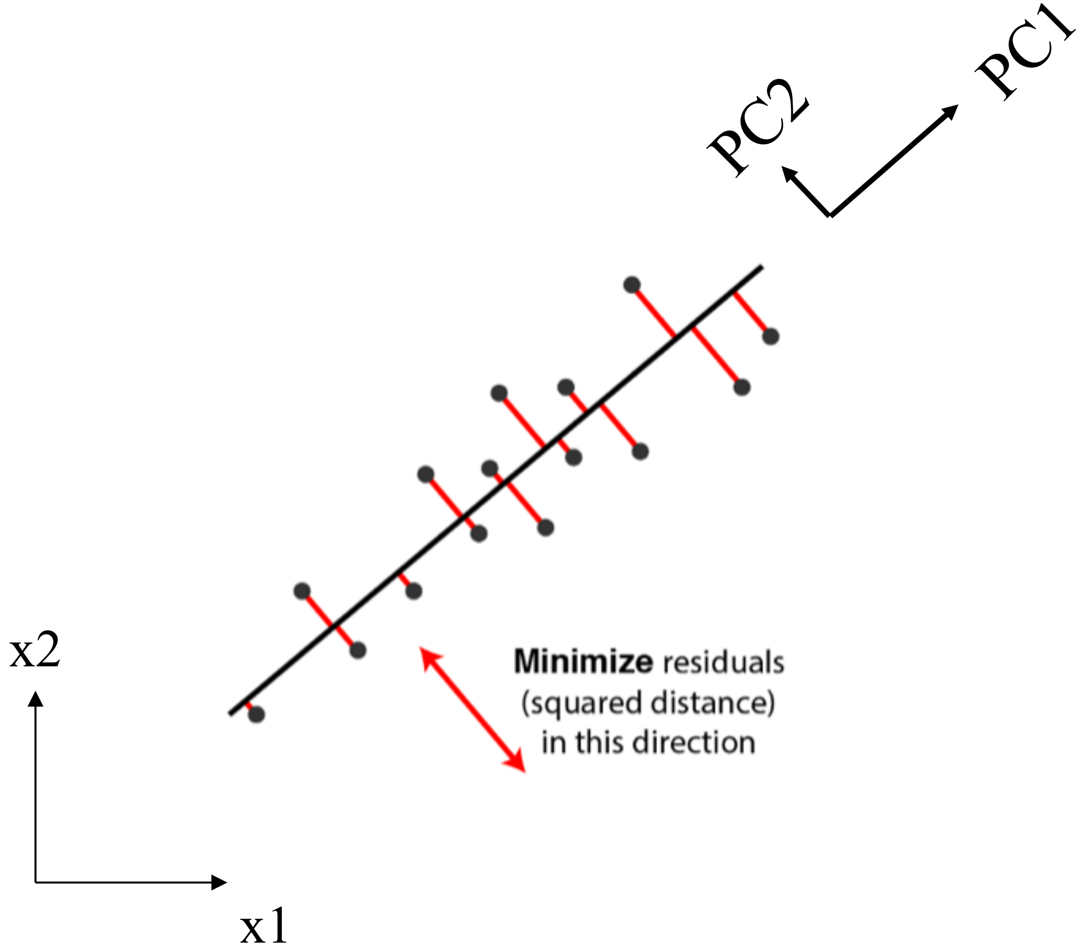


Principal Component Analysis (PCA)



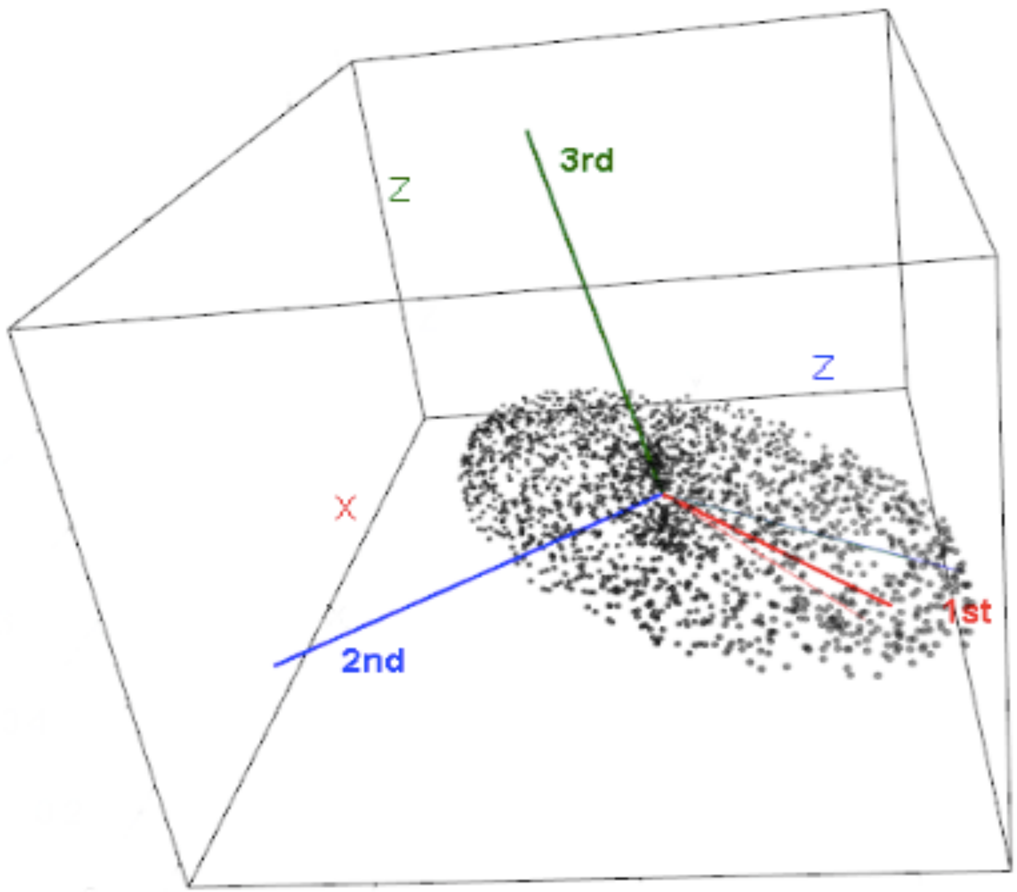
Karl Pearson 1901; Harold Hotelling 1933-1936

Principal Component Analysis (PCA)



Karl Pearson 1901

Principal Component Analysis (PCA)

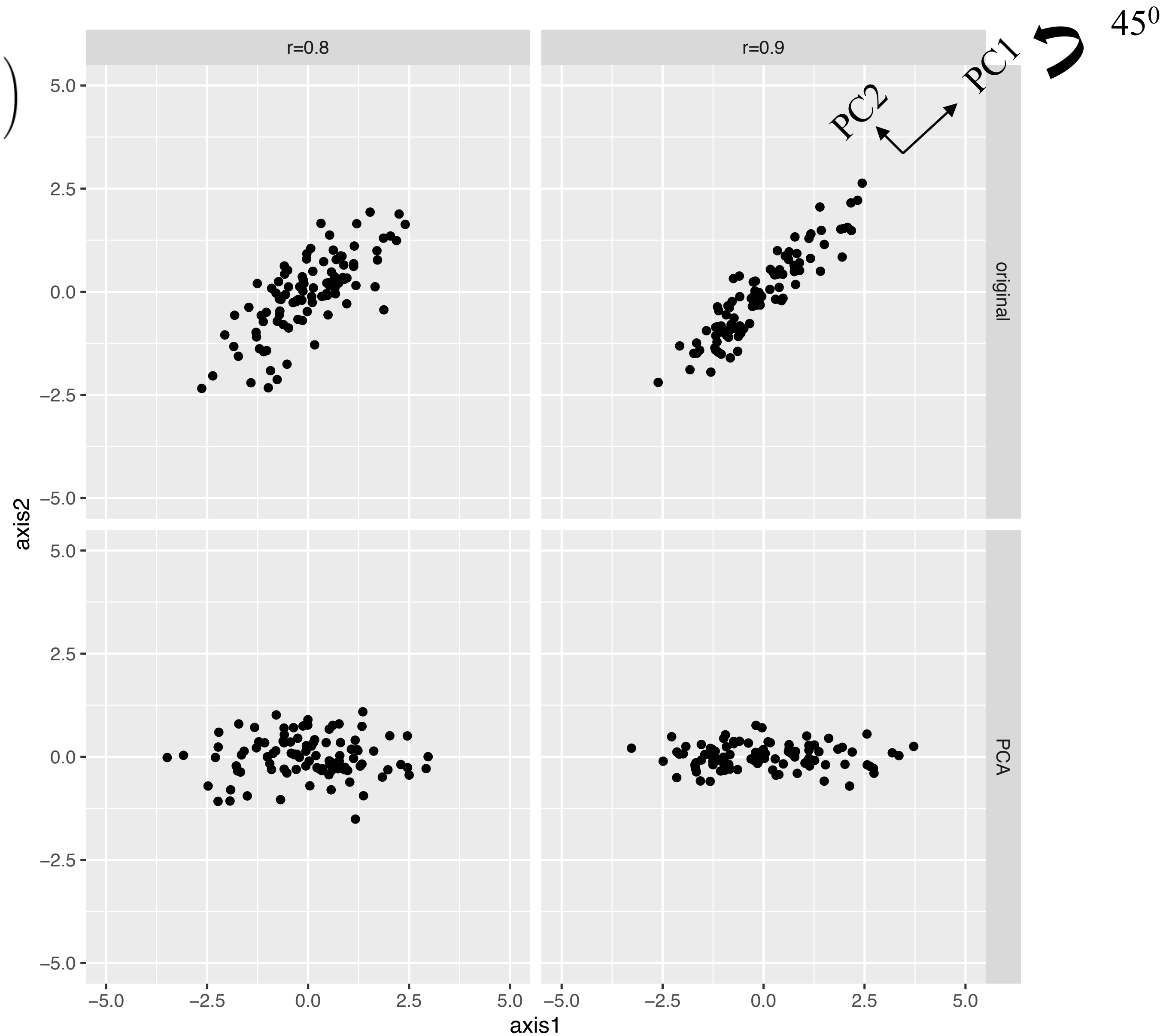


Geometric View of PCA: Rotation of Coordinates

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

r = ρ

n = 100

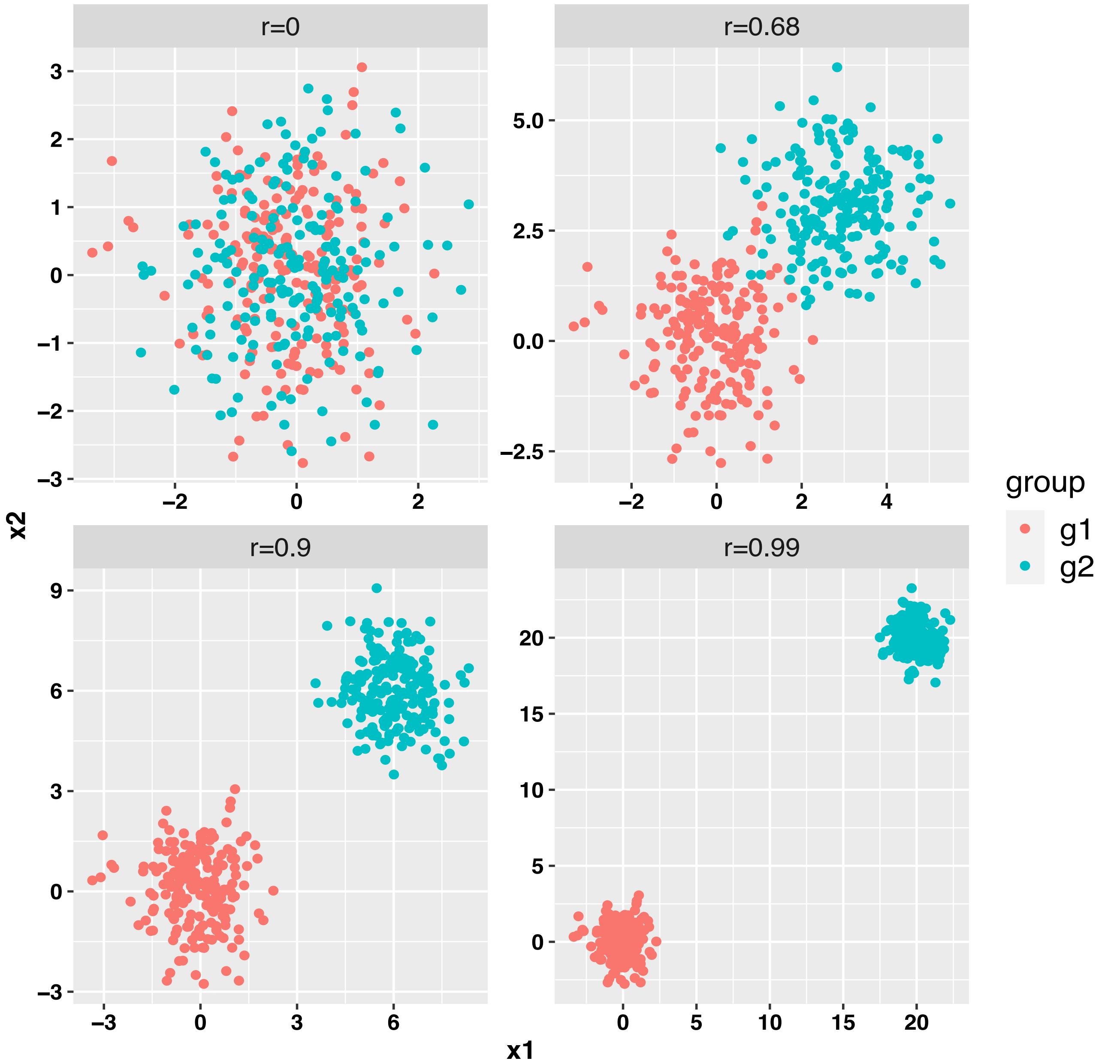


Correlation Between Variables Can Result from Heterogeneity in Sample

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

$$\rho = 0$$

	Group1		Group2	
	μ_1	μ_2	μ_1	μ_2
d1	0	0	0	0
d2	0	0	3	3
d3	0	0	6	6
d4	0	0	20	20

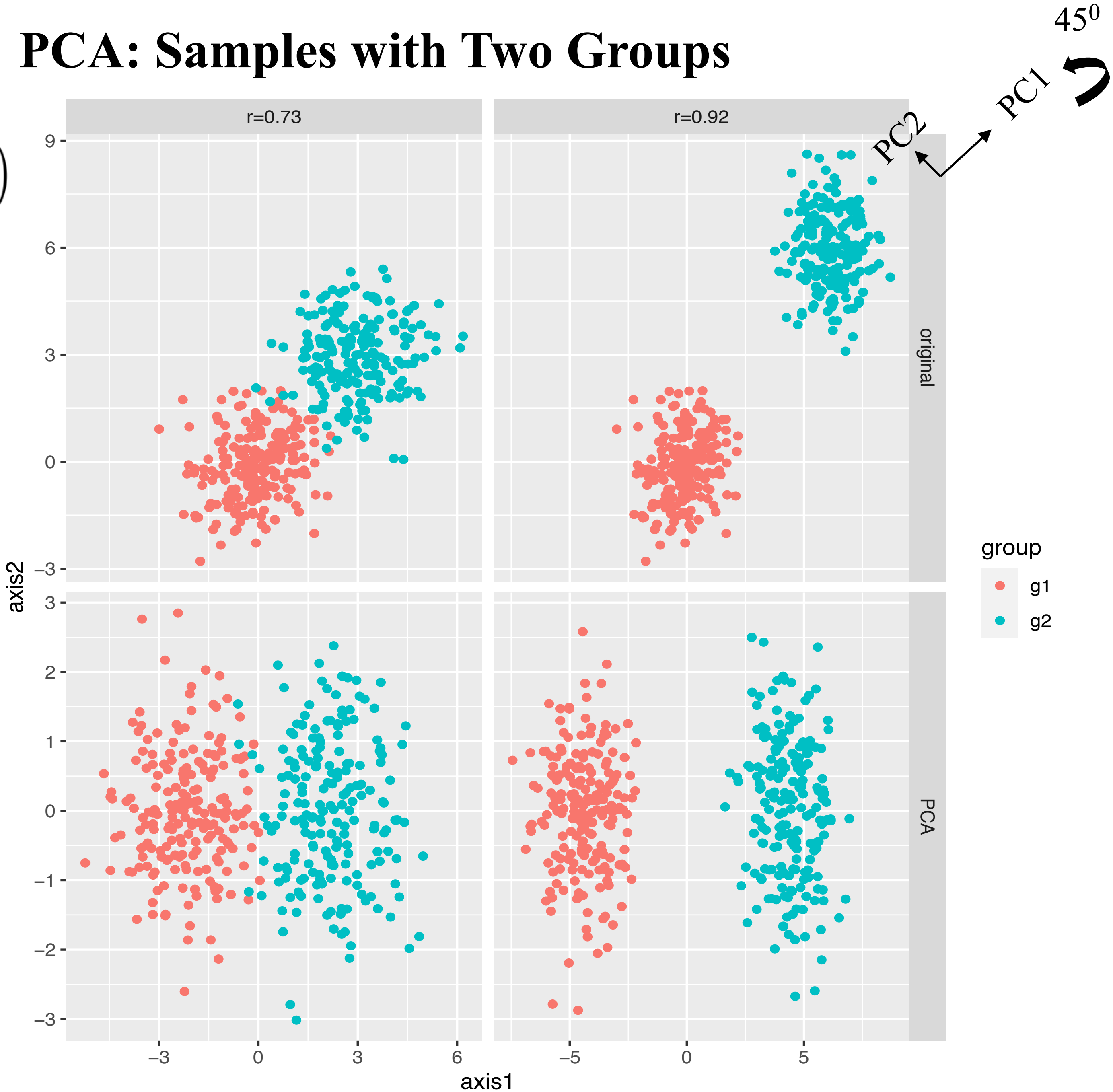


PCA: Samples with Two Groups

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

$\rho = 0$

	Group1		Group2	
	μ_1	μ_2	μ_1	μ_2
d1	0	0	3	3
d2	0	0	6	6

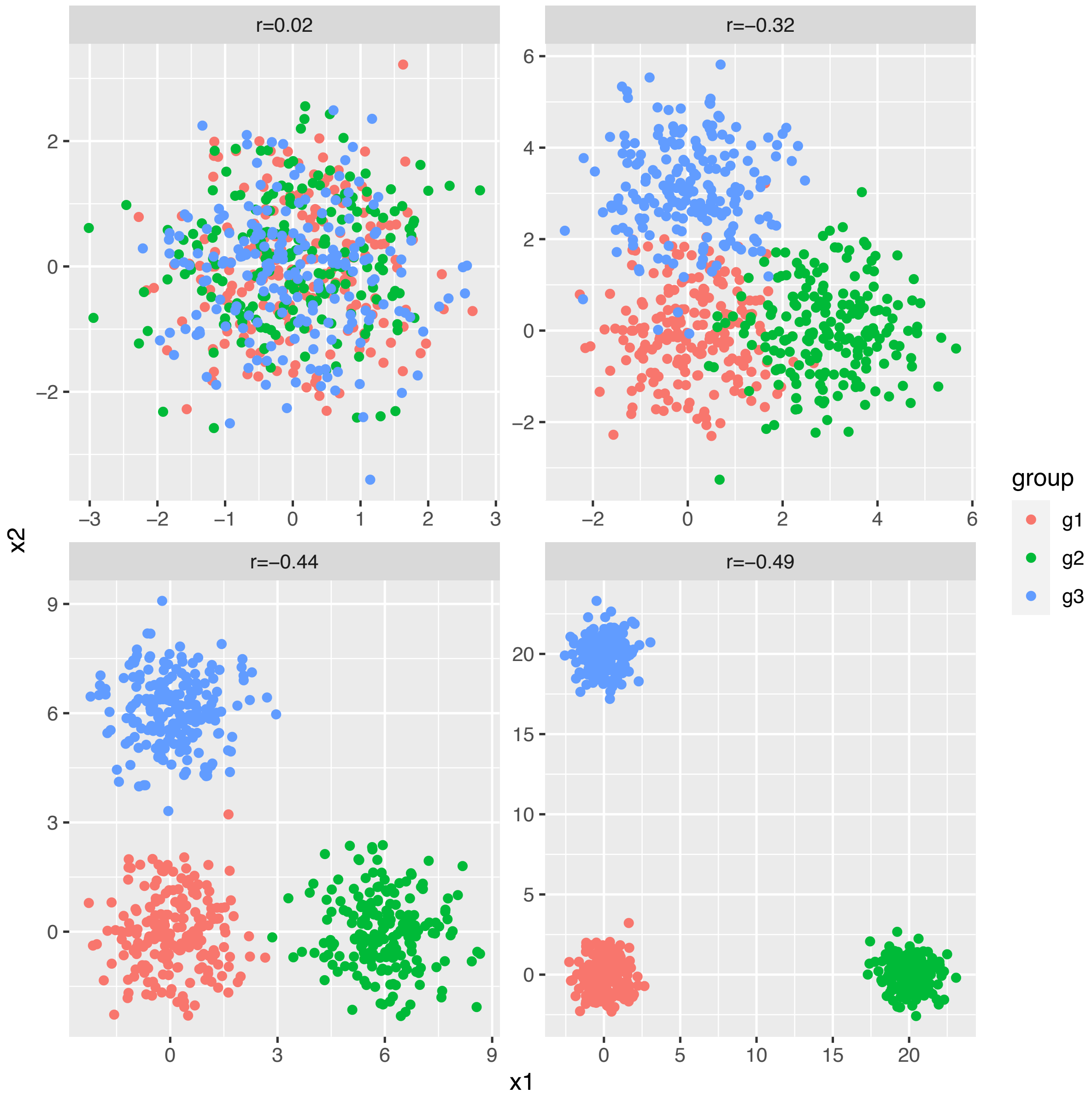


PCA: Samples with Three Groups

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

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

	Group1		Group2		Group3	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
d1	0	0	0	0	0	0
d2	0	0	3	0	0	3
d3	0	0	6	0	0	6
d4	0	0	20	0	0	20

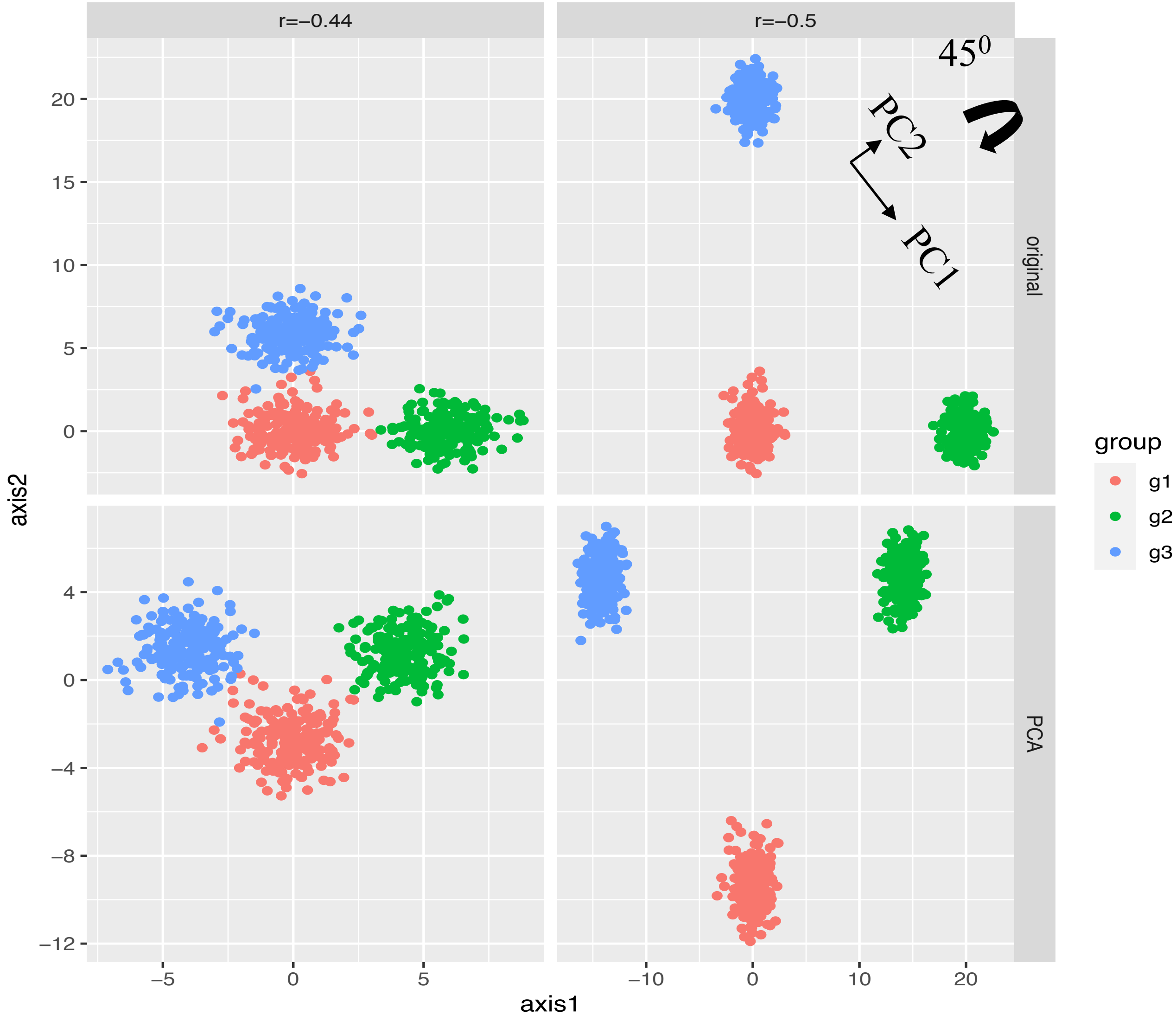


PCA: Samples with Three Groups

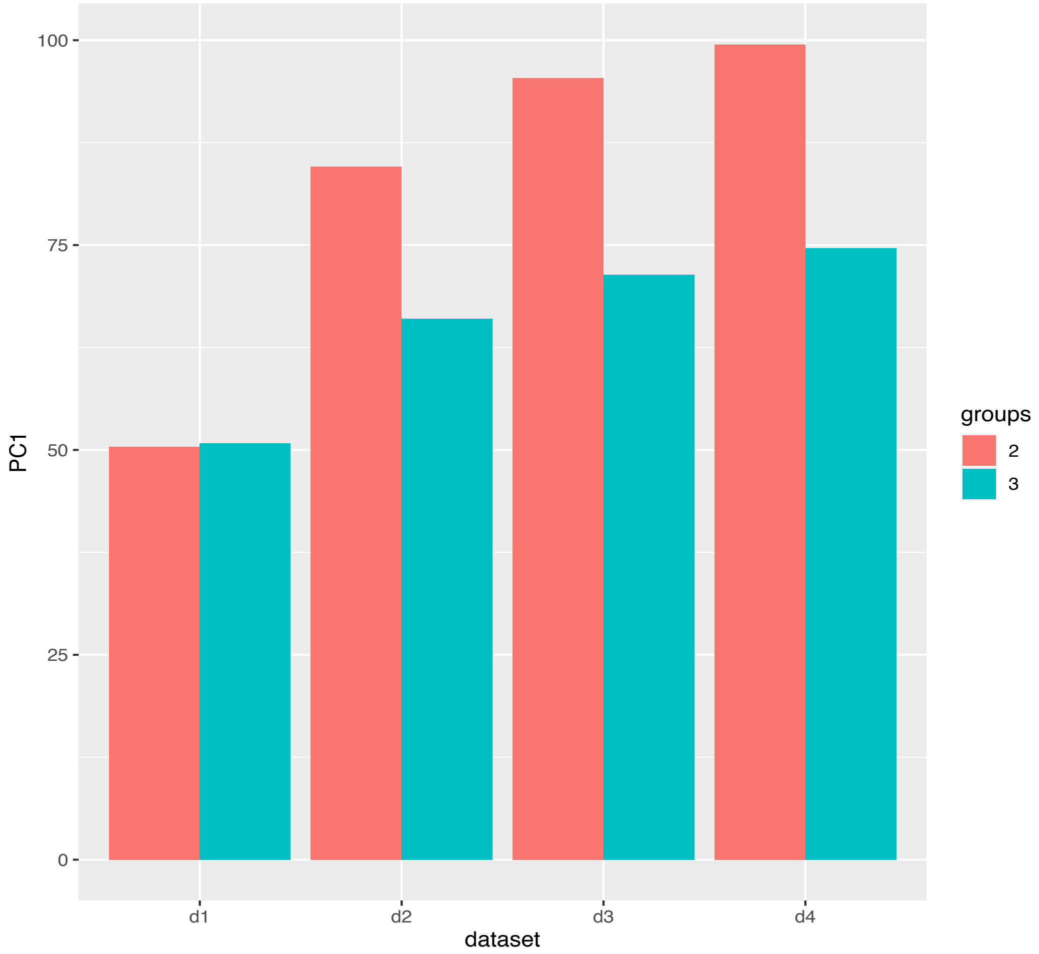
$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

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

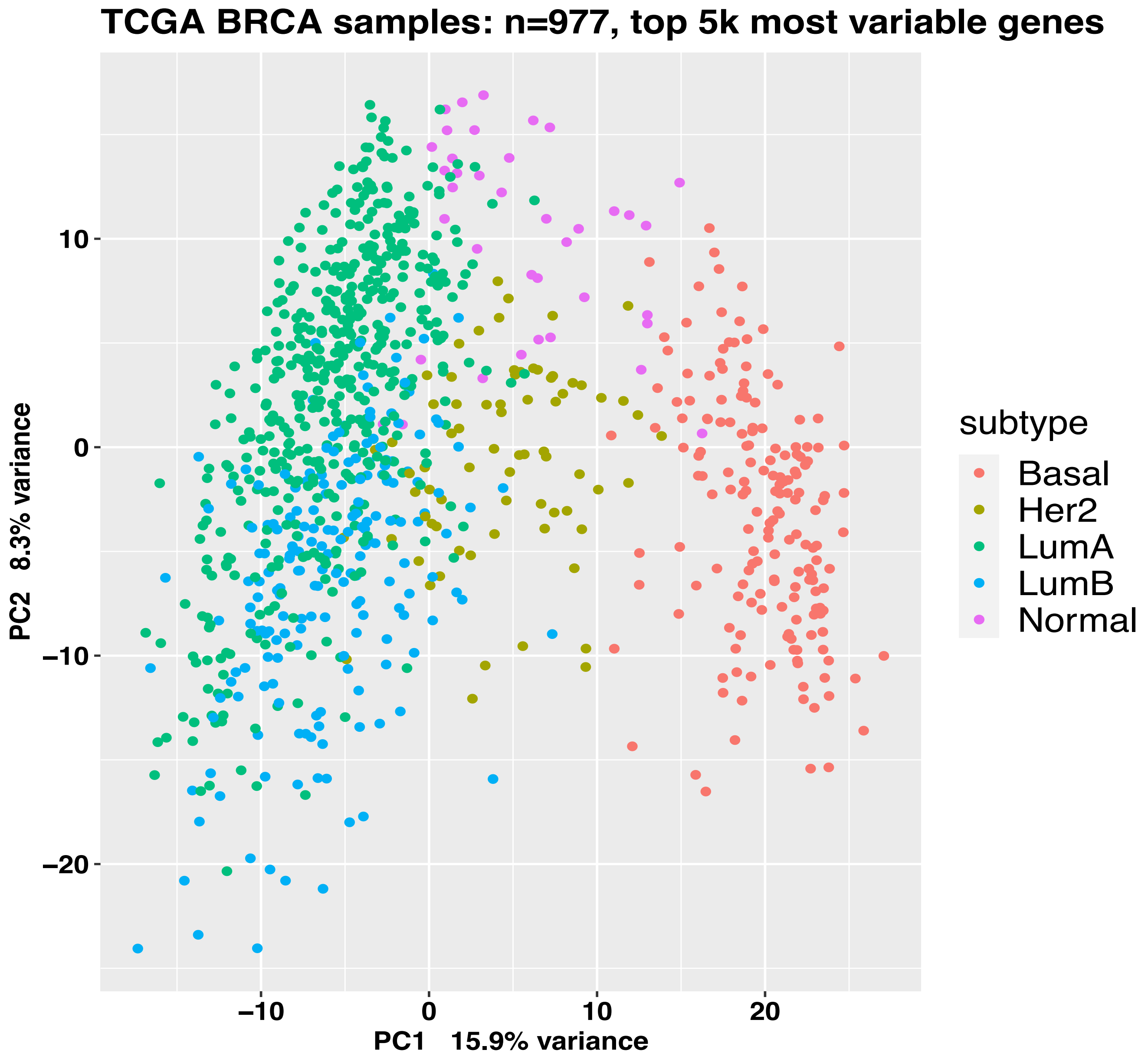
	Group1		Group2		Group3	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
d1	0	0	6	0	0	6
d2	0	0	20	0	0	20



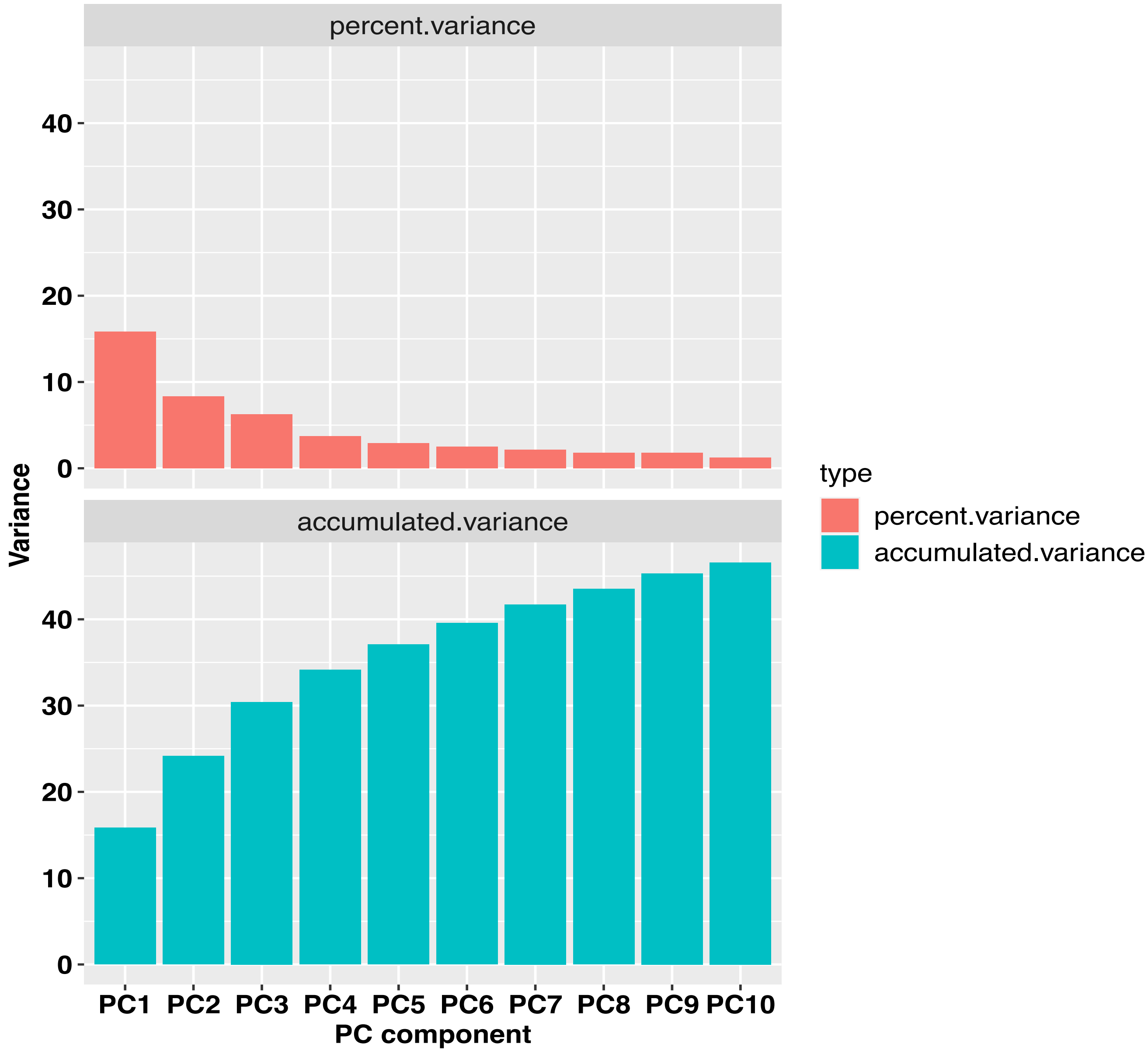
Variance Accounted for by PC1



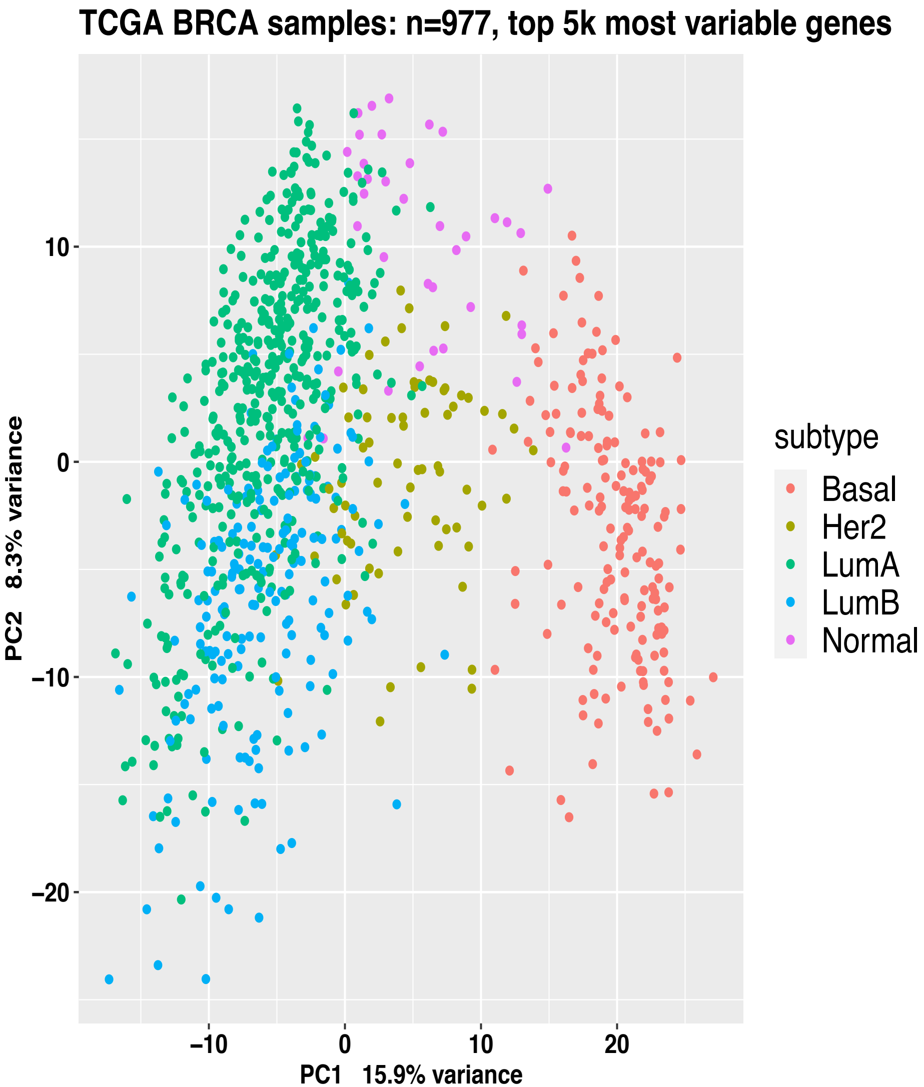
PCA Analysis of TCGA Breast Cancer RNAseq Data

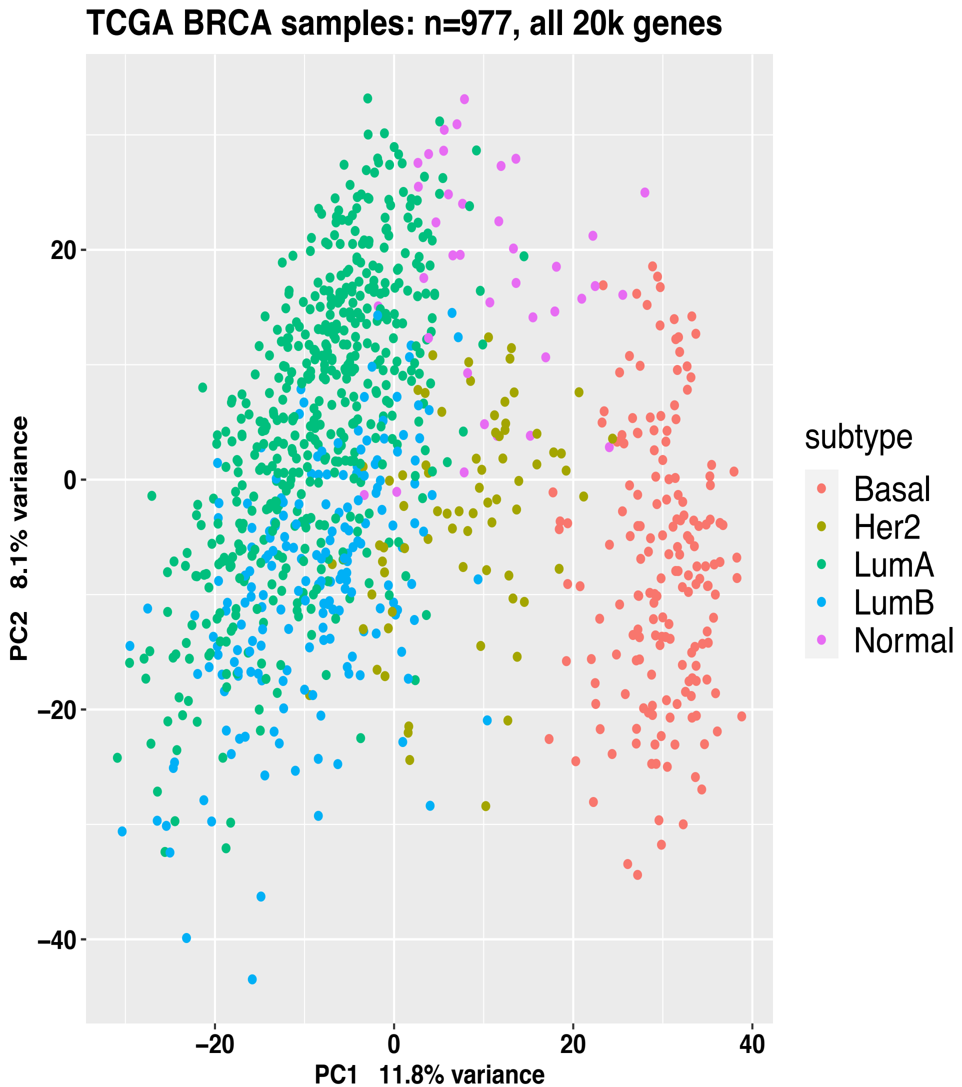


Variance of Principal Components Are Ranked from the Highest to the Lowest

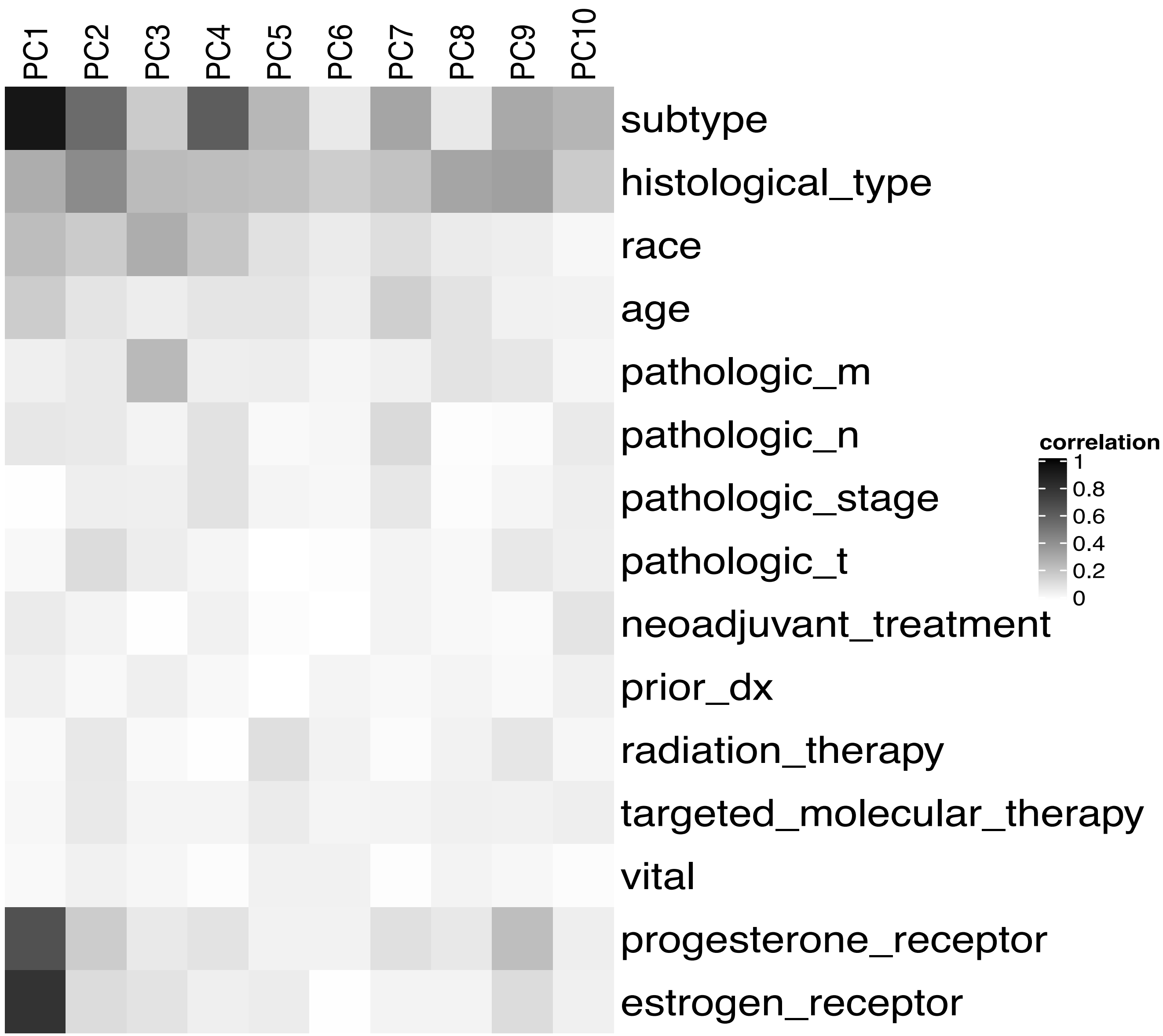


Filtering Out Genes of Low Variance Increases Percent of Variance Accounted for by PC1

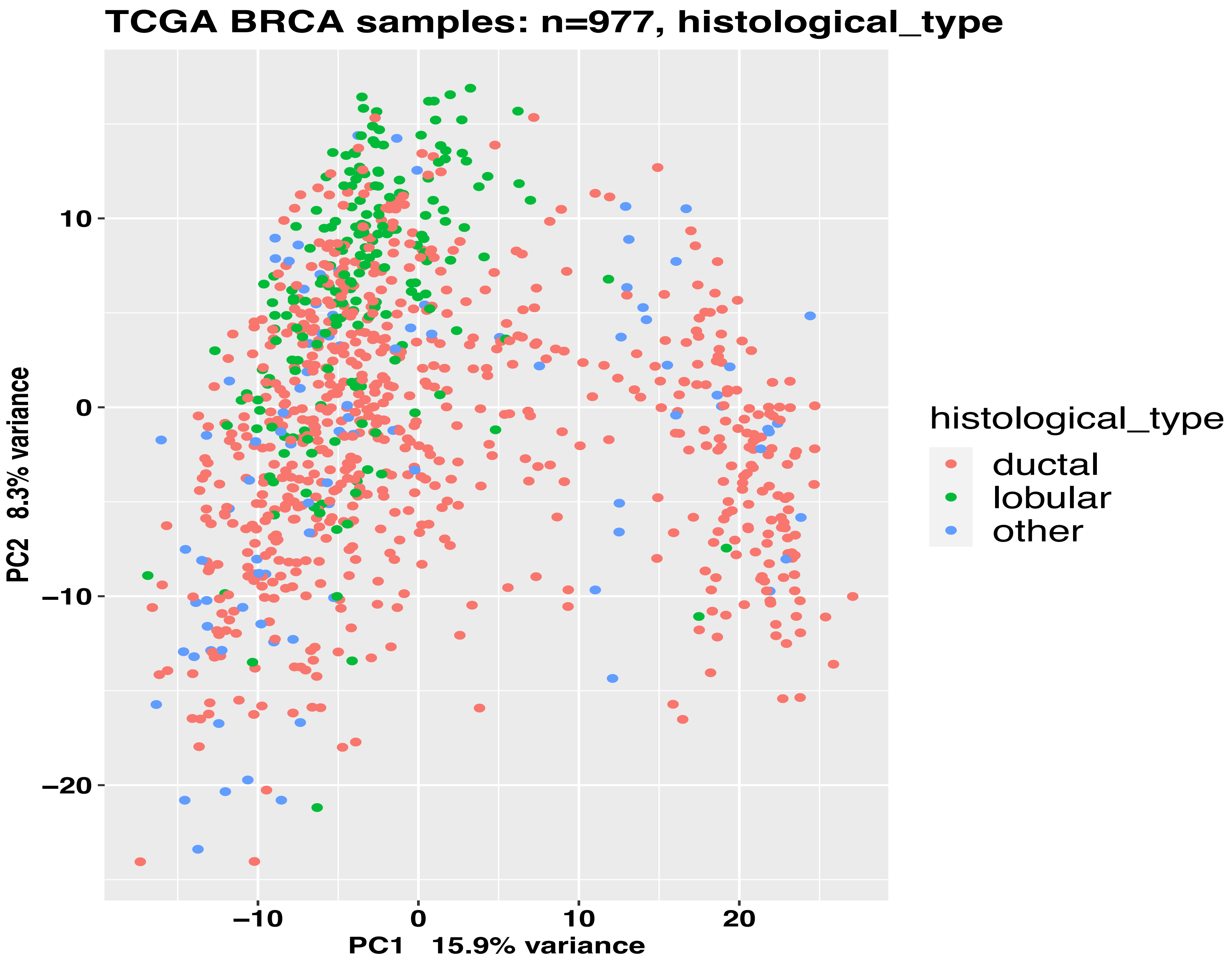




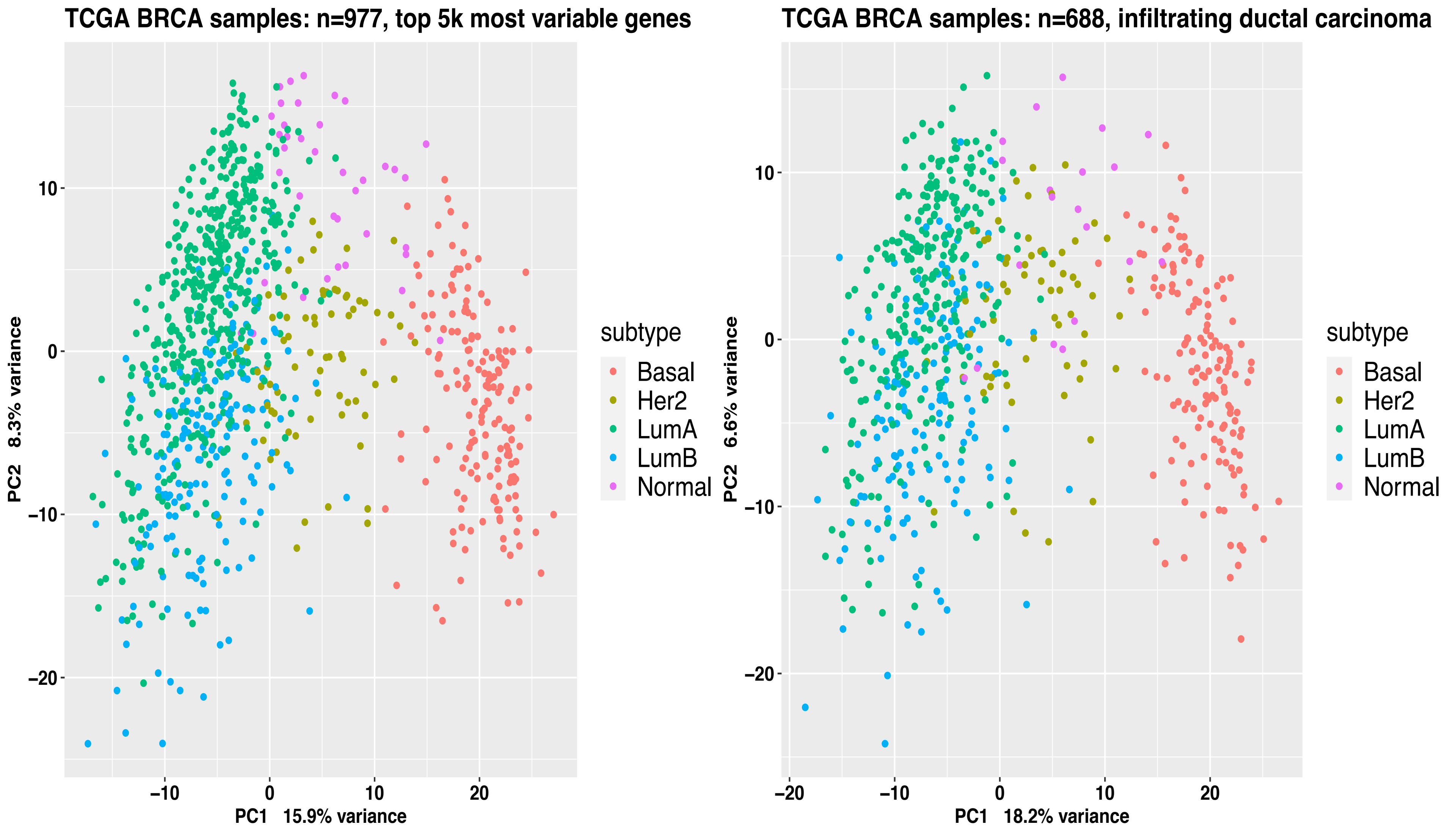
Correlation Between Principal Components and Phenotypes of Breast Cancer Data



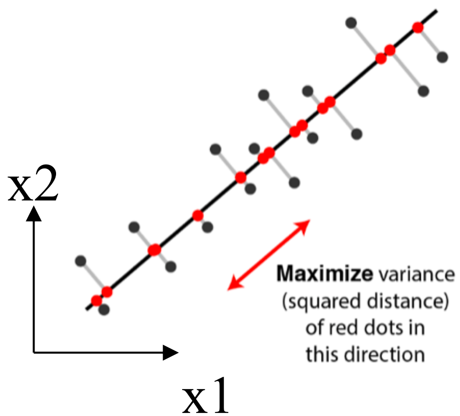
Variation in Histological Type Is Associated with PC2

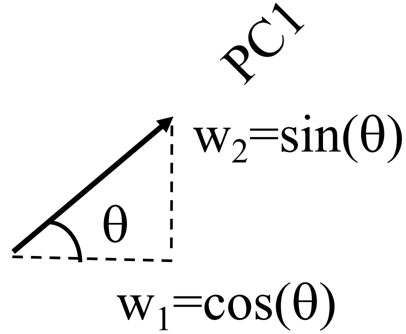


Removing Heterogeneity in Histological Type Reduces PC2 Variance and Increases PC1 Variance



Algorithm of PCA: How Does PCA Find the Direction of PC1?





$$\underbrace{\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ . & . \\ . & . \\ x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}}_{Xw} = \underbrace{\begin{bmatrix} w_1 x_{11} + w_2 x_{12} \\ w_1 x_{21} + w_2 x_{22} \\ . \\ . \\ w_n x_{21} + w_2 x_{n2} \end{bmatrix}}_{z}$$

$$z = Xw$$

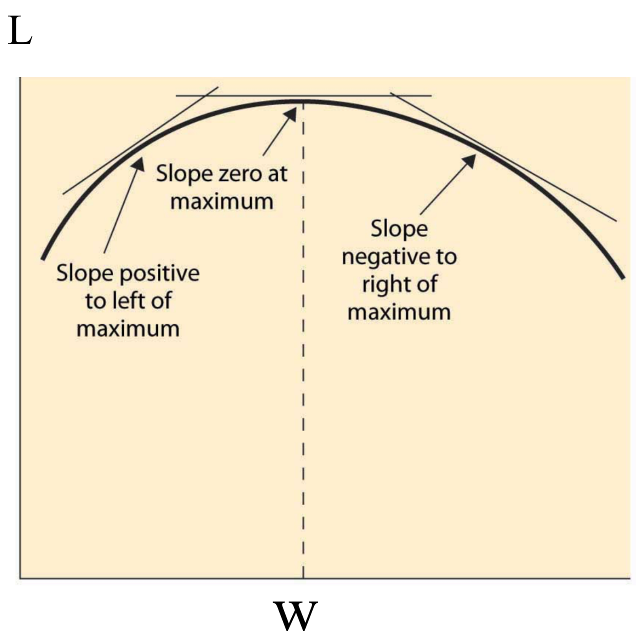
$$\mathrm{var}(z) = (Xw)^T Xw$$

$$\mathrm{var}(z) = w^T X^T X w = w^T S w$$

Choose w to maximize $w^T S w$

subject to $w^T w = 1$

The Direction of PC1 Is the Eigen Vector with the Highest Eigen Value



Choose w to maximize $w^T S w$
subject to $w^T w = 1$

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

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

$$Sw = \lambda w$$

w is the eigen vector and λ is eigen value

Variance of PCs Are Eigen Value and Are Additive

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

There are p pairs of eigen vectors and eigen values

$$\text{var}(Z) = \lambda_1 + \lambda_2 \ldots + \lambda_p$$