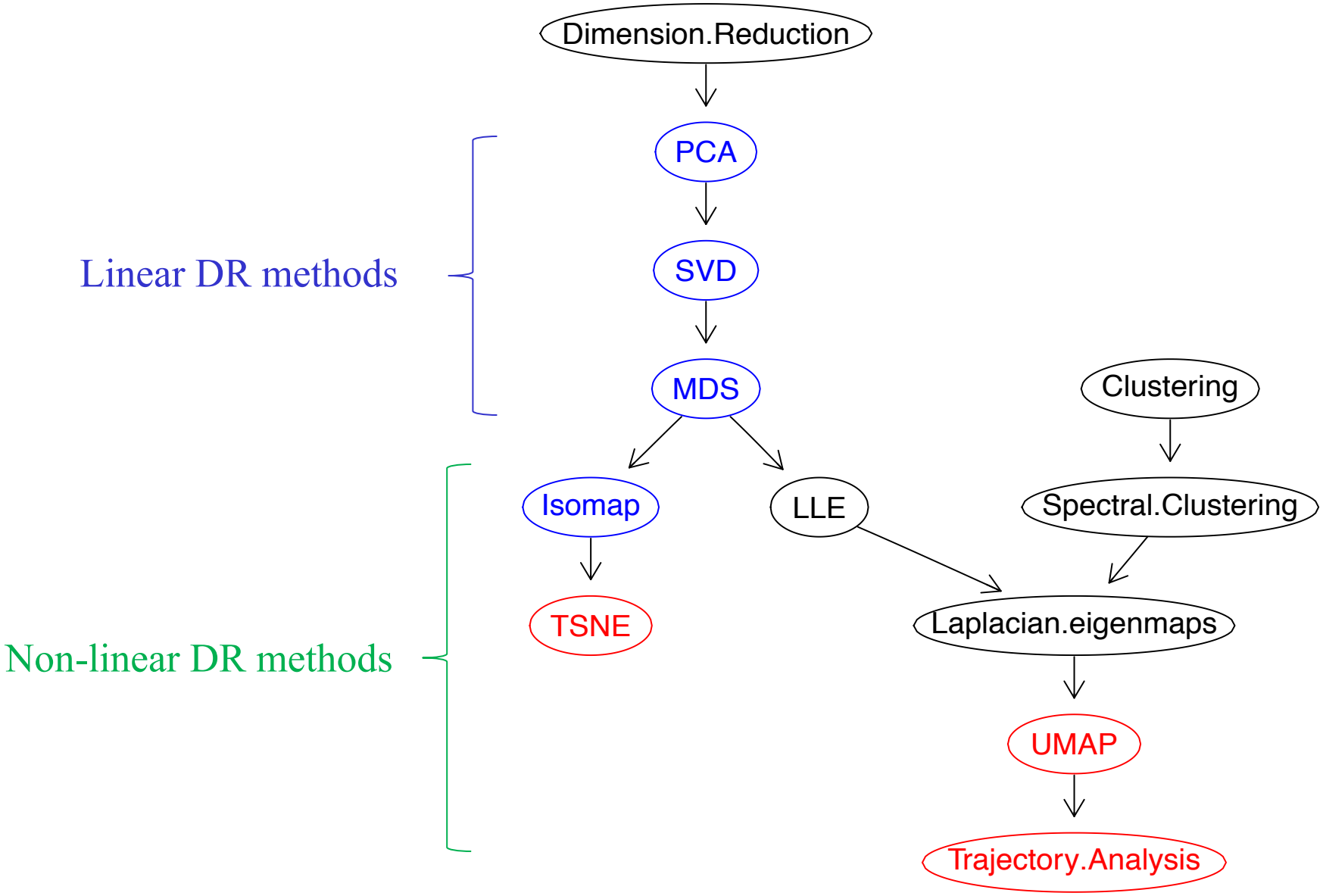
Dimension Reduction Methods:
From PCA to TSNE and UMAP

Maxwell Lee

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

May 14, 2020

Road Map for Dimension Reduction Methods



Comparison of PCA, TSNE, and UMAP

	Data type	Sample size	complexity	Performance
MNIST	image	6000	High	UMAP > TSNE > PCA
ScRNAseq	ScRNAseq	~6000	High?	UMAP >= TSNE > PCA
TCGA	Bulk RNAseq	~1000	moderate	UMAP ~ TSNE ~ PCA

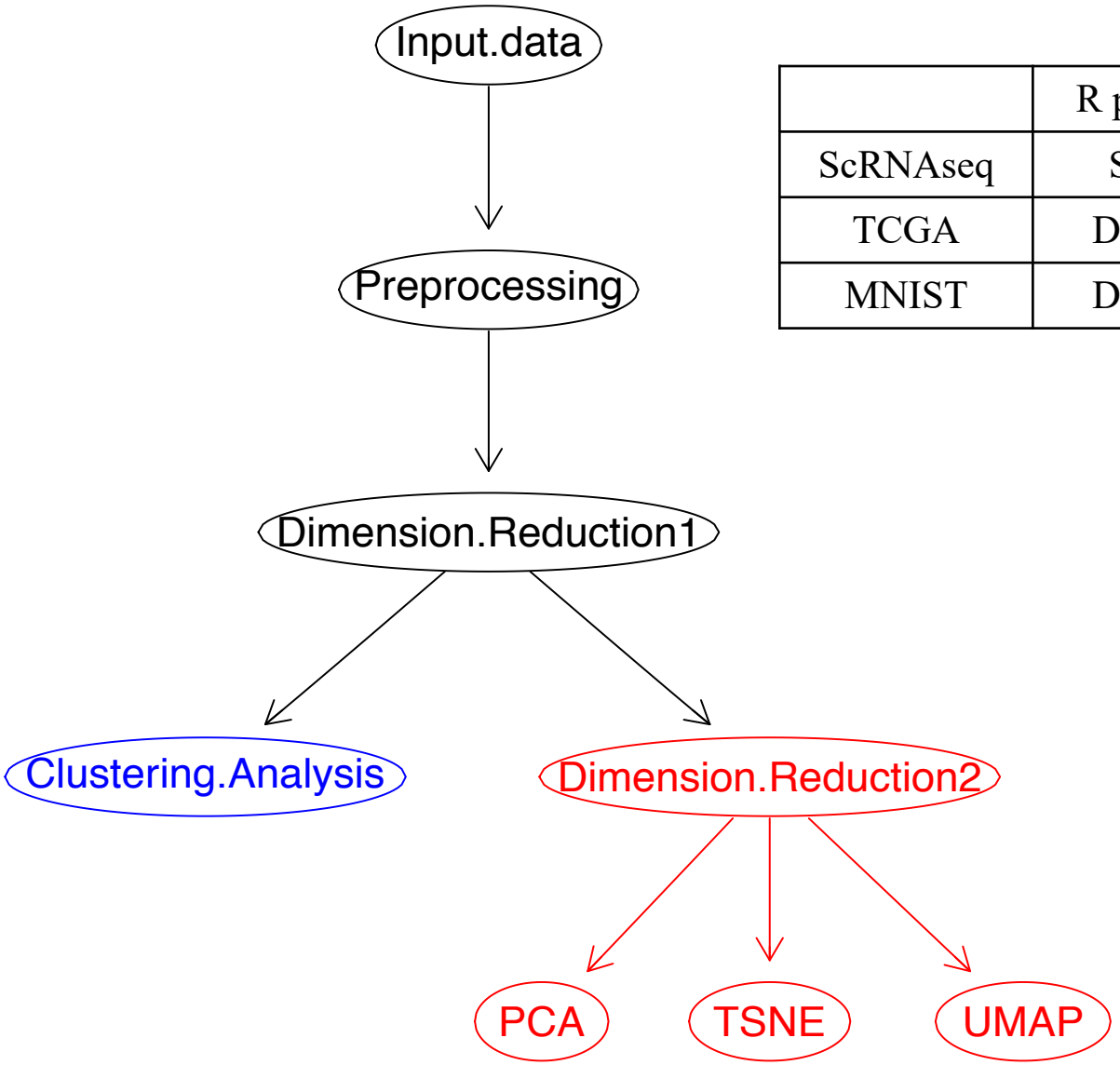
Frequently Asked Questions

1) Which method should I use, PCA, TSNE, or UMAP?

2) How many samples do I need to use these methods?

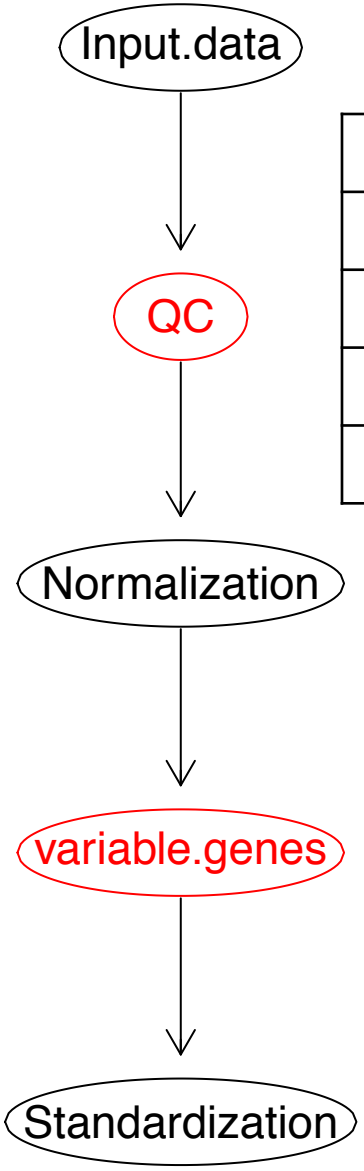
3) How to choose parameters of the analysis?

Flow Chart of PCA, TSNE, and UMAP Analyses



	R package
ScRNAseq	Seurat
TCGA	DimRed
MNIST	DimRed

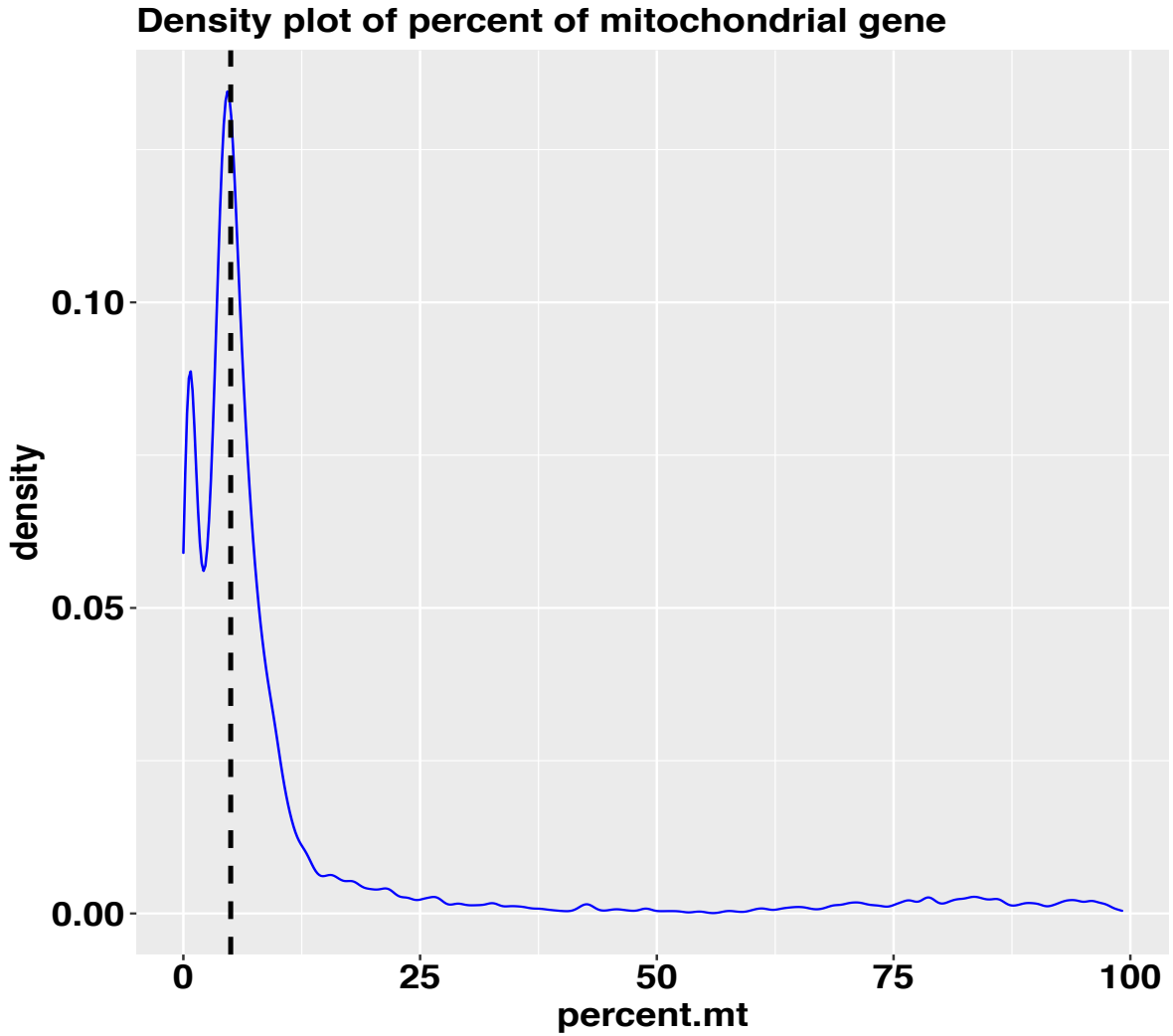
Preprocessing Steps in Seurat Package



Preprocessing	function	Description
QC	Select cells	percent.mt < 5%
Normalization	Normalizing cells	TP10K
Variable genes	Most variable genes	nfeatures = 2000
Standardization	Standardization across cells	z score

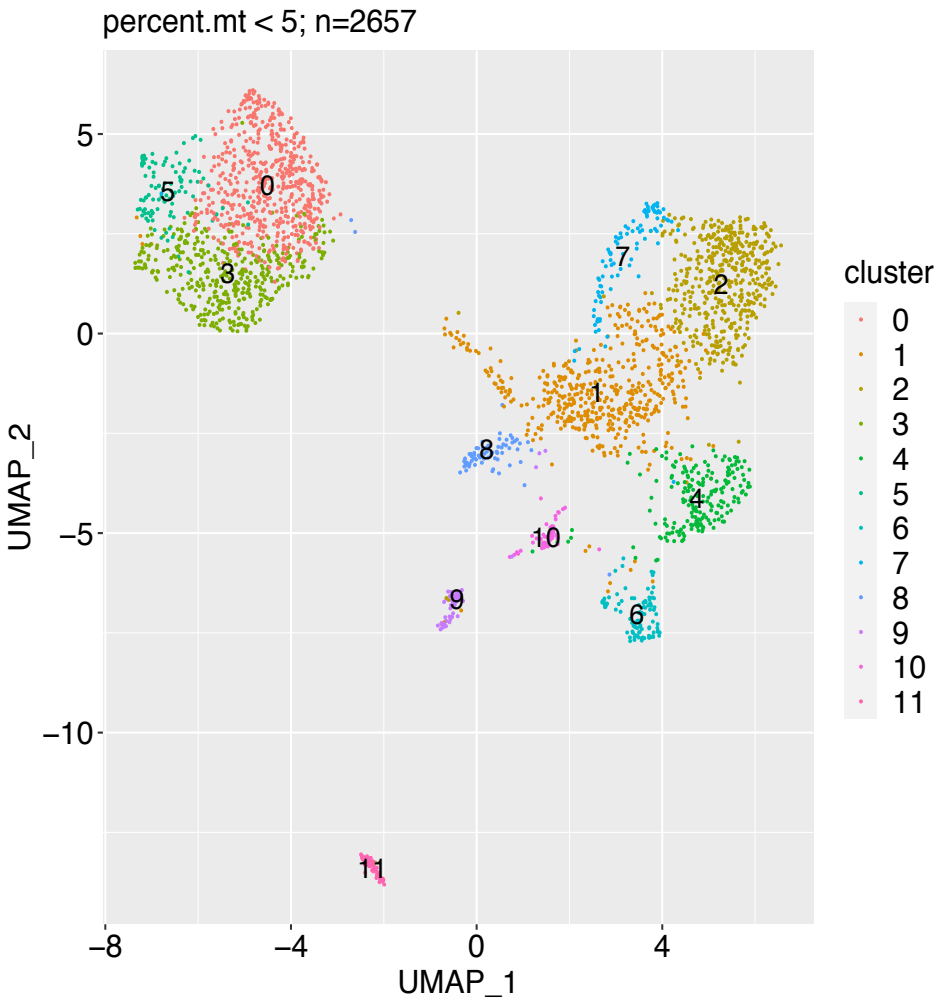
Density Plot of Percent of Mitochondrial Genes

Increased percent of mitochondrial genes is associated with cells undergoing apoptosis

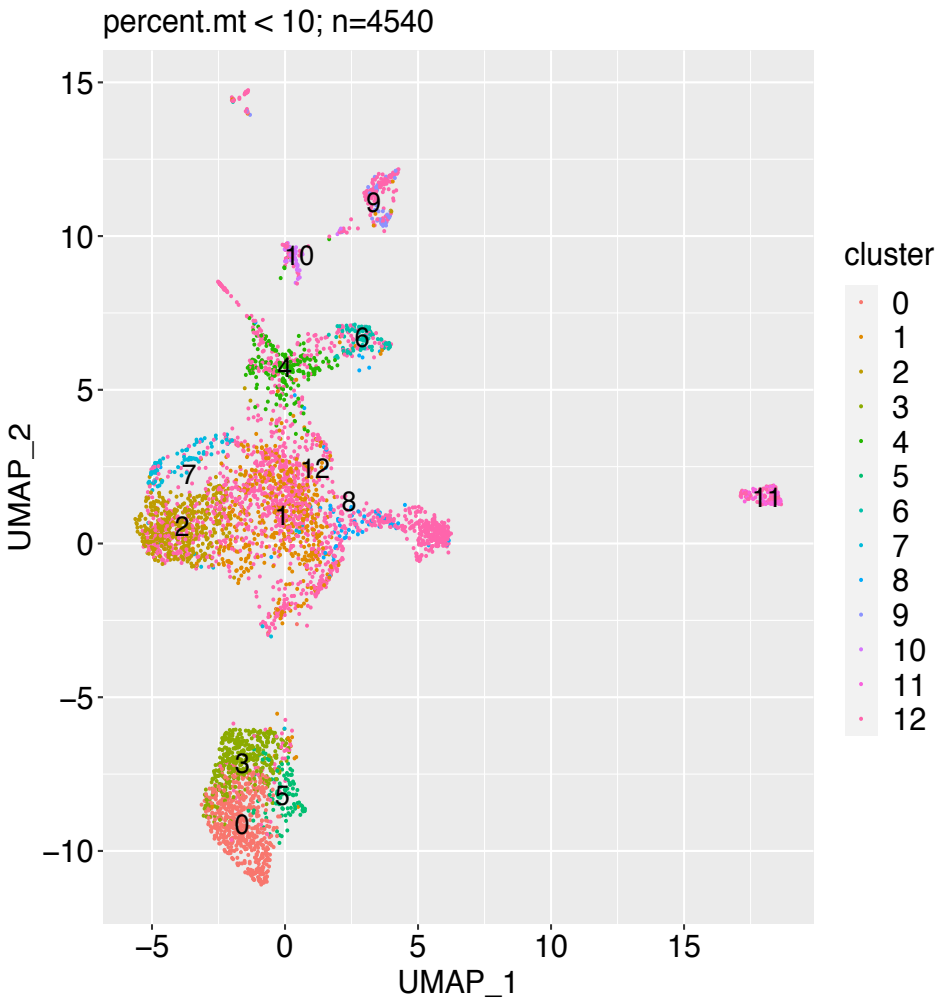


Effects of Using Percent of Mitochondrial gene Cutoff on UMAP

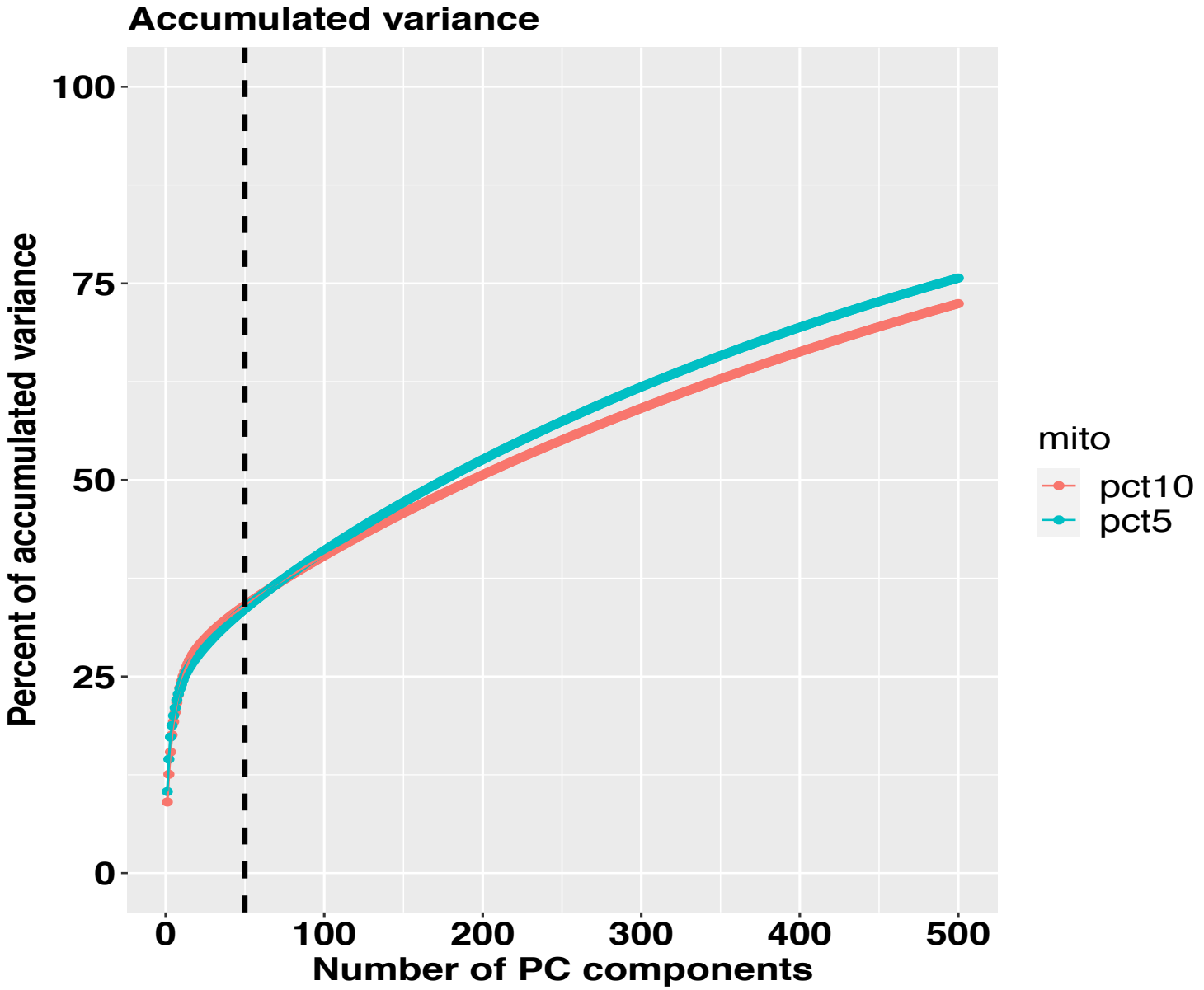
Clustering and Dimension Reduction 2



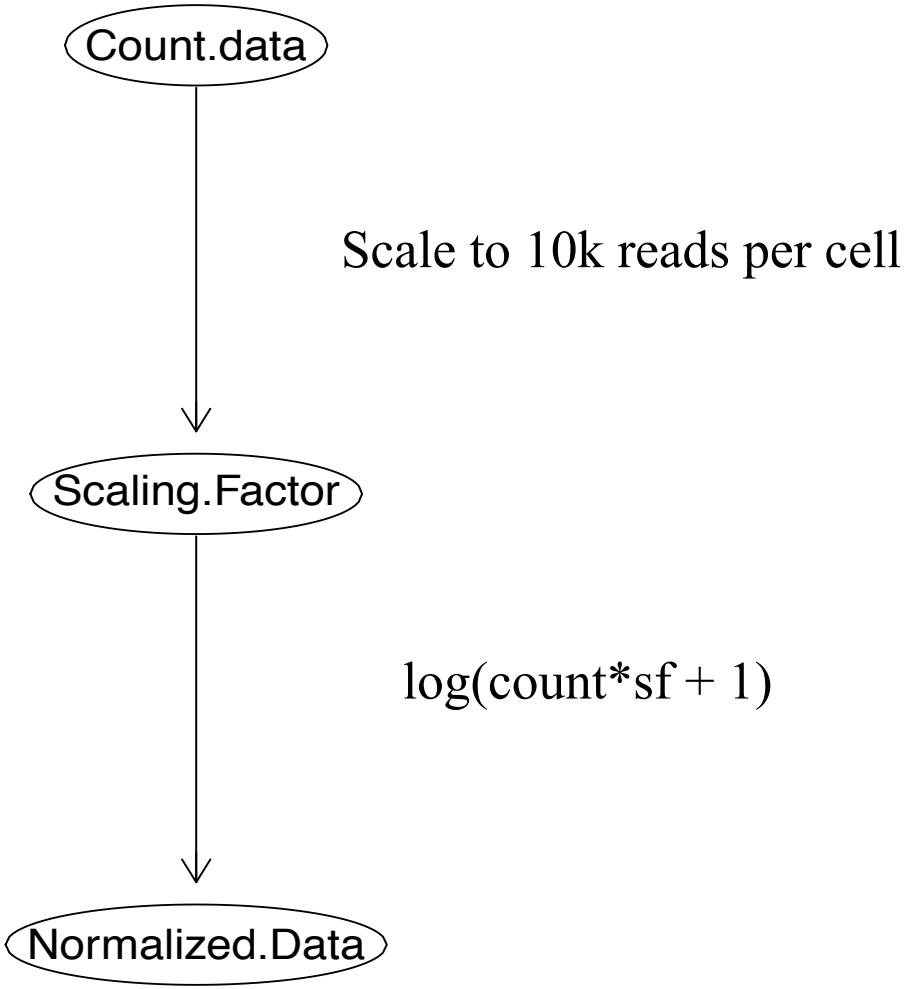
Dimension Reduction 2
Clusters 0-11 are identical to the left plot
Cluster 12 has the additional cells



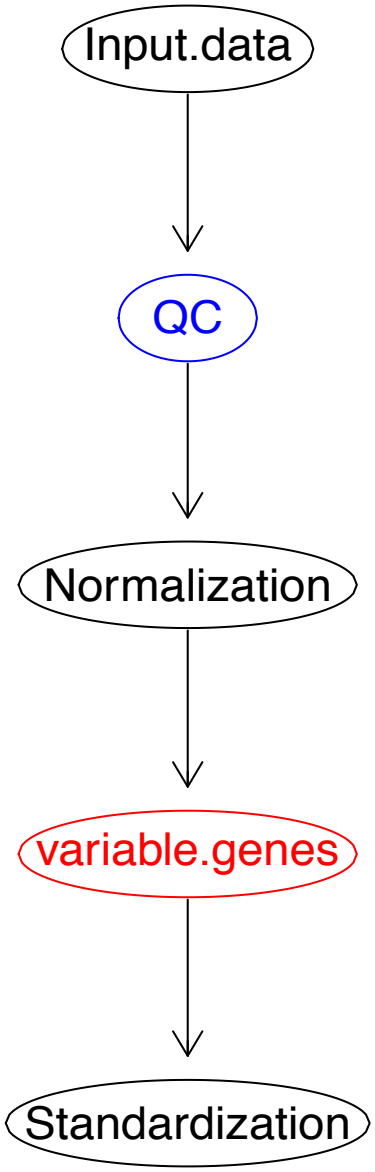
Accumulated Variance with Dimension Reduction I



Normalization in Preprocessing

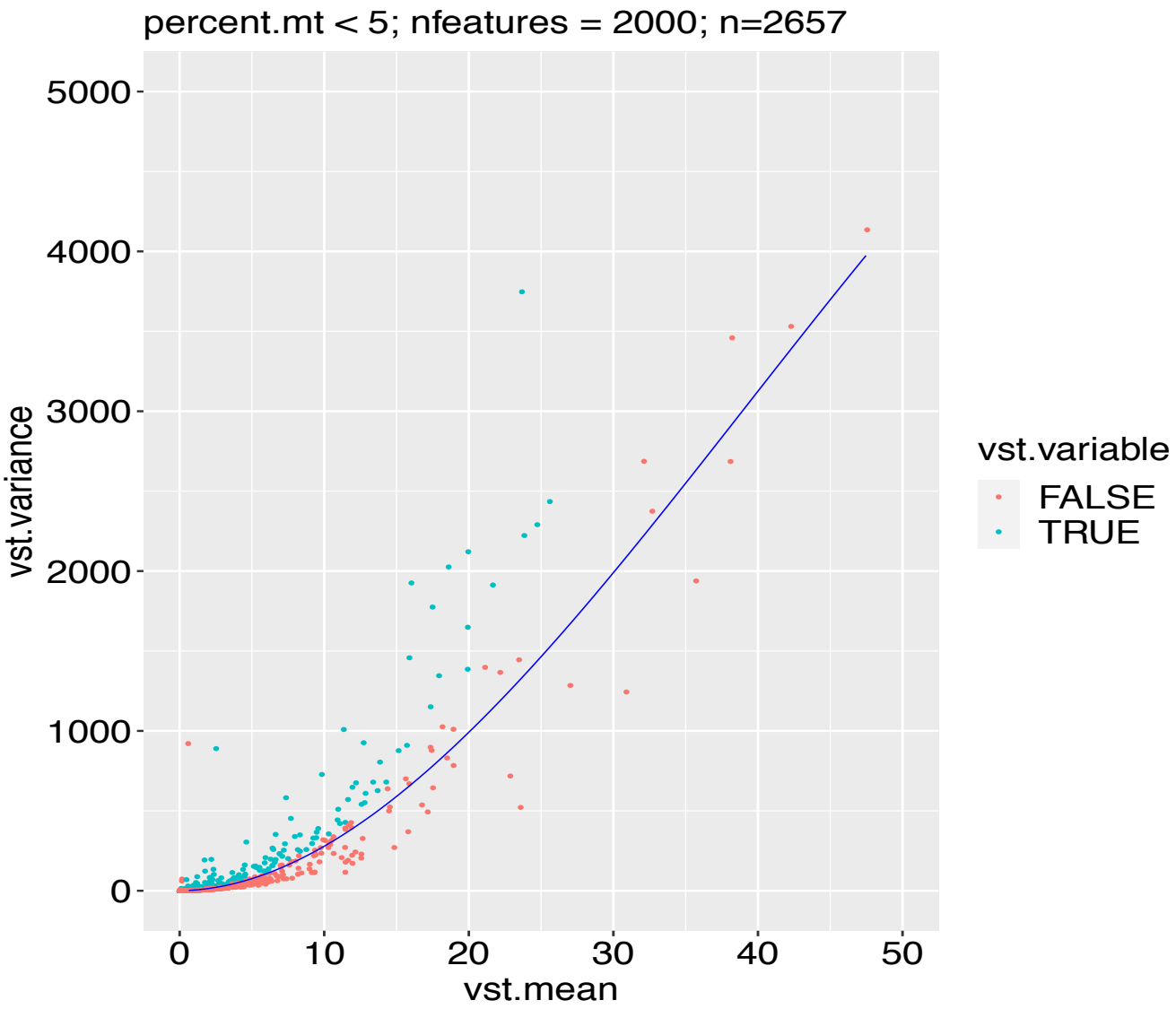


Select Most Variable Genes in Preprocessing



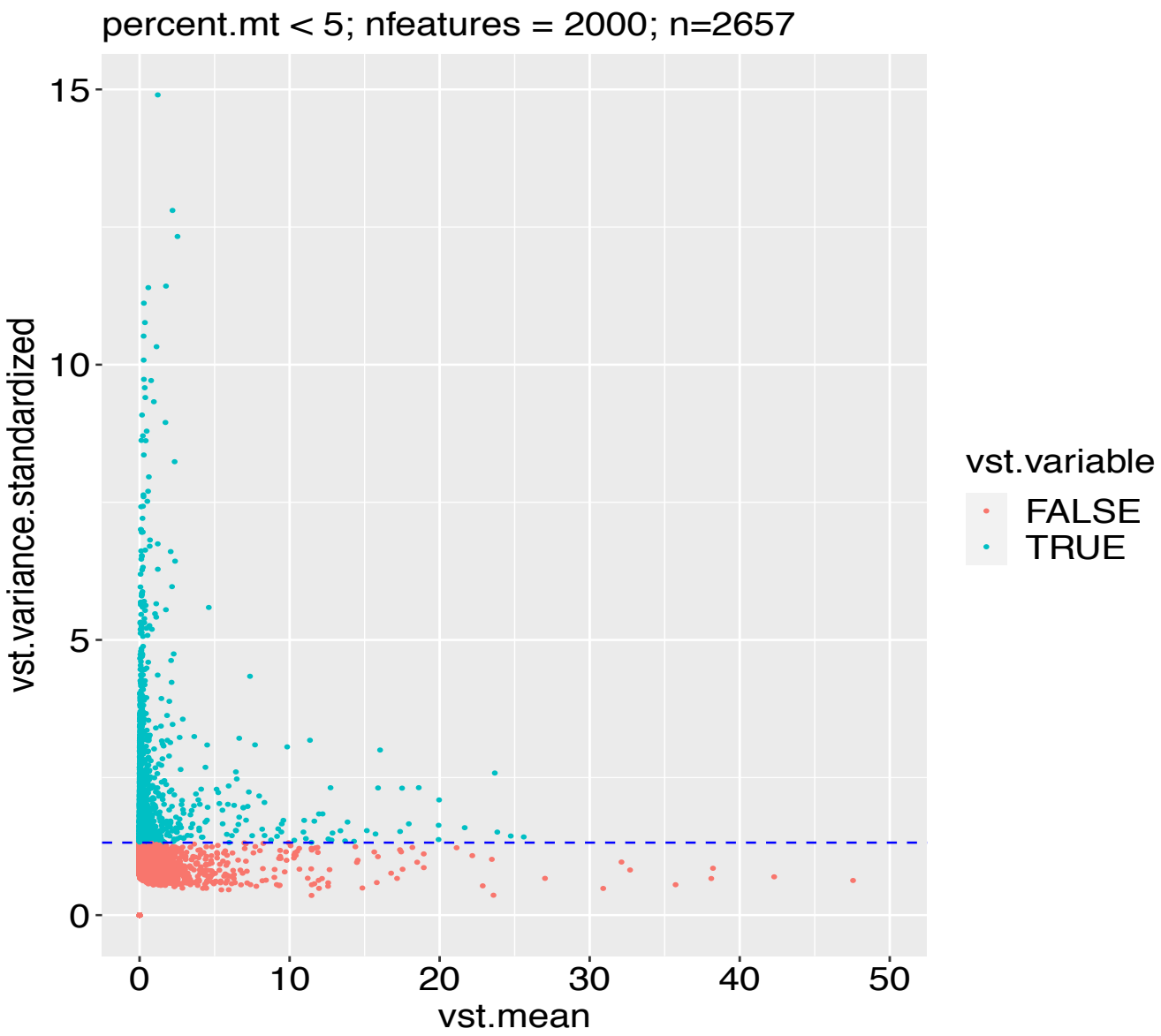
Preprocessing	function	Description
QC	Select cells	percent.mt < 5%
Normalization	Normalizing cells	TP10K
Variable genes	Most variable genes	nfeatures = 2000
Standardization	Standardization across cells	z score

How to Find Most Variable Genes?



vst: variance-stabilizing transformation

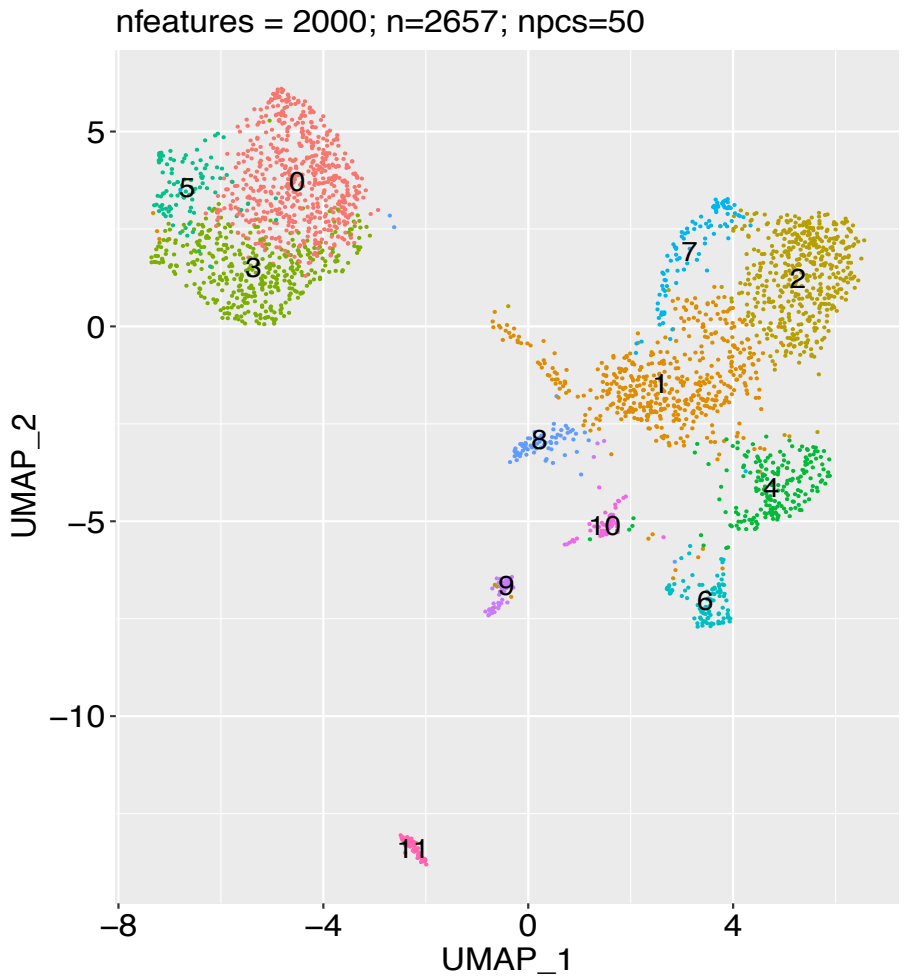
How to Find Most Variable Genes?



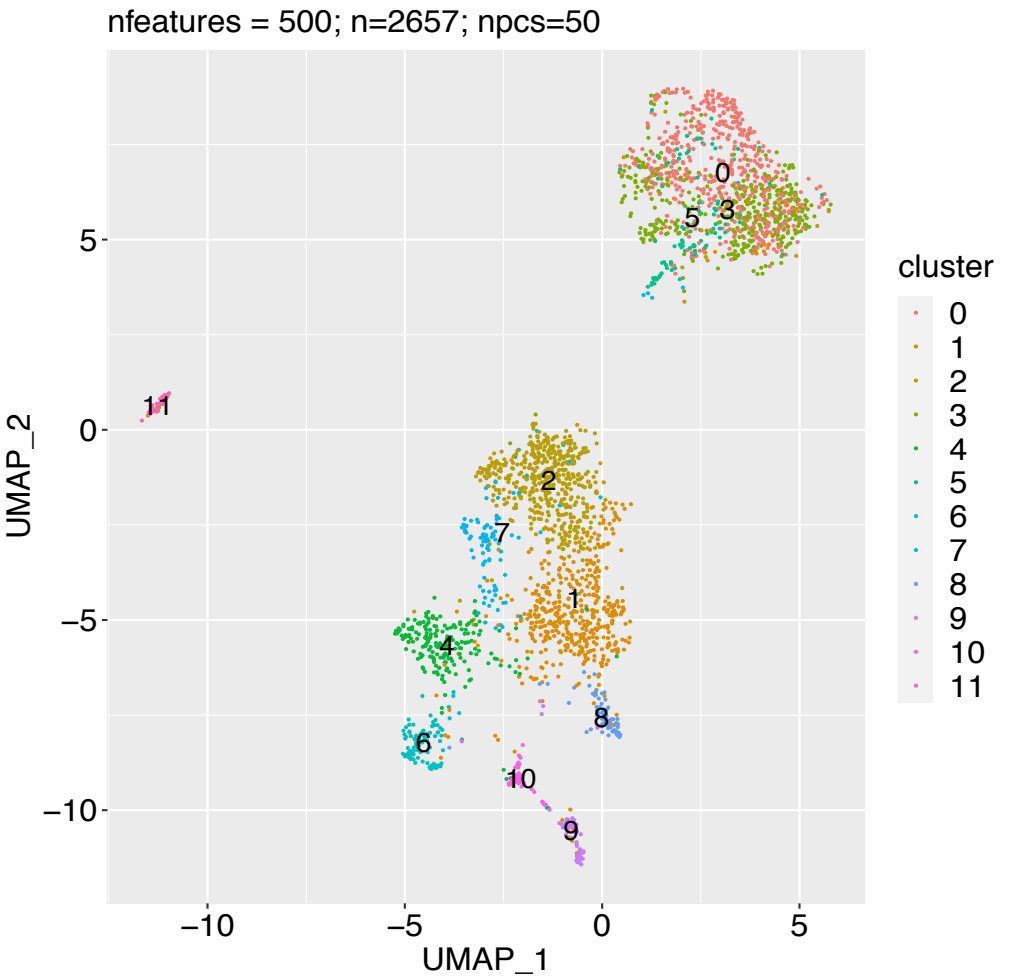
Effects of Using the Number of Genes Cutoff on UMAP

Clustering and Dimension Reduction 2
nfeatures = 2000

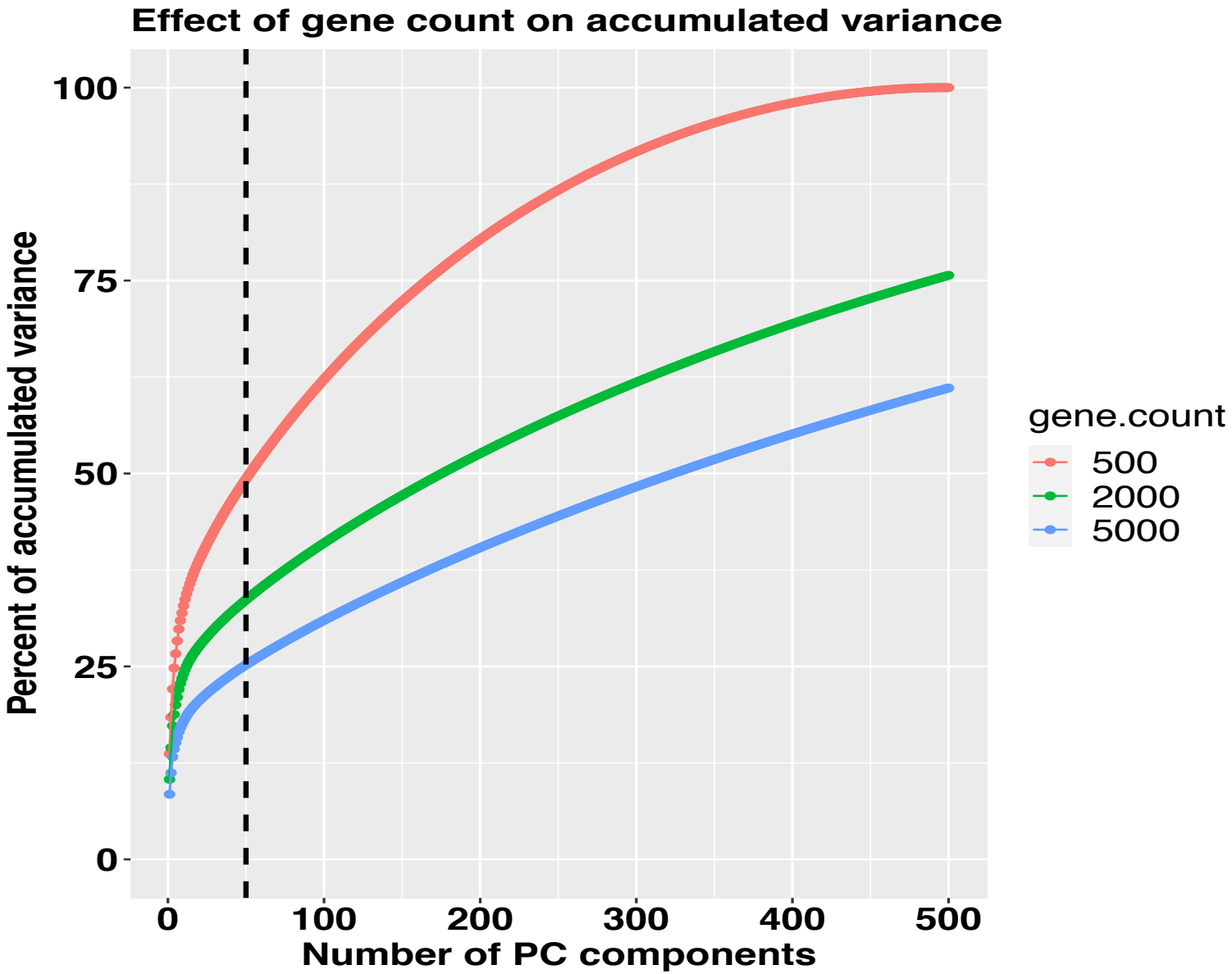
npcs: number of PCs



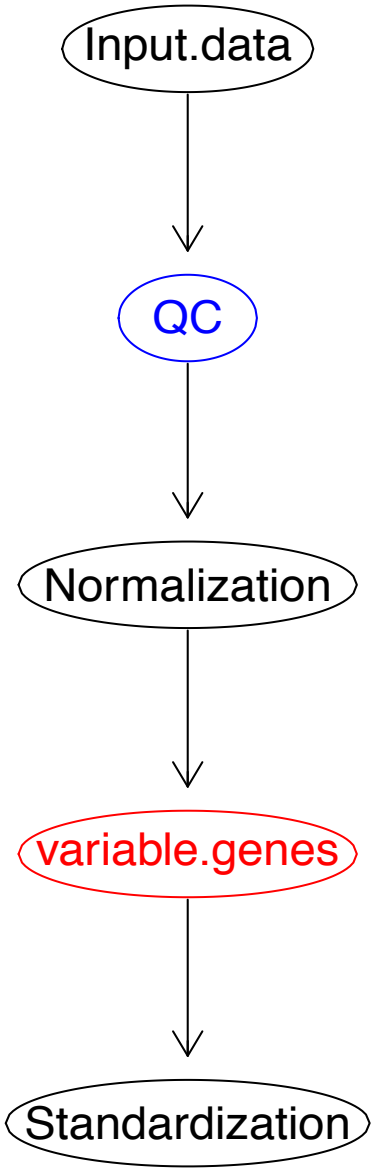
Dimension Reduction 2
nfeatures = 500
Clusters 0-11 are identical to the left plot



Accumulated Variance with Dimension Reduction I

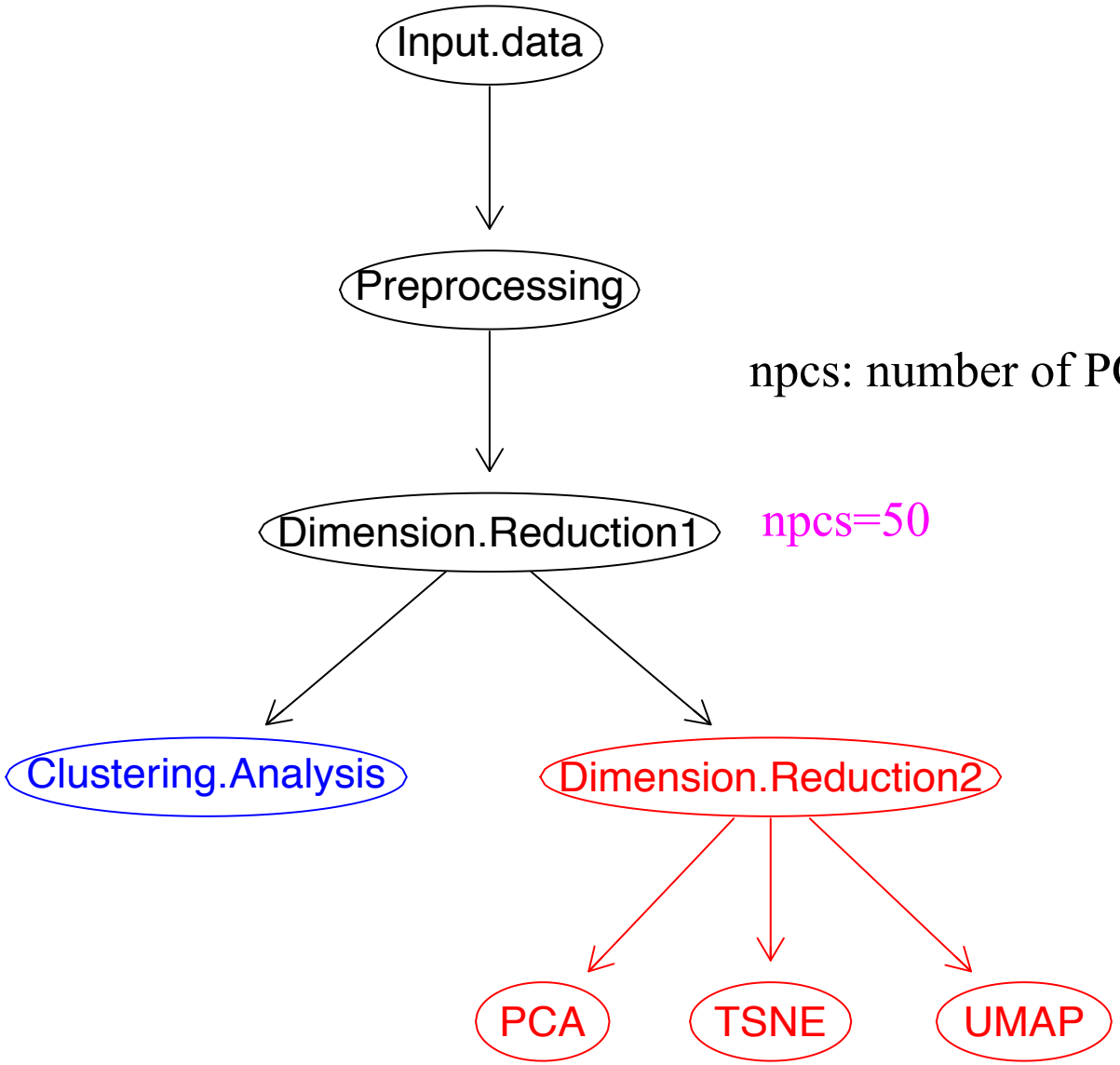


Select Most Variable Genes in Preprocessing



Preprocessing	function	Description
QC	Select cells	percent.mt < 5%
Normalization	Normalizing cells	TP10K
Variable genes	Most variable genes	nfeatures = 2000
Standardization	Standardization across cells	z score

Flow Chart of PCA, TSNE, and UMAP Analyses



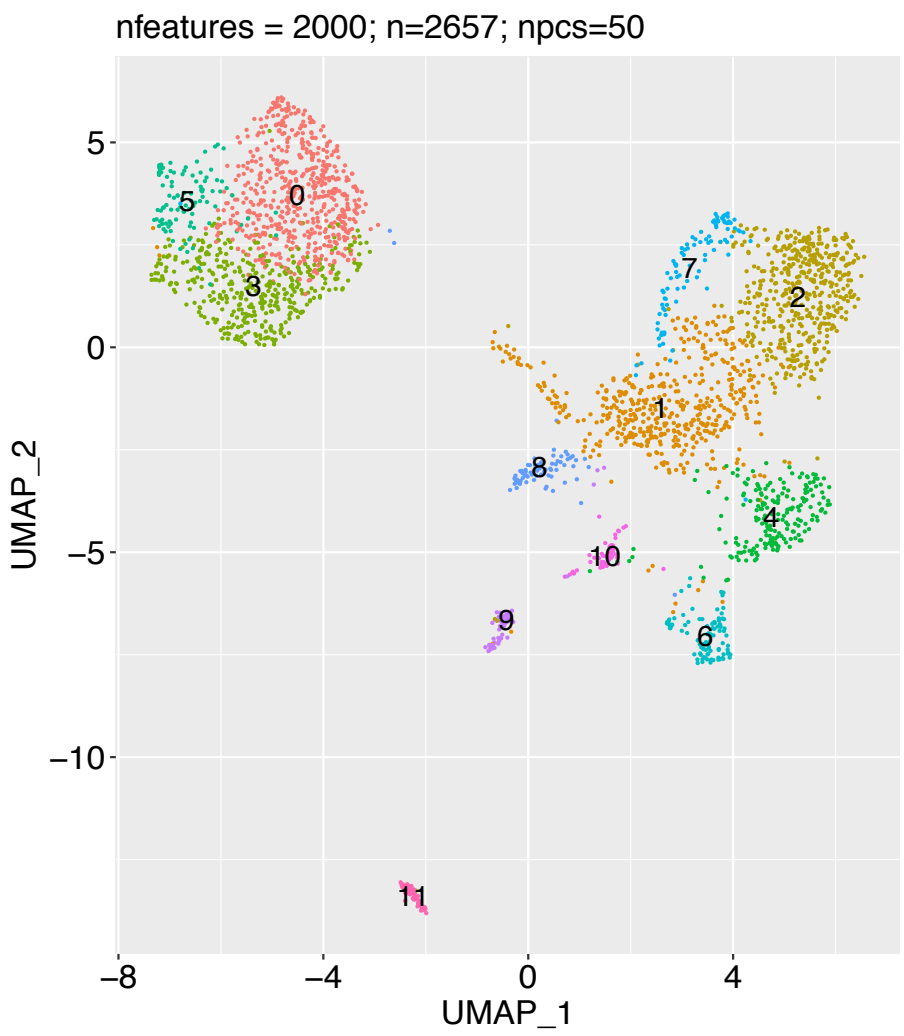
npcs: number of PC components

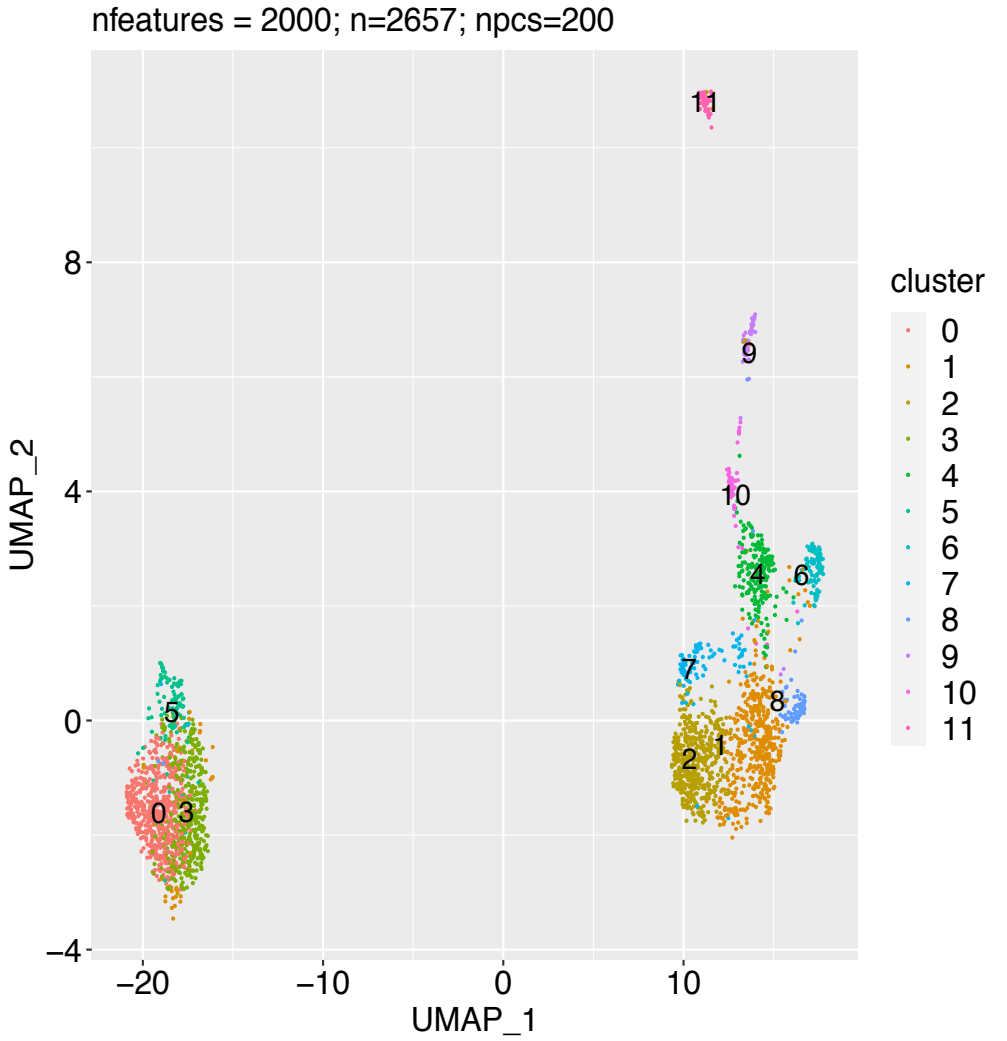
npcs=50

Effects of Number of Principal Components on UMAP

Clustering and Dimension Reduction 2
npcs = 50

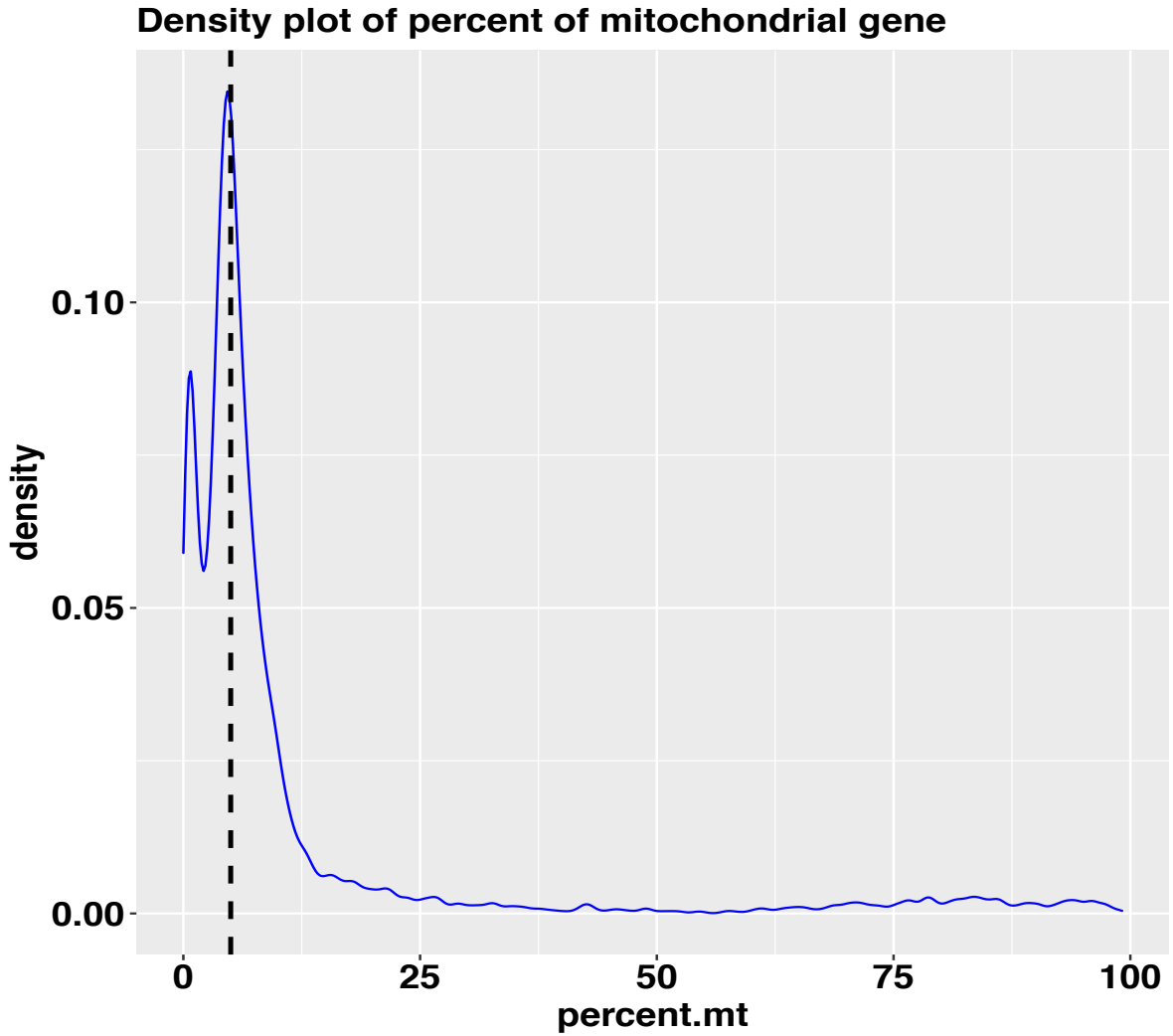
Dimension Reduction 2
npcs = 200
Clusters 0-11 are identical to the left plot





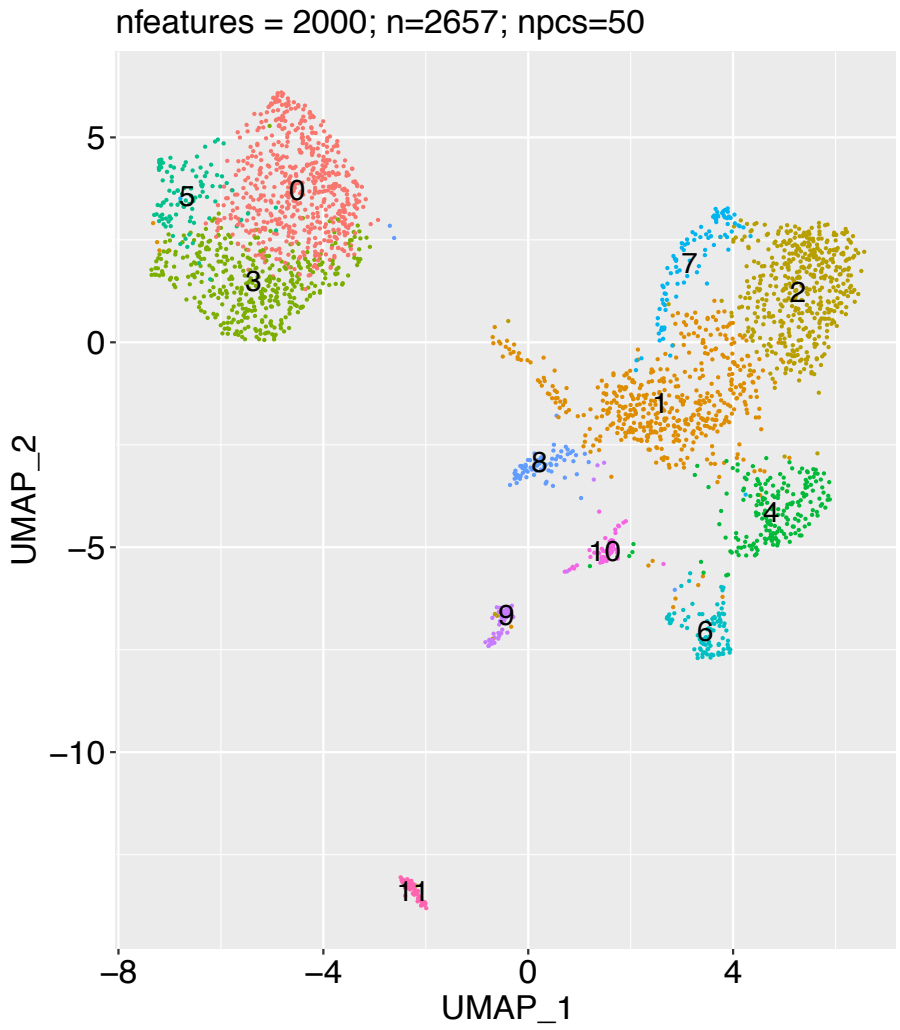
Density Plot of Percent of Mitochondrial Genes

Increased percent of mitochondrial genes is associated with cells undergoing apoptosis

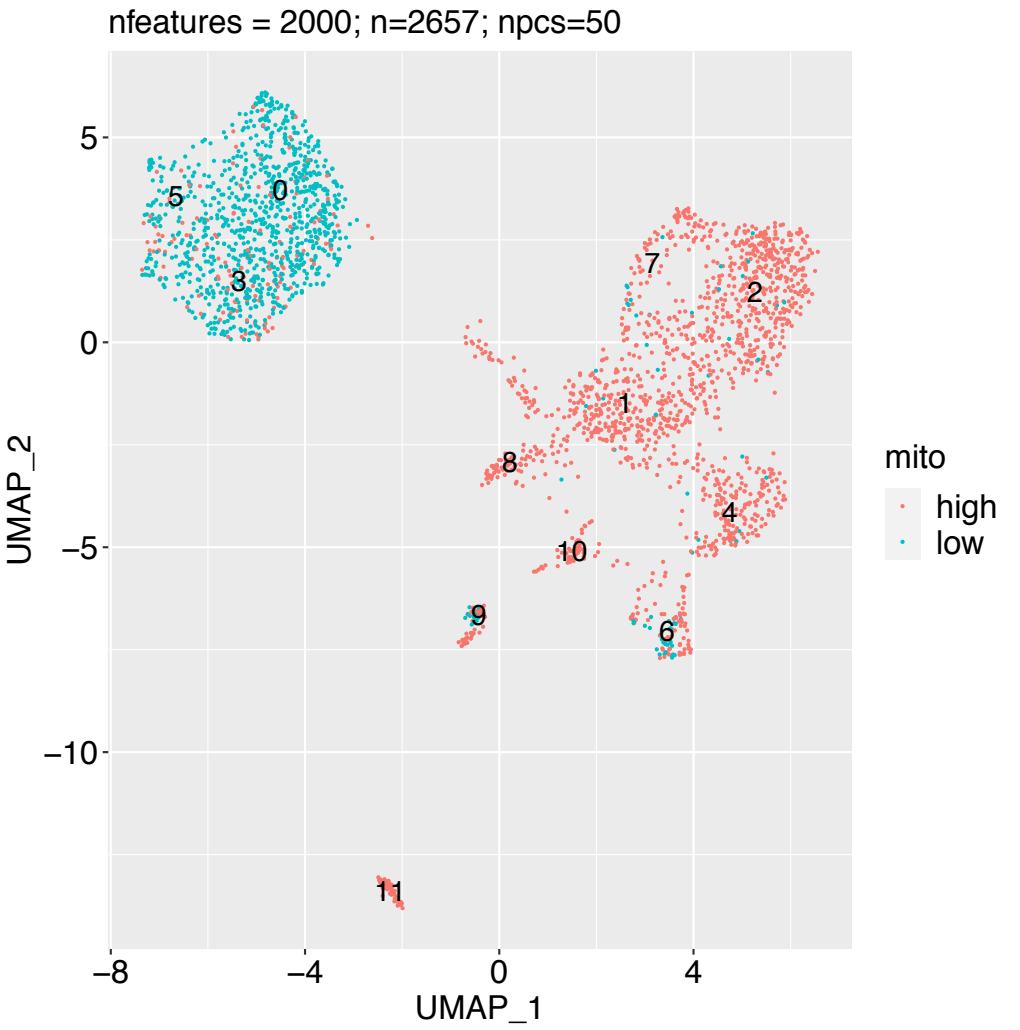


Percent of Mitochondrial Genes Low vs High

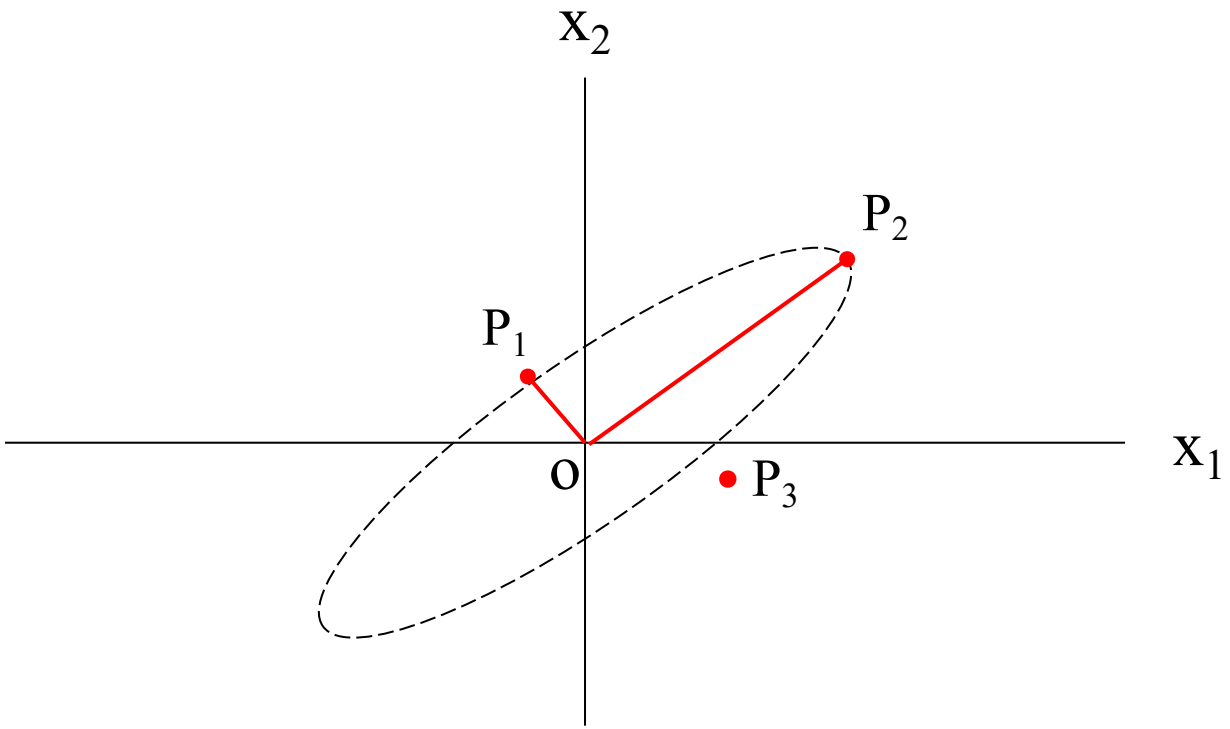
label with cluster id



label with percent of mitochondrial genes



Euclidean Distance vs Mahalanobis Distance



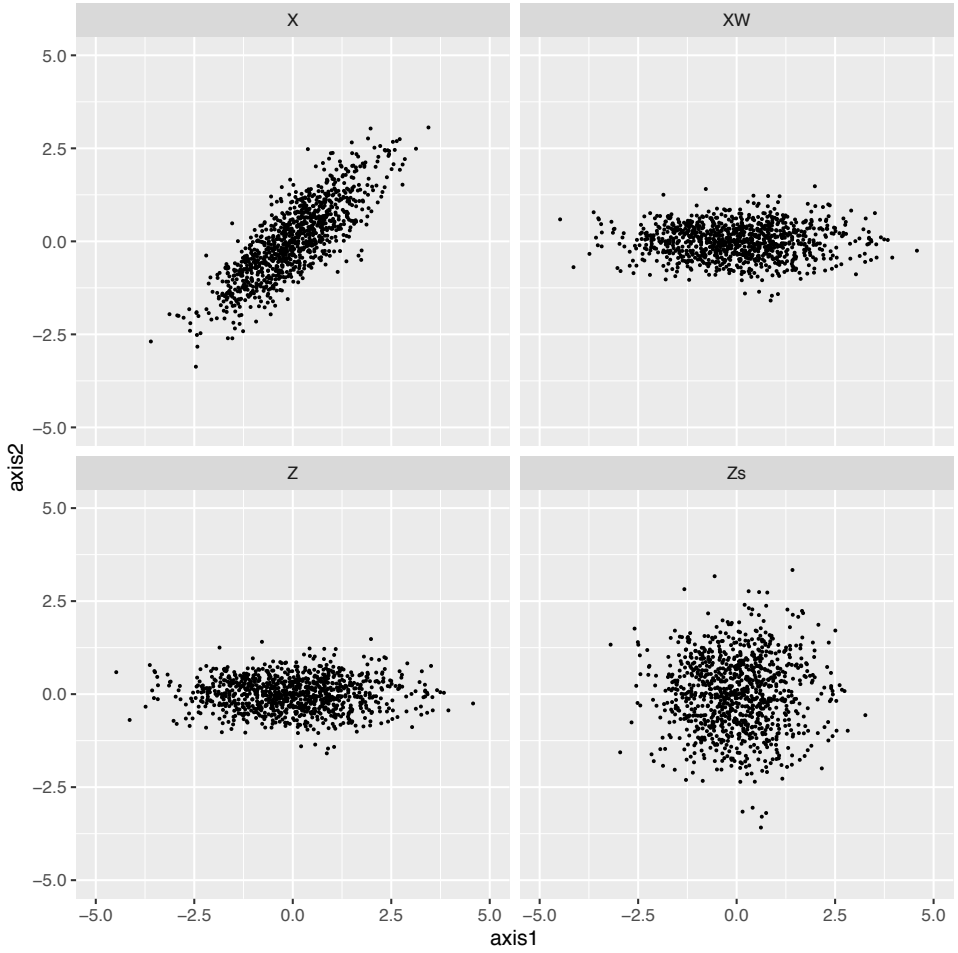
Euclidean distance: $P_1 < P_3 < P_2$

Probability: $p_1 = p_2 > p_3$

Mahalanobis distance is a statistical distance related to probability

Prasanta Chandra Mahalanobis in 1936

Multivariate Gaussian Distribution

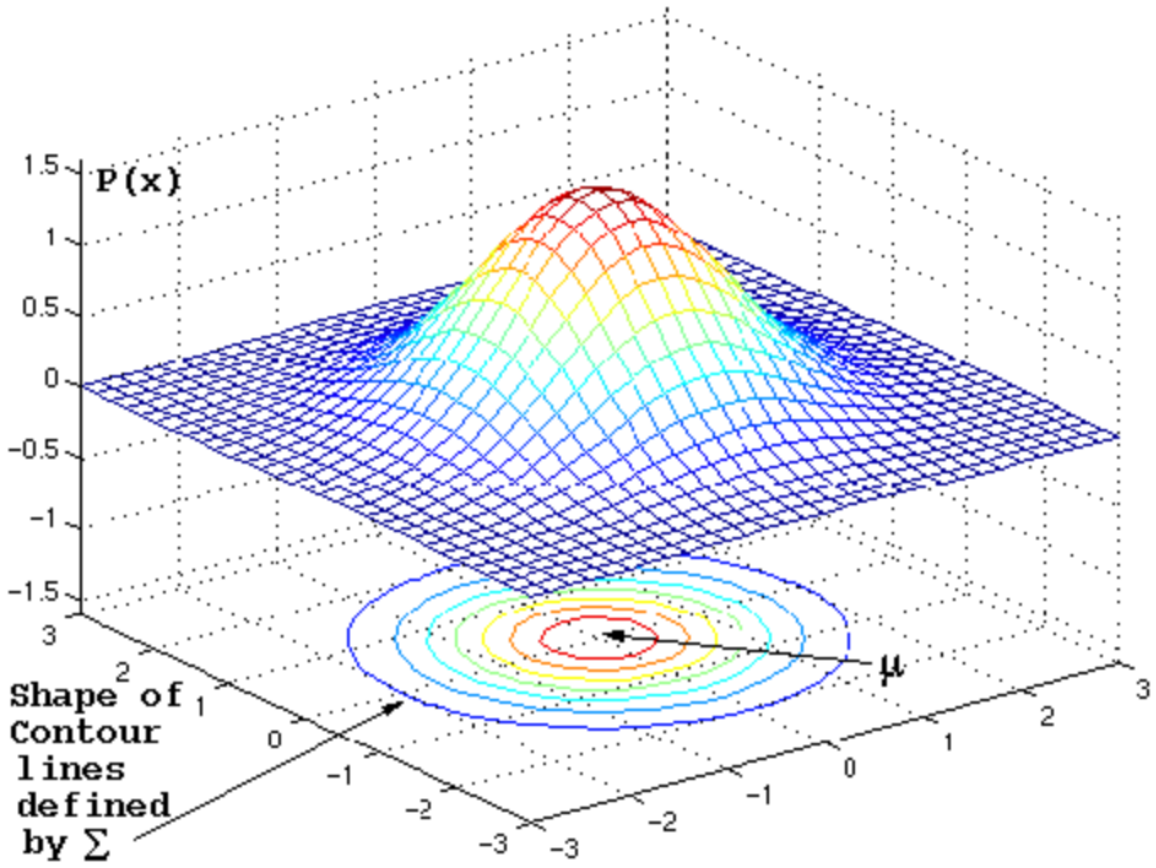


Σ: covariance matrix
Σ^{-1}: inverse of Σ
Λ: Diagonal matrix with Eigen values
W: Eigen vectors
Z: Principal Components
Z_s: Standardized Z
z: a sample from Z_s
T: Transposition
μ: mean vector

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

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

Multivariate Gaussian Distribution



T-distributed Stochastic Neighbor Embedding (TSNE)

Gaussian.kernel

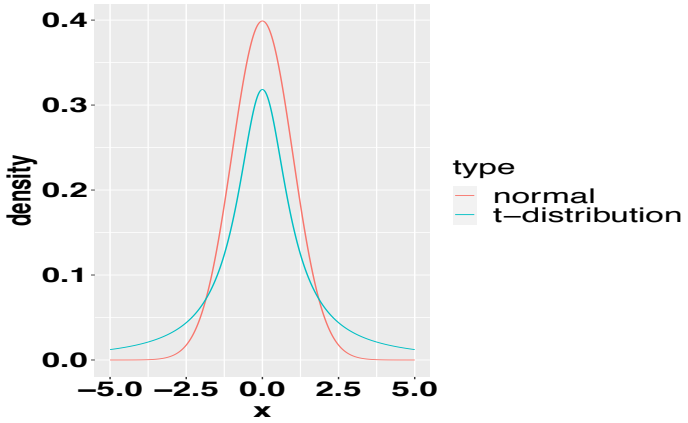
High dimension

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

Low dimension

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

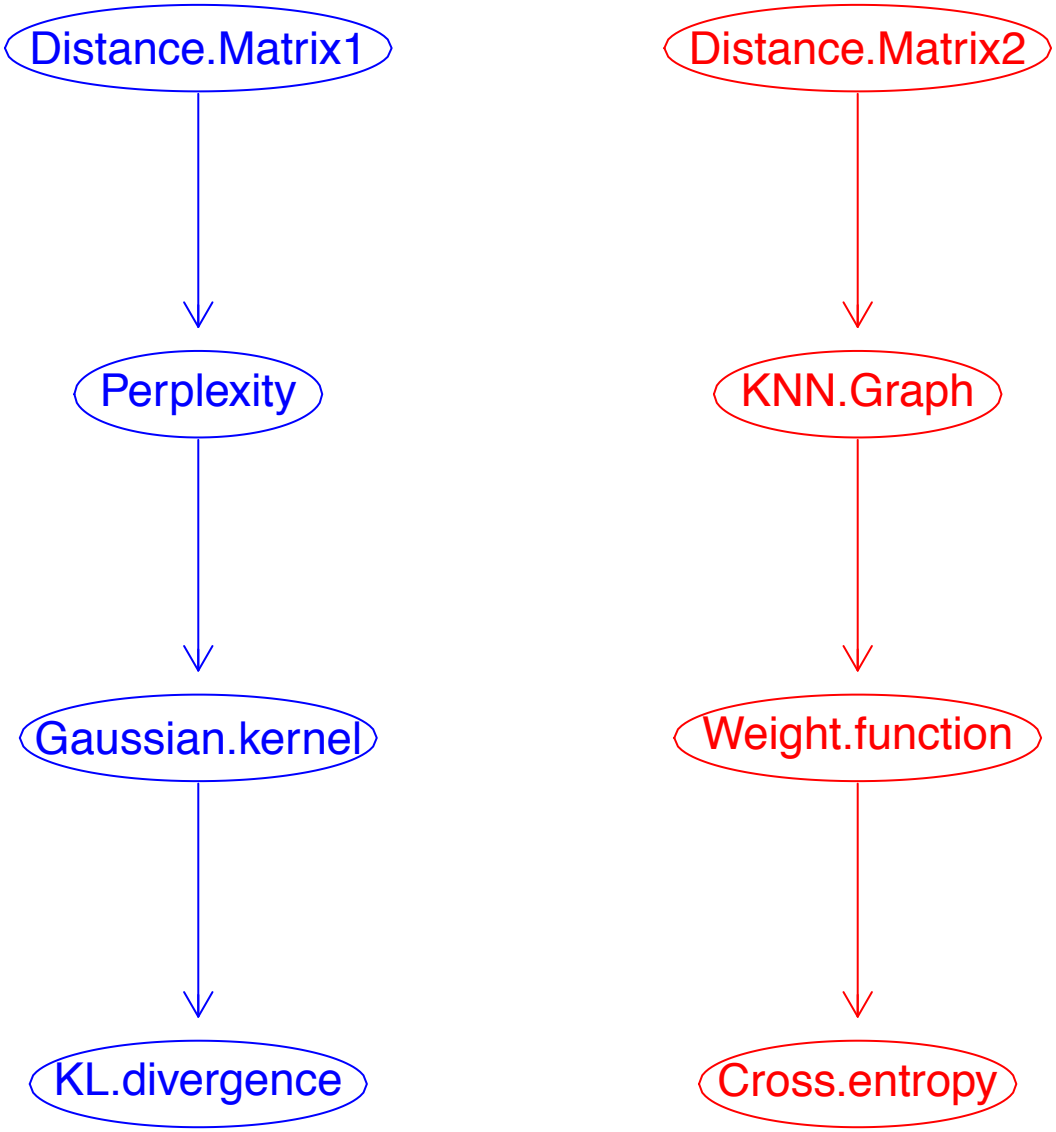


KL.Divergence

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \|y_i - y_j\|^2\right)^{-1}$$

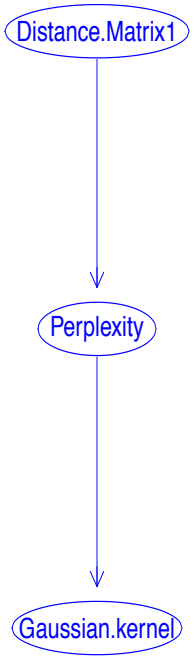
TSNE vs. UMAP


Distance.Matrix1
Perplexity
Gaussian.kernel
KL.divergence
Distance.Matrix2
KNN.Graph
Weight.function
Cross.entropy

TSNE vs. UMAP

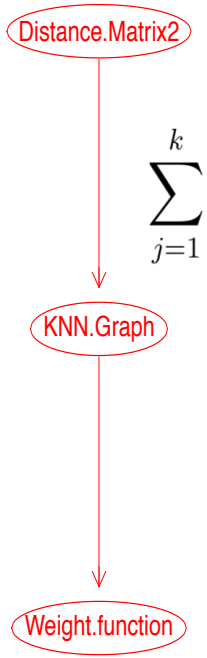
$$Perp(P_i) = 2^{H(P_i)}$$

$$H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$$



$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$



$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right) = \log_2(k)$$

$$w((x_i, x_{i_j})) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right)$$

$$B = A + A^\top - A \circ A^\top$$

ρ_i: shortest distance of x_i neighbors

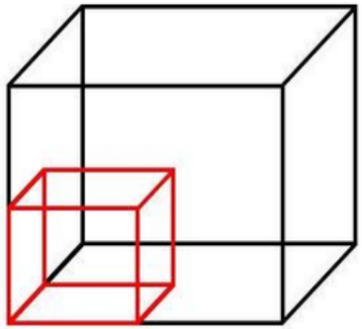
Euclidean Distance and Other Distance Metrics

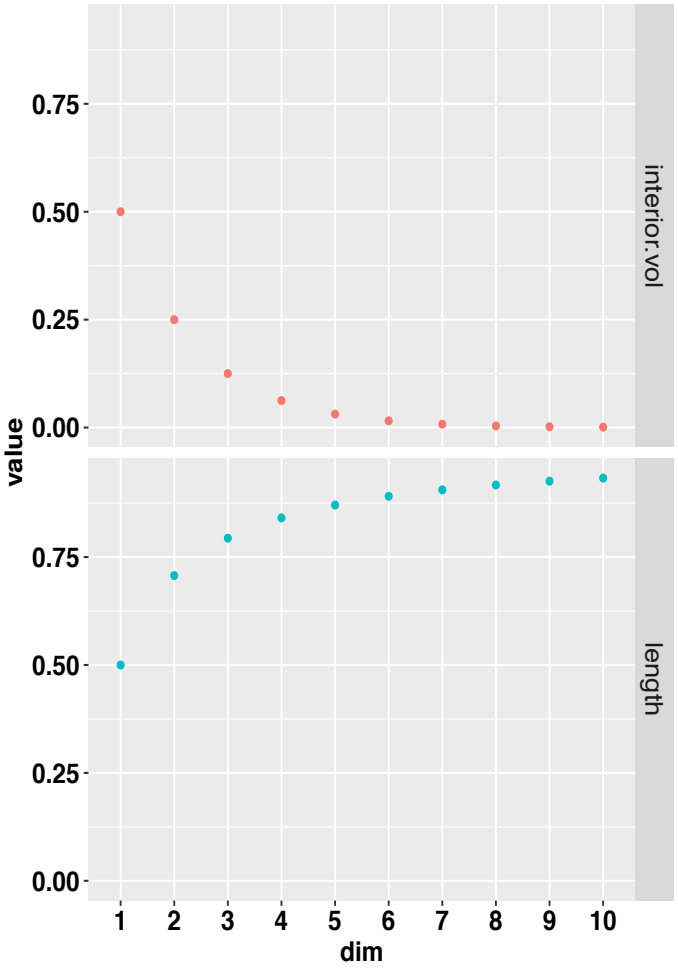
- Euclidean distance vs geodesic distance
- Euclidean distance vs Mahalanobis distance
- Curse of dimensionality

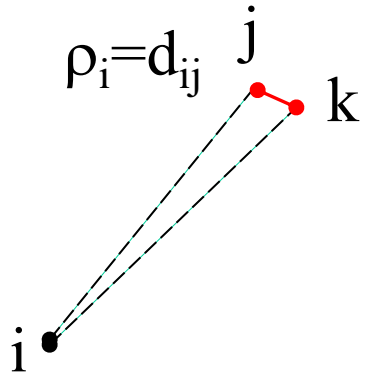
Curse of Dimensionality

(I) 50% of each dimension is sufficient to cover 25% of a 2-dimensional space

(II) 50% of each dimension is only sufficient to cover 12.5% of a 3-dimensional space





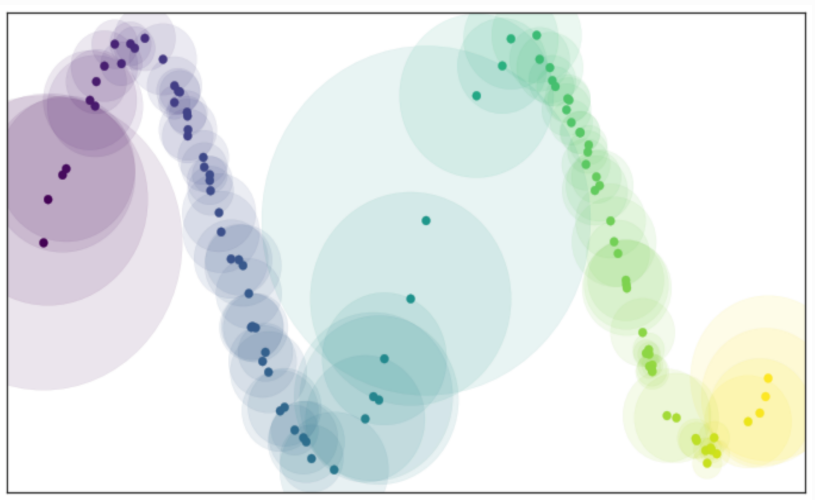


type
interior.vol
length

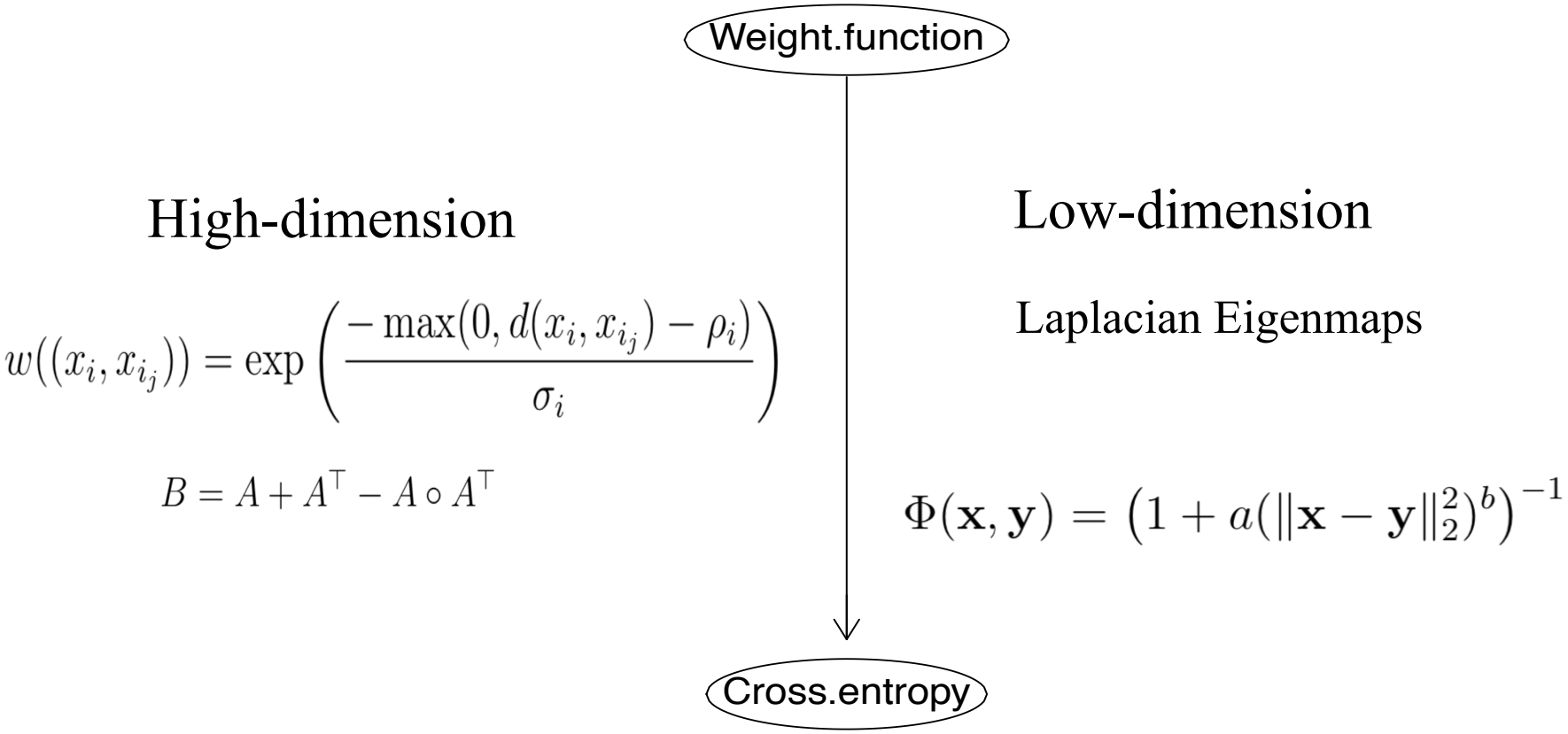
Uniform Manifold Approximation and Projection (UMAP)

$$w((x_i, x_{i_j})) = \exp\left(\frac{-\max(0, d(x_i, x_{i_j}) - \rho_i)}{\sigma_i}\right)$$

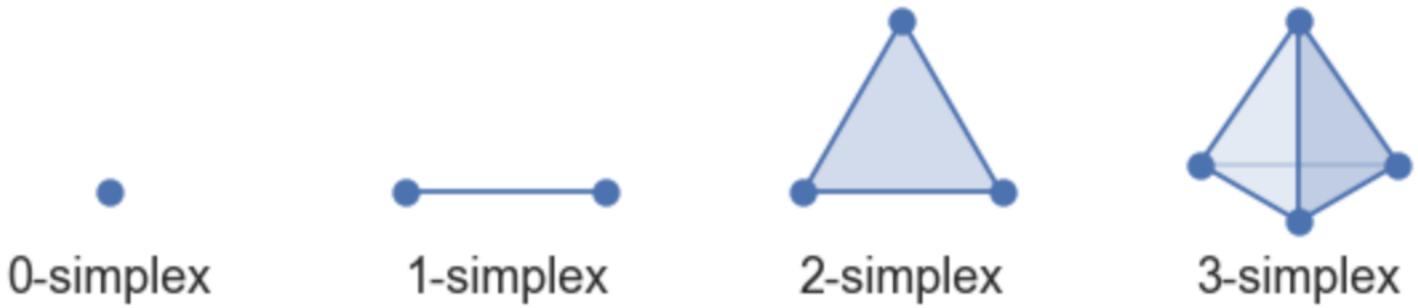
$$B = A + A^\top - A \circ A^\top$$

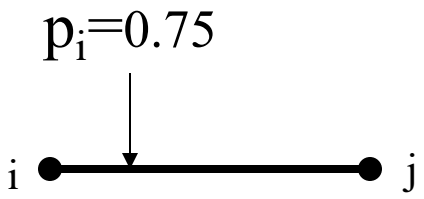


Uniform Manifold Approximation and Projection (UMAP)

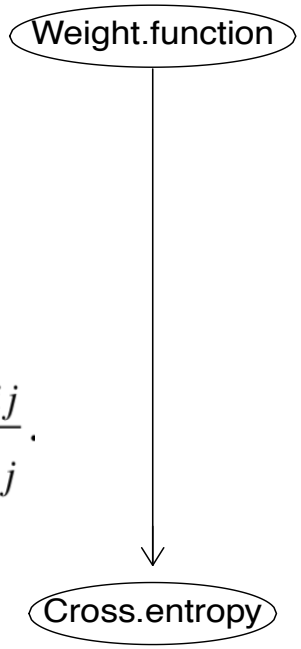


Fuzzy Simplicial Sets





Uniform Manifold Approximation and Projection (UMAP)



TSNE cost function

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

UMAP cost function

$$C((A,\mu),(A,\nu)) = \sum_{a \in A} \mu(a) \log \left(\frac{\mu(a)}{\nu(a)} \right) + (1-\mu(a)) \log \left(\frac{1-\mu(a)}{1-\nu(a)} \right)$$

Road Map for Dimension Reduction Methods