Dimensionality Reduction: Linear Discriminant Analysis and Principal Component Analysis

CMSC 678

UMBC

March 5th, 2018

Outline

Linear Algebra/Math Review

Two Methods of Dimensionality Reduction

- Linear Discriminant Analysis (LDA, LDiscA)
- Principal Component Analysis (PCA)

Covariance

covariance: how (linearly) correlated are variables

$$\sigma_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ki} - \mu_i)(x_{kj} - \mu_j)$$

Mean of variable *i*

Mean of variable *j*

covariance of variables *i* and *j*

Value of variable *i* in object k

Value of variable *j* in object k

Covariance

covariance: how (linearly) correlated are variables

$$\sigma_{ij} = \frac{1}{N-1}\sum_{k=1}^{N}(x_{ki} - \mu_i)(x_{kj} - \mu_j)$$

covariance of variables *i* and *j*

Value of variable *i* in object k

Mean of variable *i*

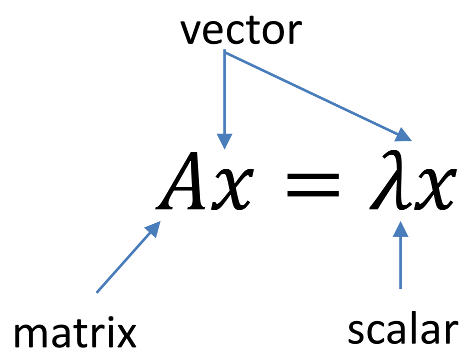
Value of variable *j* in object k

Mean of variable *j*

$$\sigma_{ij} = \sigma_{ji}$$

$$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1K} \\ \vdots & \ddots & \vdots \\ \sigma_{K1} & \cdots & \sigma_{KK} \end{pmatrix}$$

Eigenvalues and Eigenvectors



$$Ax = \lambda x$$

for a given matrix operation (multiplication):

what non-zero vector(s) change linearly?
(by a single multiplication)

Eigenvalues and Eigenvectors

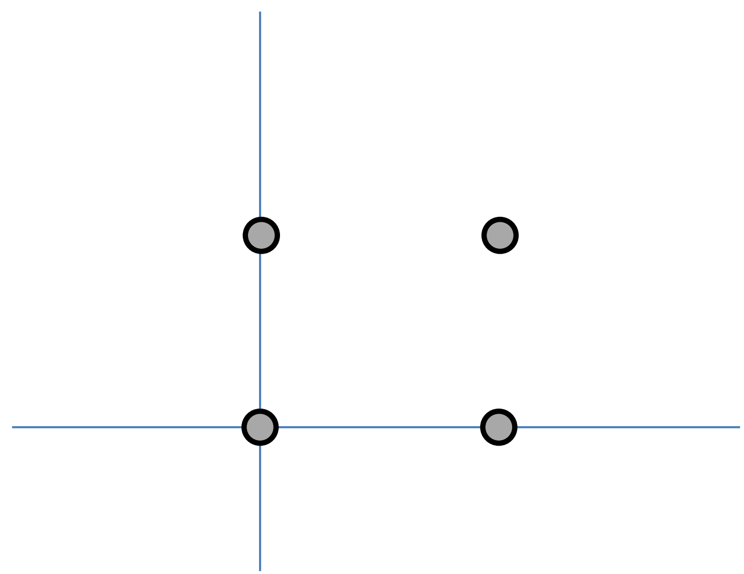
vector

$$Ax = \lambda x$$

matrix

scalar

$$A = \begin{pmatrix} 1 & 5 \\ 0 & 1 \end{pmatrix}$$



Eigenvalues and Eigenvectors

vector

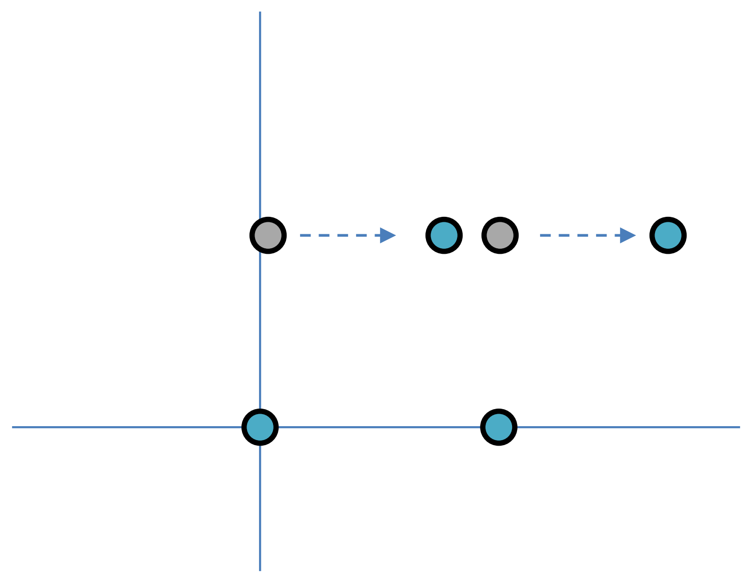
$$Ax = \lambda x$$

matrix   scalar

$$\begin{pmatrix} x + 5y \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}$$

$$A = \begin{pmatrix} 1 & 5 \\ 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 5 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x + 5y \\ y \end{pmatrix}$$



Eigenvalues and Eigenvectors

vector

$$Ax = \lambda x$$

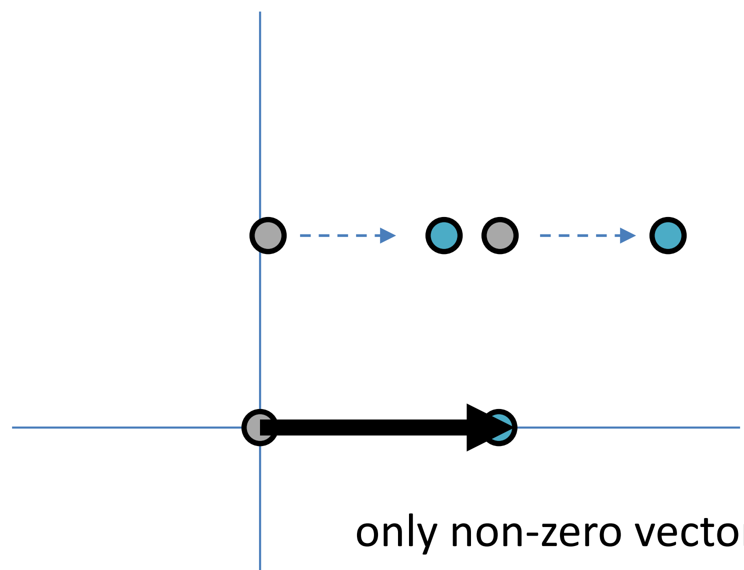
matrix

scalar

$$\begin{pmatrix} x + 5y \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{pmatrix} 1 & 5 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1 \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$A = \begin{pmatrix} 1 & 5 \\ 0 & 1 \end{pmatrix}$$



only non-zero vector to scale

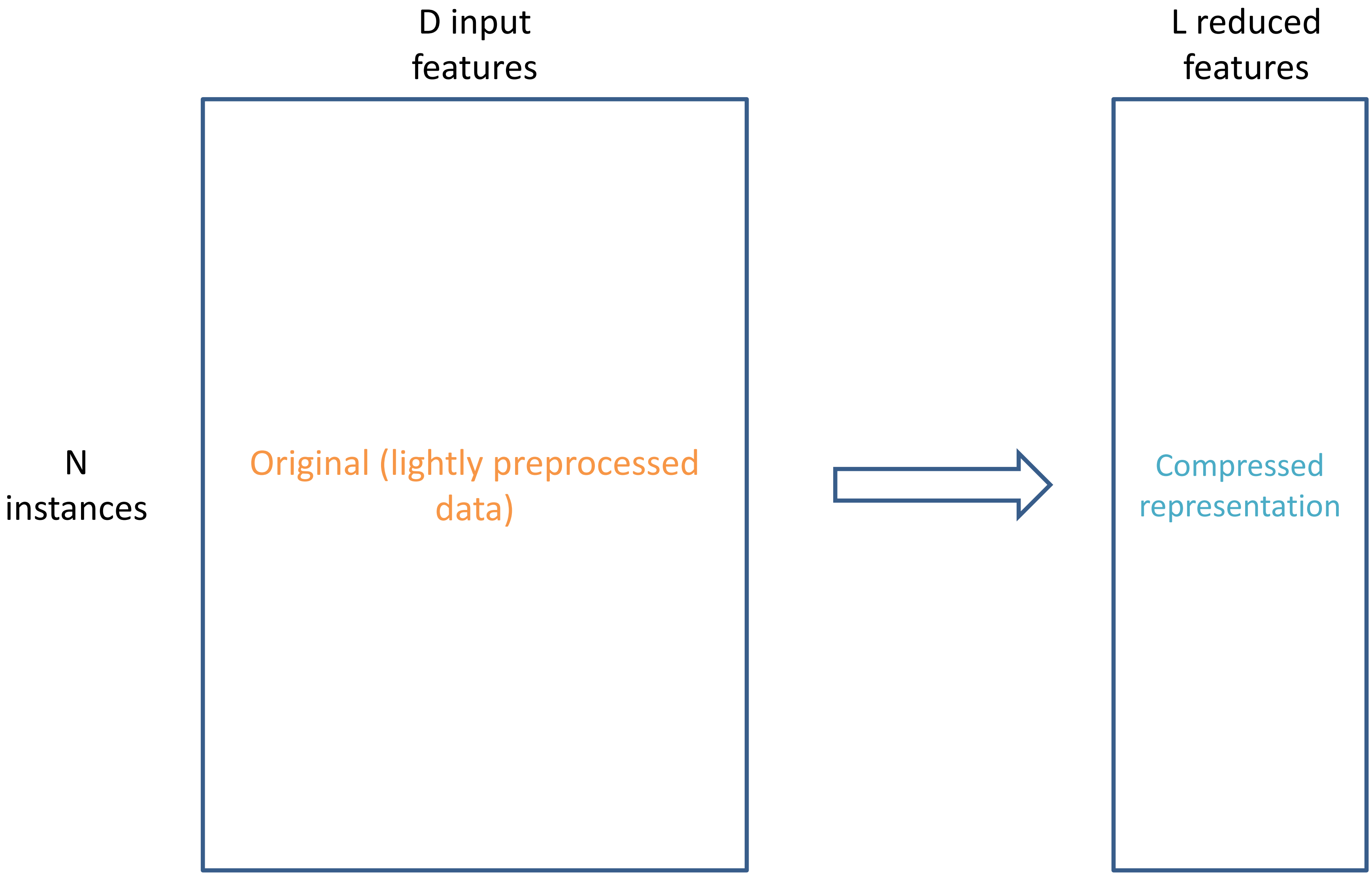
Outline

Linear Algebra/Math Review

Two Methods of Dimensionality Reduction

- Linear Discriminant Analysis (LDA, LDiscA)
- Principal Component Analysis (PCA)

Dimensionality Reduction



Dimensionality Reduction

clarity of representation vs. ease of understanding

oversimplification: loss of important or relevant information

Courtesy Antano Žilinsko

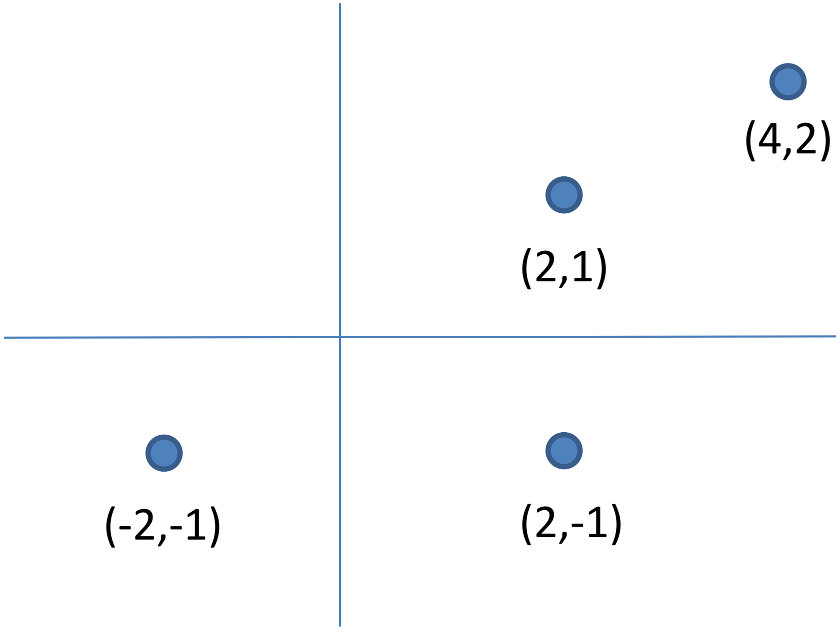
Why “maximize” the variance?

How can we efficiently summarize? We maximize the variance within our summarization

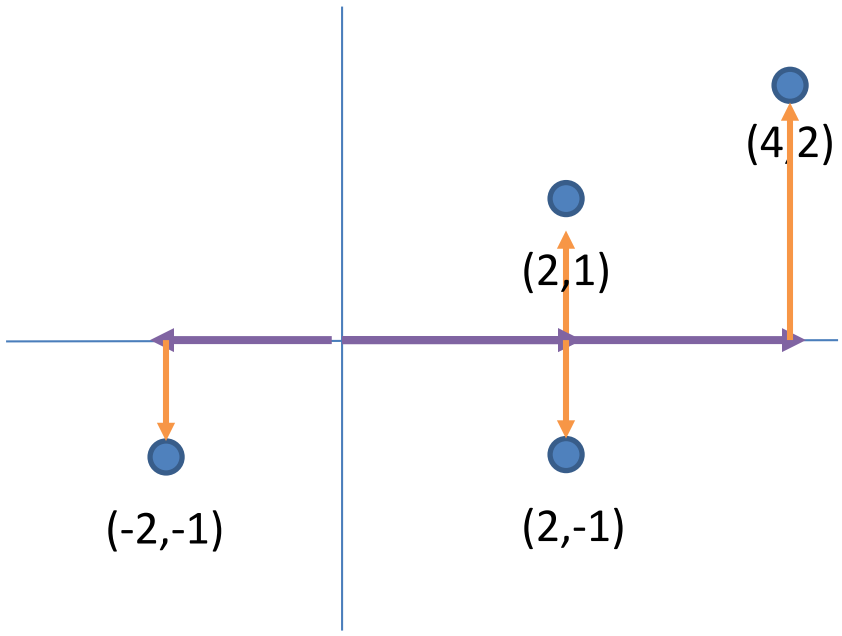
We don’t increase the variance in the dataset

How can we capture the most information with the fewest number of axes?

Summarizing Redundant Information

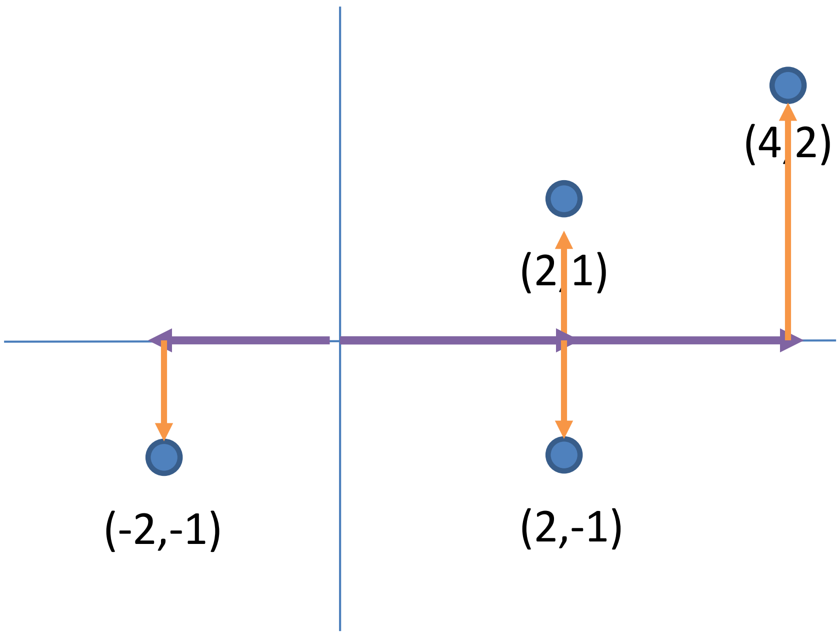


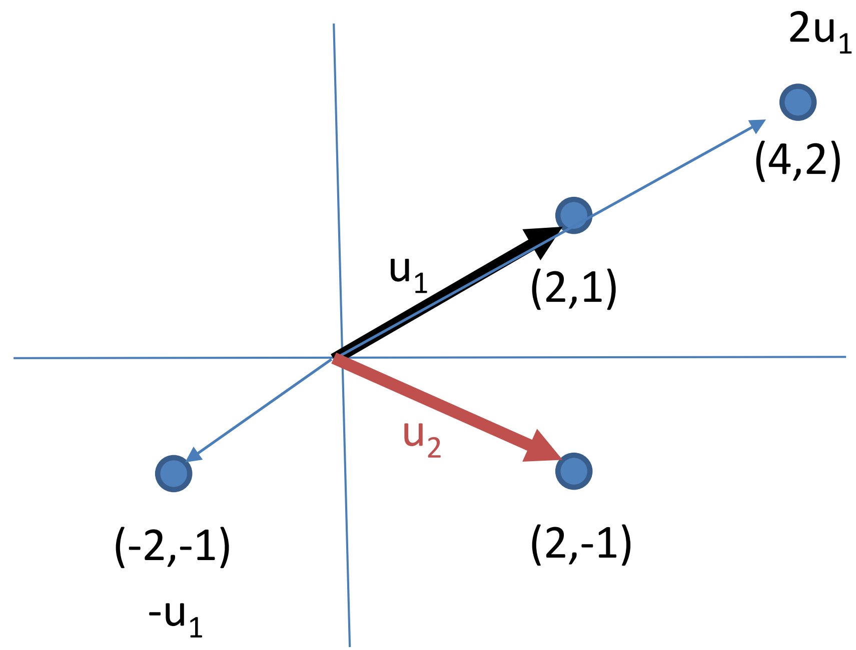
Summarizing Redundant Information



(2,1) = 2*(1,0) + 1*(0,1)

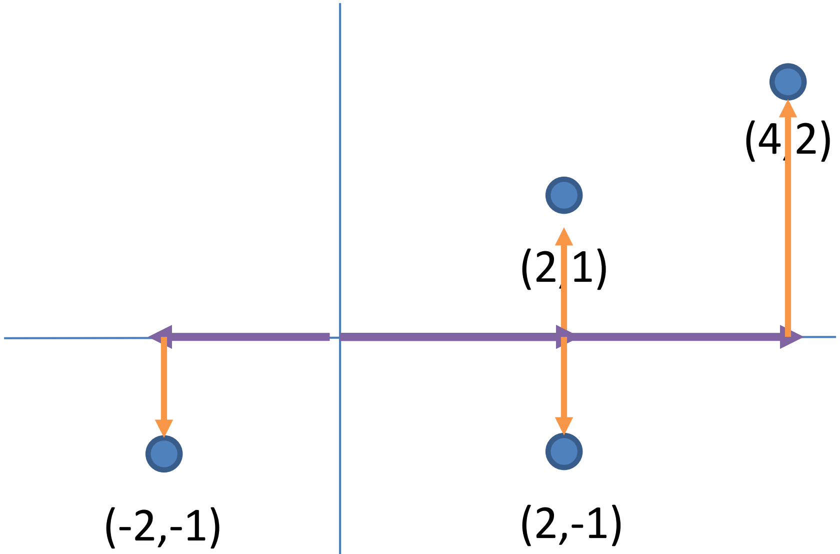
Summarizing Redundant Information

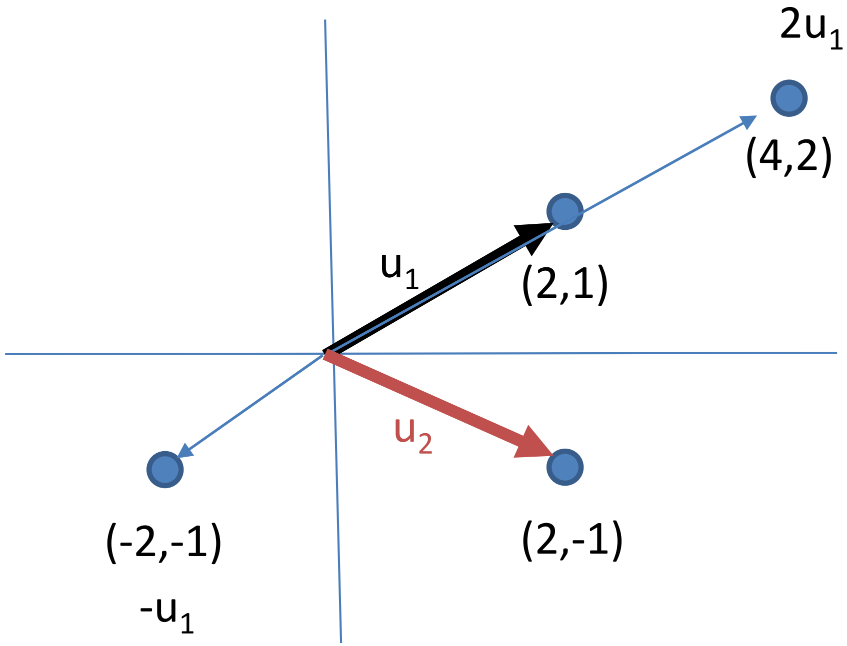




(2,1) = 1*(2,1) + 0*(2,-1)
(4,2) = 2*(2,1) + 0*(2,-1)

Summarizing Redundant Information





(2,1) = 1*(2,1) + 0*(2,-1)
(4,2) = 2*(2,1) + 0*(2,-1)

(Is it the most general? These vectors aren't orthogonal)

Outline

Linear Algebra/Math Review

Two Methods of Dimensionality Reduction

- Linear Discriminant Analysis (LDA, LDiscA)
- Principal Component Analysis (PCA)

Linear Discriminant Analysis (LDA, LDiscA) and Principal Component Analysis (PCA)

Summarize D-dimensional input data by uncorrelated axes

Uncorrelated axes are also called principal components

Use the first L components to account for as much variance as possible

Geometric Rationale of LDiscA & PCA

Objective: to rigidly rotate the axes of the D-dimensional space to new positions (principal axes):

ordered such that principal axis 1 has the highest variance, axis 2 has the next highest variance, ... , and axis D has the lowest variance

covariance among each pair of the principal axes is zero (the principal axes are uncorrelated)

Courtesy Antano Žilinsko

Remember: MAP Classifiers are Optimal for Classification

$$\min_{\mathbf{w}} \sum_i \mathbb{E}_{\widehat{y_i}}[\ell^{0/1}(y, \widehat{y_i})] \rightarrow \max_{\mathbf{w}} \sum_i p(\hat{y}_i = y_i | x_i)$$

$$p(\hat{y}_i = y_i | x_i) \propto p(x_i | \hat{y}_i) p(\hat{y}_i)$$

posterior *class-conditional likelihood* *class prior*

$x_i \in \mathbb{R}^D$

Linear Discriminant Analysis

MAP Classifier where:

1. class-conditional likelihoods are Gaussian

2. common covariance among class likelihoods

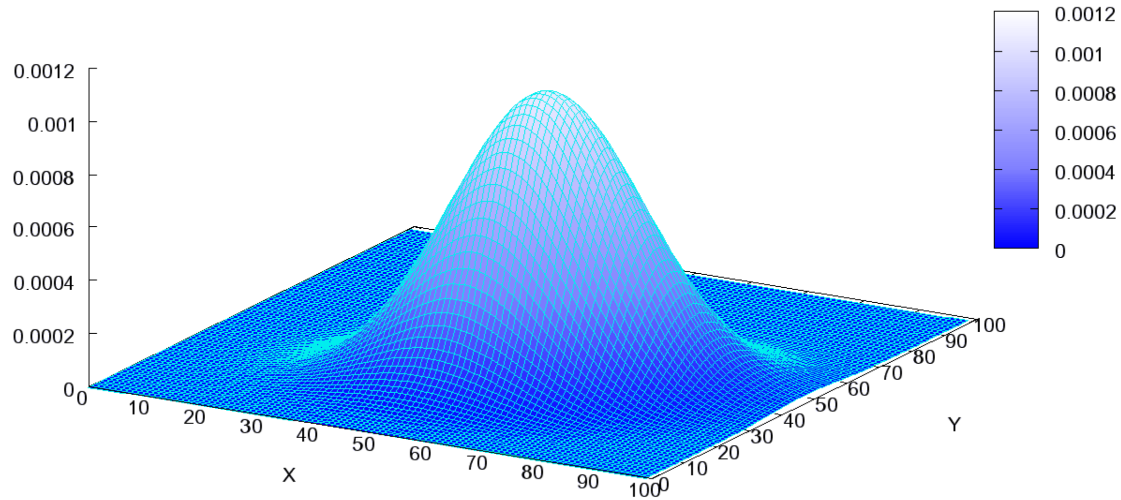
LDiscA: (1) What if likelihoods are Gaussian

$$p(\hat{y}_i = y_i | x_i) \propto p(x_i | \hat{y}_i) p(\hat{y}_i)$$

$$p(x_i | k) = \mathcal{N}(\mu_k, \Sigma_k)$$

$$= \frac{\exp\left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)\right)}{(2\pi)^{D/2} |\Sigma_k|^{1/2}}$$

Multivariate Normal Distribution



https://upload.wikimedia.org/wikipedia/commons/5/57/Multivariate_Gaussian.png

LDiscA: (2) Shared Covariance

$$\log \frac{p(\hat{y}_i = k | x_i)}{p(\hat{y}_i = l | x_i)} = \log \frac{p(x_i | k)}{p(x_i | l)} + \log \frac{p(k)}{p(l)}$$

LDiscA: (2) Shared Covariance

$$\log \frac{p(\hat{y_i} = k | x_i)}{p(\hat{y_i} = l | x_i)} = \log \frac{p(x_i | k)}{p(x_i | l)} + \log \frac{p(k)}{p(l)}$$

$$= \log \frac{p(k)}{p(l)} + \log \left[\frac{\frac{\exp\left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)\right)}{(2\pi)^{D/2} |\Sigma_k|^{1/2}}}{\frac{\exp\left(-\frac{1}{2}(x_i - \mu_l)^T \Sigma_l^{-1} (x_i - \mu_l)\right)}{(2\pi)^{D/2} |\Sigma_l|^{1/2}}} \right]$$

LDiscA: (2) Shared Covariance

$$\log \frac{p(\hat{y}_i = k | x_i)}{p(\hat{y}_i = l | x_i)} = \log \frac{p(x_i | k)}{p(x_i | l)} + \log \frac{p(k)}{p(l)}$$

$$= \log \frac{p(k)}{p(l)} + \log \left[\frac{\frac{\exp\left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma^{-1} (x_i - \mu_k)\right)}{\cancel{(2\pi)^{D/2} |\Sigma_k|^{1/2}}}}{\frac{\exp\left(-\frac{1}{2}(x_i - \mu_l)^T \Sigma^{-1} (x_i - \mu_l)\right)}{\cancel{(2\pi)^{D/2} |\Sigma_l|^{1/2}}}} \right]$$

$$\Sigma_l = \Sigma_k$$

LDiscA: (2) Shared Covariance

$$\log \frac{p(\hat{y}_i = k | x_i)}{p(\hat{y}_i = l | x_i)} = \log \frac{p(x_i | k)}{p(x_i | l)} + \log \frac{p(k)}{p(l)}$$

$$= \log \frac{p(k)}{p(l)} - \frac{1}{2}(\mu_k - \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x_i^T \Sigma^{-1} (\mu_k - \mu_l)$$

linear in x_i

(check for yourself: why did the quadratic x_i terms cancel?)

LDiscA: (2) Shared Covariance

$$\log \frac{p(\hat{y}_i = k | x_i)}{p(\hat{y}_i = l | x_i)} = \log \frac{p(x_i | k)}{p(x_i | l)} + \log \frac{p(k)}{p(l)}$$

$$= \log \frac{p(k)}{p(l)} - \frac{1}{2}(\mu_k - \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x_i^T \Sigma^{-1} (\mu_k - \mu_l)$$

$$= x_i^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log p(k)$$

$$+ x_i^T \Sigma^{-1} \mu_l - \frac{1}{2} \mu_l^T \Sigma^{-1} \mu_l + \log p(l)$$

linear in x_i
(check for yourself: why did the quadratic x_i terms cancel?)

rewrite only in terms of x_i (data) and single-class terms

Classify via Linear Discriminant Functions

$$\delta_k(x_i) = x_i^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log p(k)$$

$\arg\max_k \; \delta_k(x_i)$ $\xrightarrow[\text{to}]{\text{equivalent}}$ MAP classifier

LDiscA

Parameters to learn: $\{p(k)\}_k, \{\mu_k\}_k, \Sigma$

$$p(k) \propto N_k$$

number of items labeled with class k

LDiscA

Parameters to learn: $\{p(k)\}_k, \{\mu_k\}_k, \Sigma$

$$p(k) \propto N_k$$

$$\mu_k = \frac{1}{N_k} \sum_{i:y_i=k} x_i$$

LDiscA

Parameters to learn: $\{p(k)\}_k, \{\mu_k\}_k, \Sigma$

$$p(k) \propto N_k$$

$$\mu_k = \frac{1}{N_k} \sum_{i:y_i=k} x_i$$

$$\Sigma = \frac{1}{N-K} \sum_k \text{scatter}_k = \frac{1}{N-K} \sum_k \left[\sum_{i:y_i=k} (x_i - \mu_k)(x_i - \mu_k)^T \right]$$

one option for Σ

within-class covariance

Computational Steps for Full-Dimensional LDiscA

1. Compute means, priors, and covariance

Computational Steps for Full-Dimensional LDiscA

1. Compute means, priors, and covariance
2. Diagonalize covariance

$$\Sigma = \mathrm{UDU}^{\mathrm{T}}$$

Eigen decomposition

K x K orthonormal matrix (eigenvectors)

diagonal matrix of eigenvalues

Computational Steps for Full-Dimensional LDiscA

1. Compute means, priors, and covariance
2. Diagonalize covariance

$$\Sigma = \mathrm{UDU}^{\mathrm{T}}$$

3. Sphere the data

$$\mathrm{X}^{*} = D^{\frac{-1}{2}} U^{T} X$$

Computational Steps for Full-Dimensional LDiscA

1. Compute means, priors, and covariance
2. Diagonalize covariance

$$\Sigma = \mathrm{UDU}^{\mathrm{T}}$$

3. Sphere the data (get unit covariance)

$$\mathrm{X}^* = D^{\frac{-1}{2}} U^T X$$

4. Classify according to linear discriminant functions $\delta_k(x_i^*)$

Two Extensions to LDiscA

Quadratic Discriminant Analysis (QDA)

Keep separate covariances per class

$$\delta_k(x_i) = -\frac{1}{2}(x_i - \mu_k)^{\mathrm{T}} \Sigma_{\mathrm{k}}^{-1} (x_i - \mu_k) + \log p(k) - \frac{\log|\Sigma_k|}{2}$$

Two Extensions to LDiscA

Quadratic Discriminant Analysis (QDA)

Keep separate covariances per class

$$\delta_k(x_i) = -\frac{1}{2}(x_i - \mu_k)^{\mathrm{T}}\Sigma_{\mathrm{k}}^{-1}(x_i - \mu_k) + \log p(k) - \frac{\log|\Sigma_k|}{2}$$

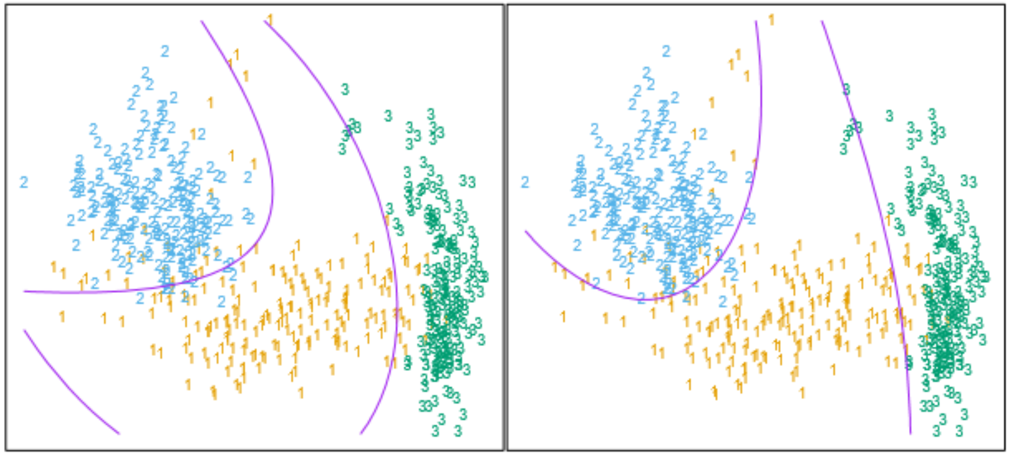
Regularized LDiscA

Interpolate between shared covariance estimate (LDiscA) and class-specific estimate (QDA)

$$\Sigma_k(\alpha) = \alpha\Sigma_k + (1-\alpha)\Sigma$$

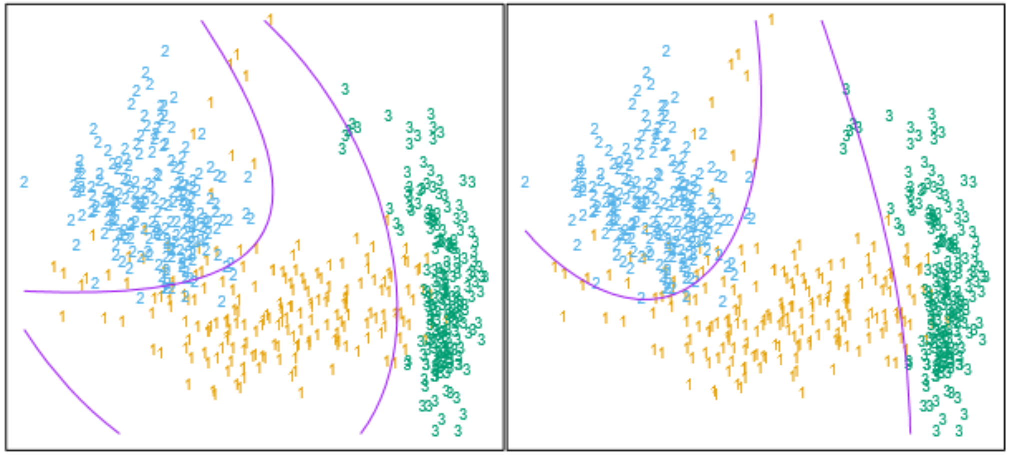
Vowel Classification

LDiscA (left) vs. QDA (right)



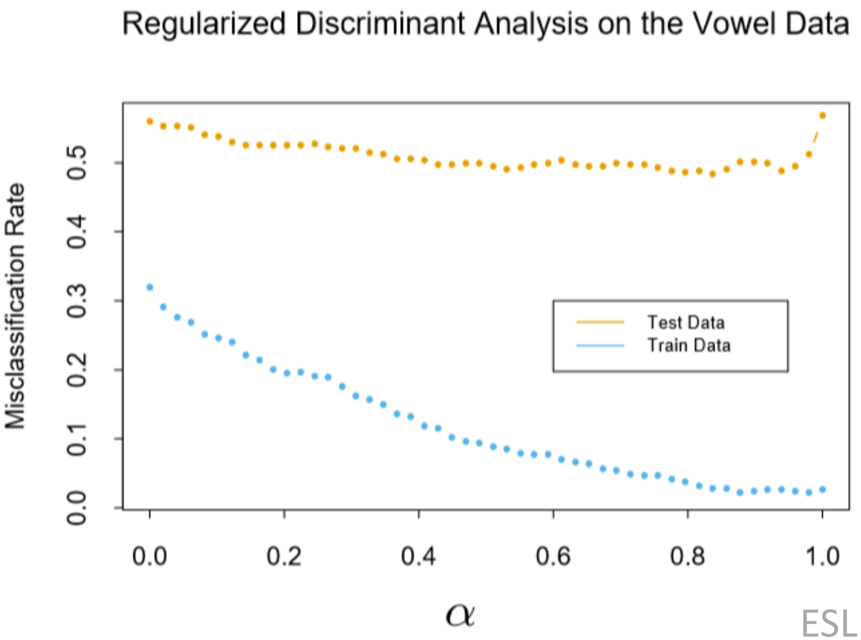
Vowel Classification

LDiscA (left) vs. QDA (right)



Regularized LDiscA

$$\Sigma_k(\alpha) = \alpha\Sigma_k + (1-\alpha)\Sigma$$



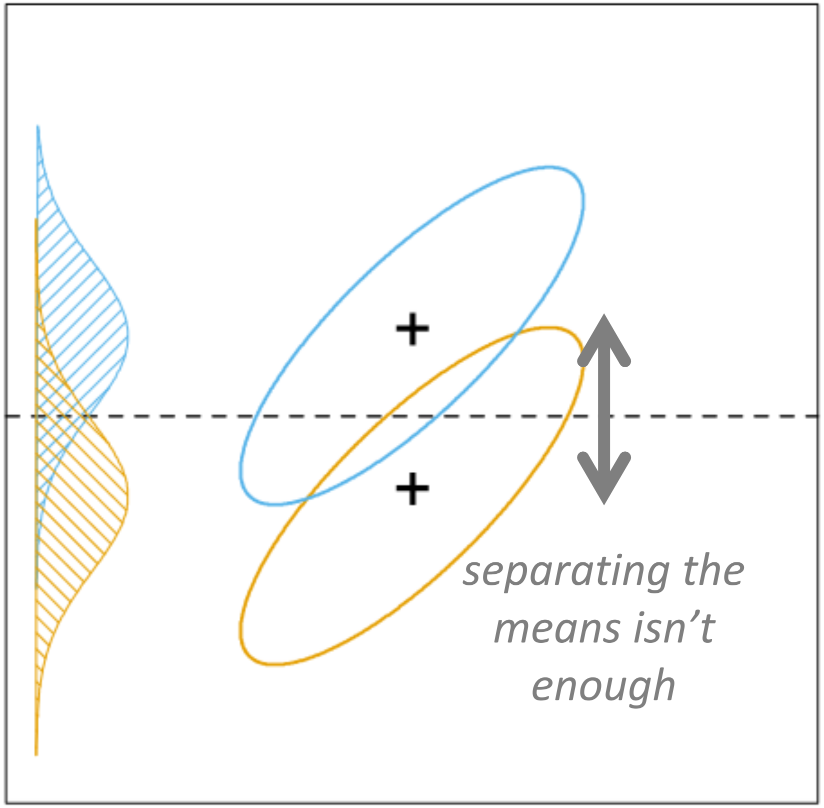
LDA for Dimensionality Reduction

Classifying D-dimensional inputs (features) into K-dimensional space (labels)

Can we view the data faithfully (optimally) in smaller dimensions?

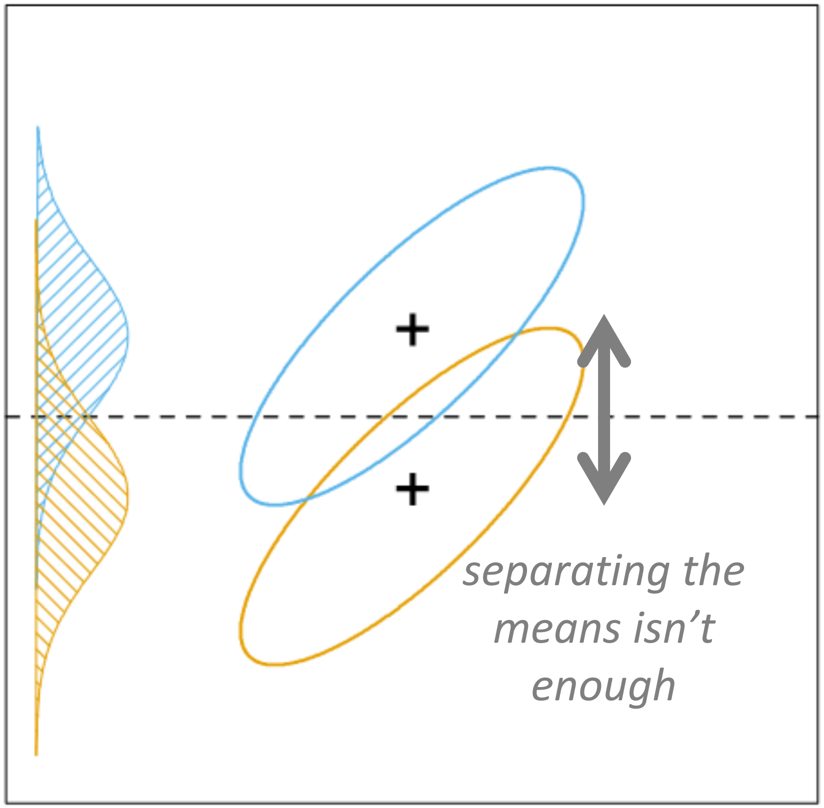
Fisher's optimal: spread out the centroids (means)

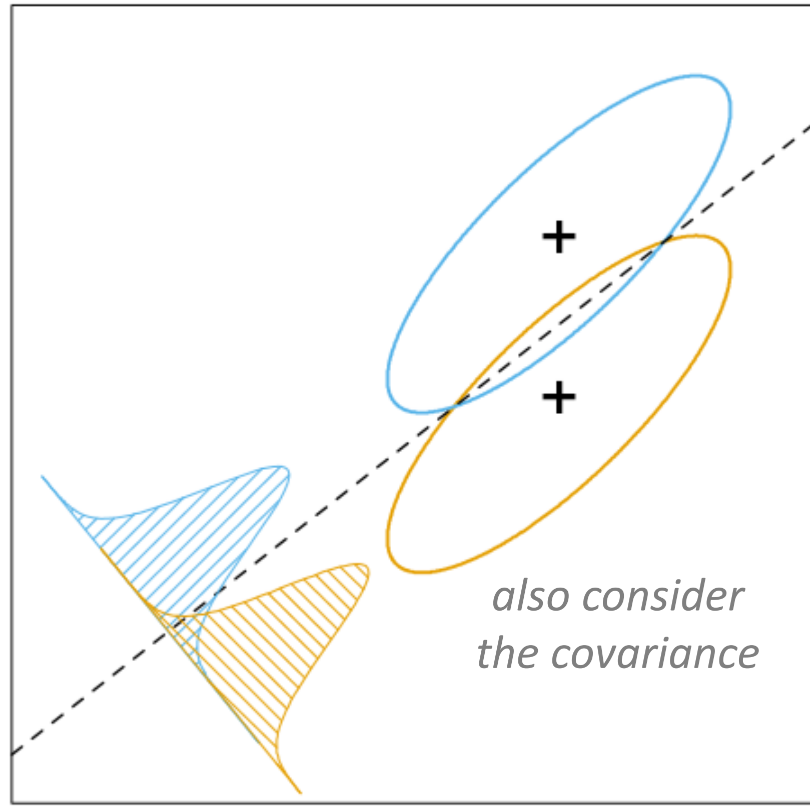
Fisher's Argument



"Find a linear combination such that the between-class variance is maximized relative to the within-class variance" (ESL, 4.3)

Fisher's Argument





"Find a linear combination such that the between-class variance is maximized relative to the within-class variance" (ESL, 4.3)

L-Dimensional LDiscA

"Find a linear combination such that the between-class variance is maximized relative to the within-class variance" (ESL, 4.3)

$$\max \frac{u^T B u}{u^T \Sigma u}$$

$$\mathrm{B} = \sum_k (\mu_k - \mu)(\mu_k - \mu)^T$$

between-class scatter (covariance)

L-Dimensional LDiscA

"Find a linear combination such that the between-class variance is maximized relative to the within-class variance" (ESL, 4.3)

$$\max \frac{u^T B u}{u^T \Sigma u} \longrightarrow \max u^T B u \ \text{ s.t. } u^T \Sigma u = 1$$

$$\mathrm{B} = \sum_k (\mu_k - \mu)(\mu_k - \mu)^T$$

between-class scatter (covariance)

L-Dimensional LDiscA

"Find a linear combination such that the between-class variance is maximized relative to the within-class variance" (ESL, 4.3)

generalized eigenvalue problem

$$\max \frac{u^T B u}{u^T \Sigma u} \longrightarrow \max u^T B u \;\; \text{s.t.}\; u^T \Sigma u = 1$$

first (largest) eigenvector

L-Dimensional LDiscA

"Find a linear combination such that the between-class variance is maximized relative to the within-class variance" (ESL, 4.3)

find the *next* largest eigenvector

$$\max u_2^T B u_2$$
$$\text{s.t. } u_2^T \Sigma u_2 = 1, u_1^T u_2 = 0$$

L-Dimensional LDiscA

"Find a linear combination such that the between-class variance is maximized relative to the within-class variance" (ESL, 4.3)

and the *next* largest eigenvector….

$$\max u_3^T B u_3$$
$$\text{s.t.}\, u_3^T \Sigma u_3 = 1,$$
$$u_1^T u_2 = 0,$$
$$u_1^T u_3 = 0,$$
$$u_2^T u_3 = 0$$

L-Dimensional LDiscA

1. Compute means μ, priors, and common covariance Σ

$$\Sigma = \frac{1}{N-K} \sum_k \text{scatter}_k = \frac{1}{N-K} \sum_k \left[\sum_{i:y_i=k} (x_i - \mu_k)(x_i - \mu_k)^T \right]$$

L-Dimensional LDiscA

1. Compute means μ, priors, and common covariance Σ

$$\Sigma = \frac{1}{N-K}\sum_{k} \text{scatter}_k = \frac{1}{N-K}\sum_{k}\left[\sum_{i:y_i=k}(x_i-\mu_k)(x_i-\mu_k)^T\right]$$

2. Compute the between-class scatter (covariance)

$$\mathrm{B} = \sum_{k}(\mu_k-\mu)(\mu_k-\mu)^T$$

L-Dimensional LDiscA

1. Compute means μ, priors, and common covariance Σ

$$\Sigma = \frac{1}{N-K}\sum_k \text{scatter}_k = \frac{1}{N-K}\sum_k \left[\sum_{i:y_i=k} (x_i - \mu_k)(x_i - \mu_k)^T\right]$$

2. Compute the between-class scatter (covariance)

$$\mathrm{B} = \sum_k (\mu_k - \mu)(\mu_k - \mu)^T$$

3. Compute the eigen decomposition of B

$$B = V D_B V^T$$

L-Dimensional LDiscA

1. Compute means μ, priors, and common covariance Σ

$$\Sigma = \frac{1}{N-K}\sum_k \text{scatter}_k = \frac{1}{N-K}\sum_k \left[\sum_{i:y_i=k} (x_i - \mu_k)(x_i - \mu_k)^T\right]$$

2. Compute the between-class scatter (covariance)

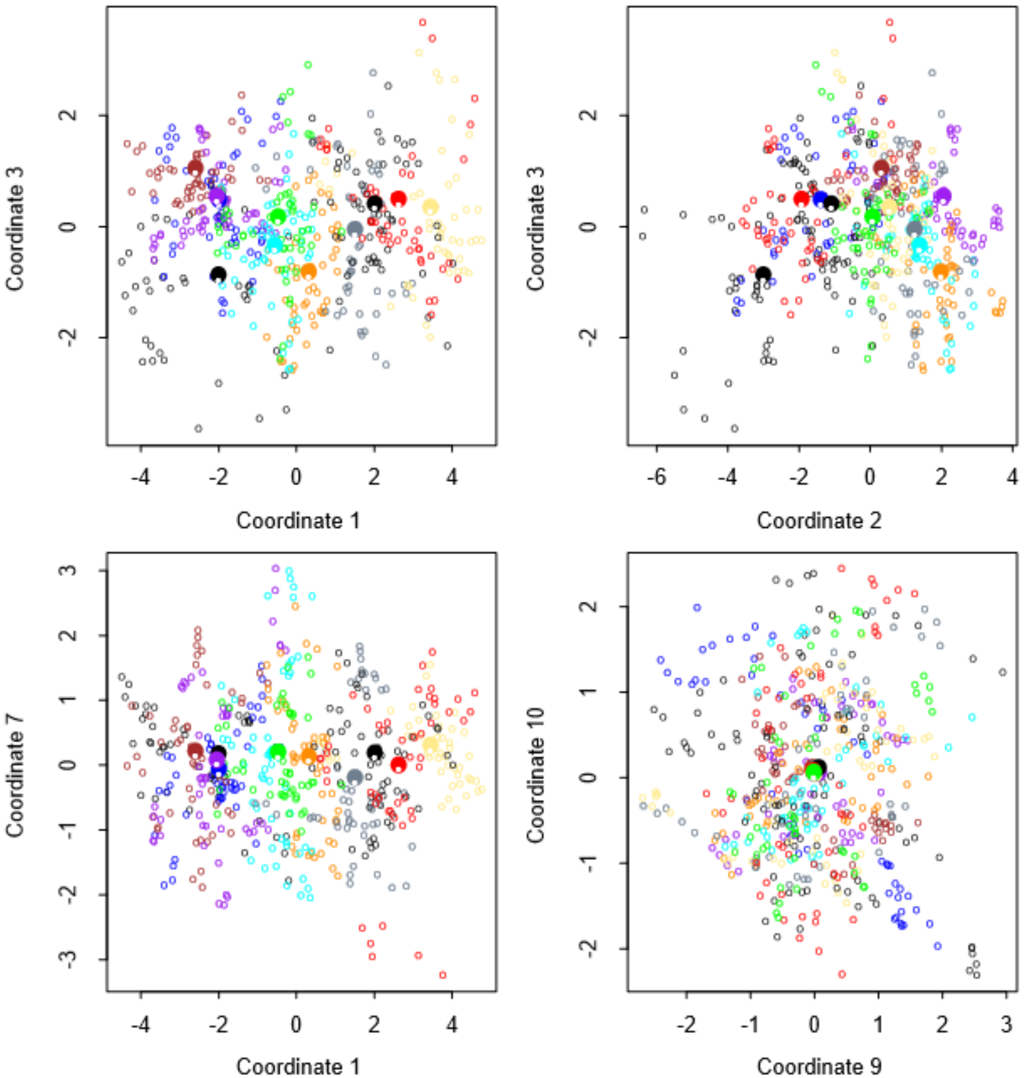
$$\mathrm{B} = \sum_k (\mu_k - \mu)(\mu_k - \mu)^T$$

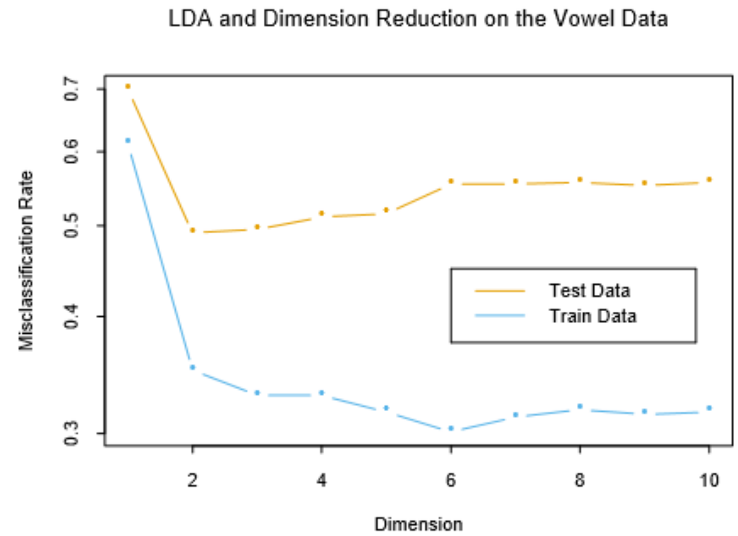
3. Compute the eigen decomposition of B

$$B = V D_B V^T$$

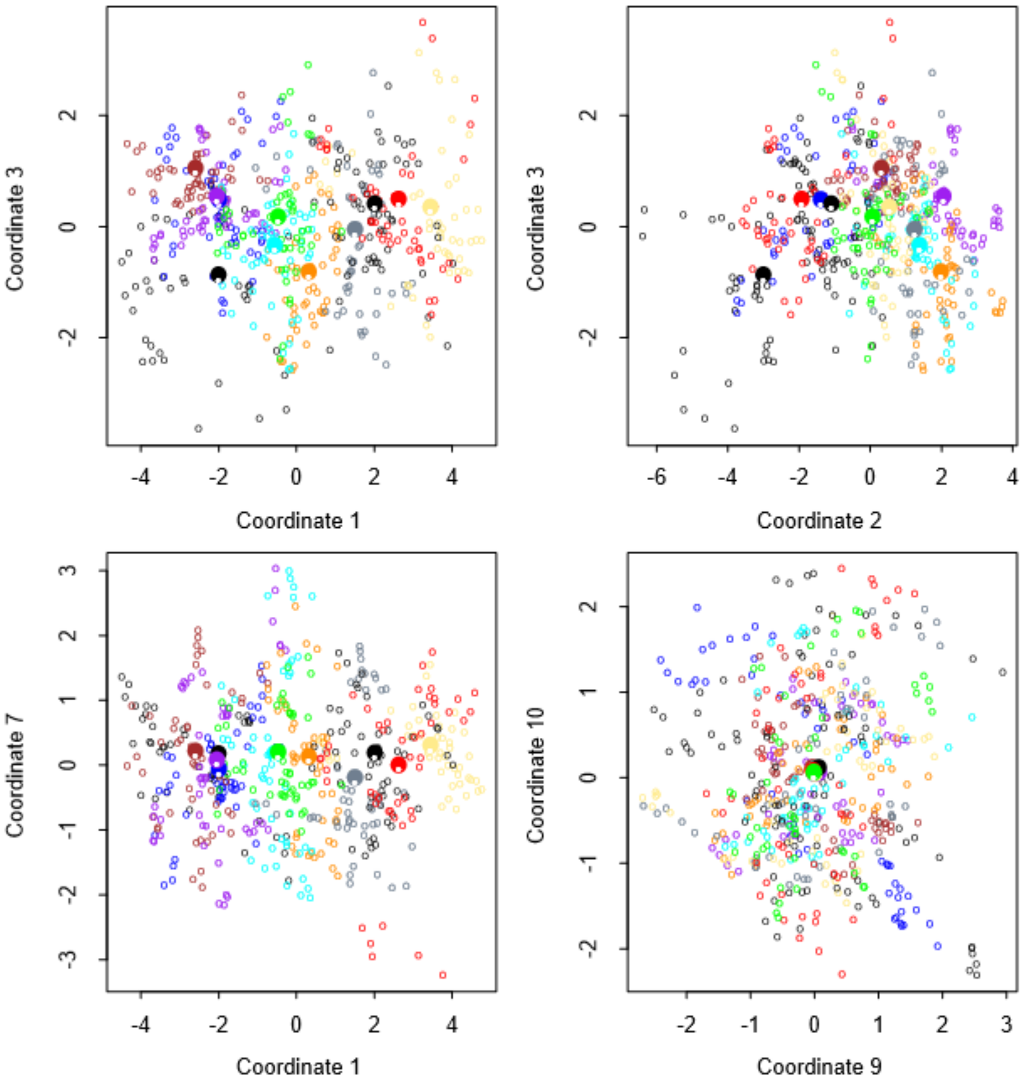
4. Take the top L eigenvectors from V

Vowel Classification





Vowel Classification



Supervised → Unsupervised

Supervised learning: learning with a teacher

You had training data which was (feature, label) pairs and the goal was to learn a mapping from features to labels

Supervised → Unsupervised

Supervised learning: learning with a teacher

- You had training data which was (feature, label) pairs and the goal was to learn a mapping from features to labels

Unsupervised learning: learning without a teacher

- Only features and no labels

Why is unsupervised learning useful?

- Visualization — dimensionality reduction
  - lower dimensional features might help learning
- Discover hidden structures in the data: clustering

Outline

Linear Algebra/Math Review

Two Methods of Dimensionality Reduction

- Linear Discriminant Analysis (LDA, LDiscA)
- Principal Component Analysis (PCA)

Geometric Rationale of LDiscA & PCA

Objective: to rigidly rotate the axes of the D-dimensional space to new positions (principal axes):

- ordered such that principal axis 1 has the highest variance, axis 2 has the next highest variance, ... , and axis D has the lowest variance

- covariance among each pair of the principal axes is zero (the principal axes are uncorrelated)

LDiscA: but now without the labels!

Adapted from Antano Žilinsko

L-Dimensional PCA

1. Compute mean μ, priors, and common covariance Σ

$$\Sigma = \frac{1}{N} \sum_{i:y_i=k} (x_i - \mu)(x_i - \mu)^T \qquad \mu = \frac{1}{N} \sum_i x_i$$

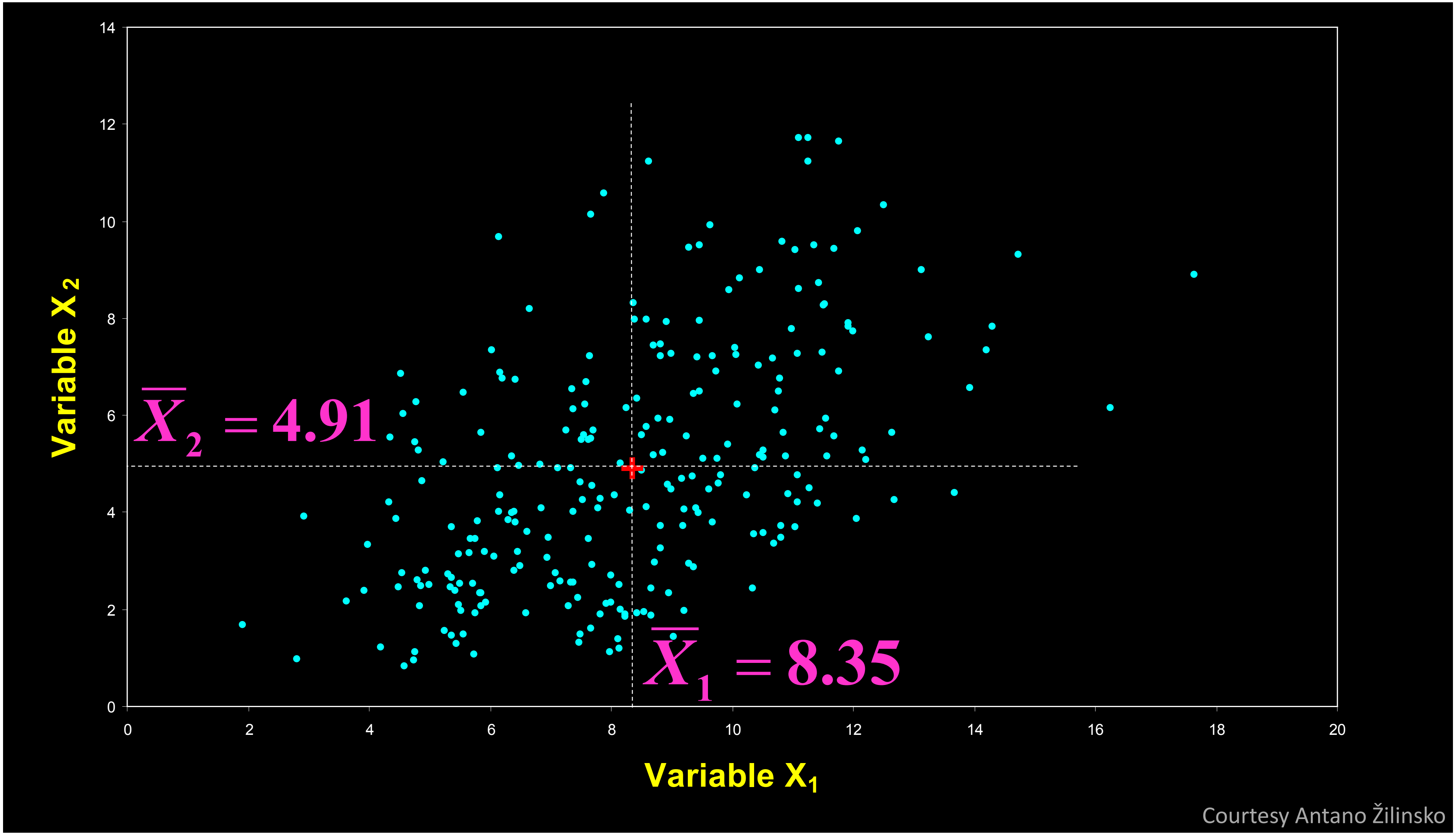
2. Sphere the data (zero-mean, unit covariance)
3. Compute the (top L) eigenvectors, from sphere-d data, via V

$$X^* = V D_B V^T$$

4. Project the data

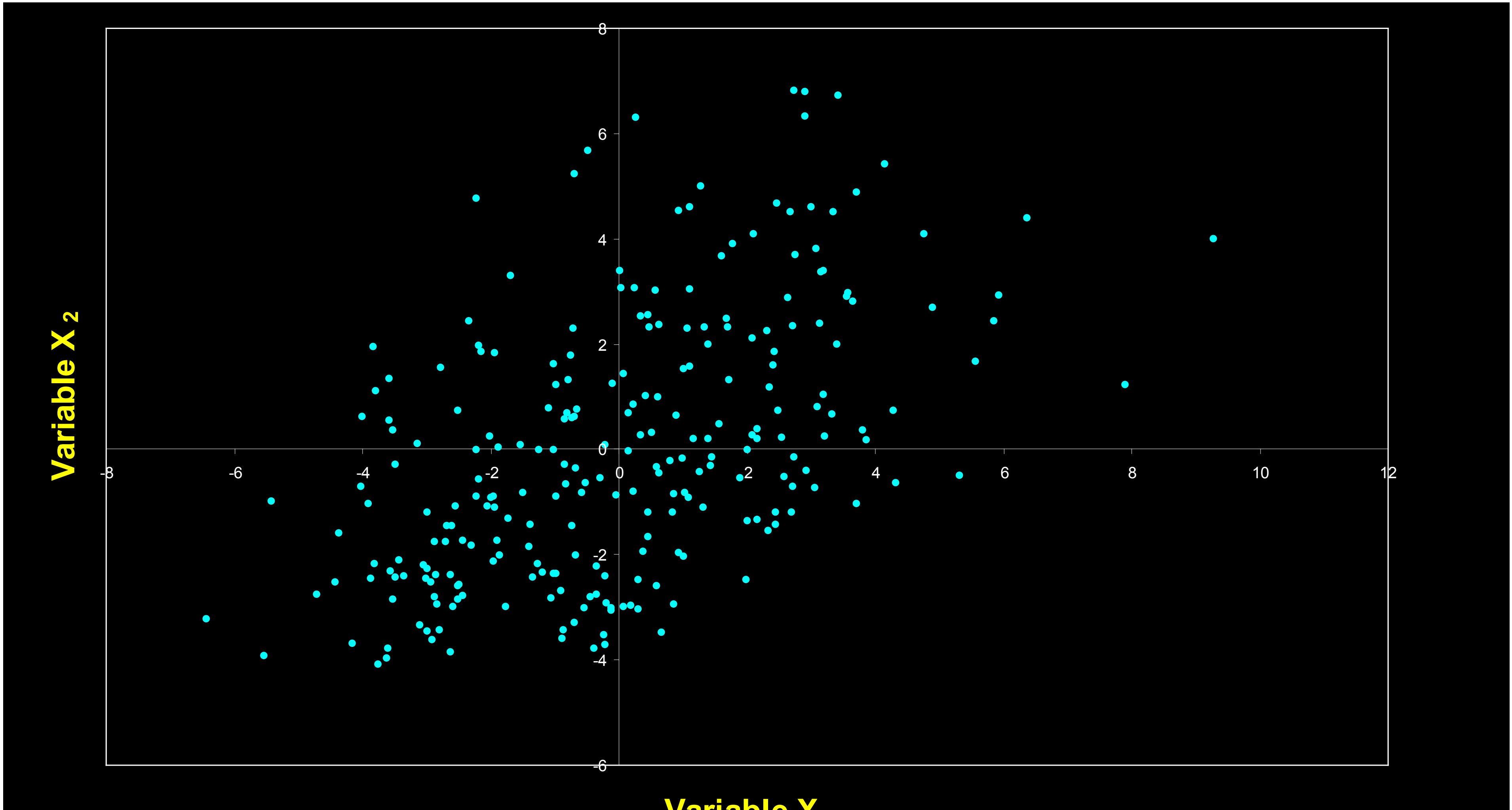
2D Example of PCA

variables X_1 and X_2 have positive covariance & each has a similar variance



Configuration is Centered

subtract the component-wise mean



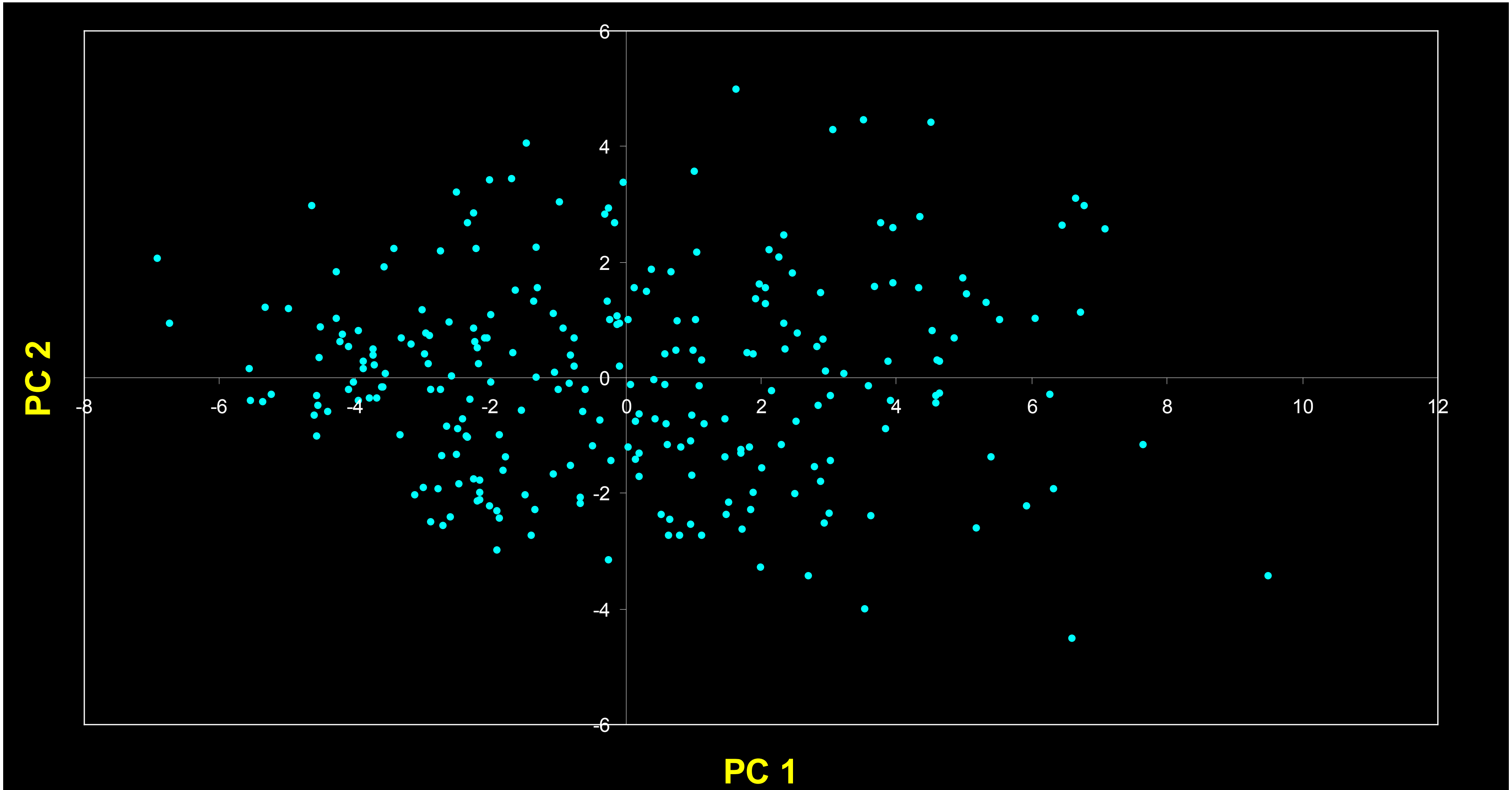
Courtesy Antano Žilinsko

Compute Principal Components

PC 1 has the highest possible variance (9.88)

PC 2 has a variance of 3.03

PC 1 and PC 2 have zero covariance.



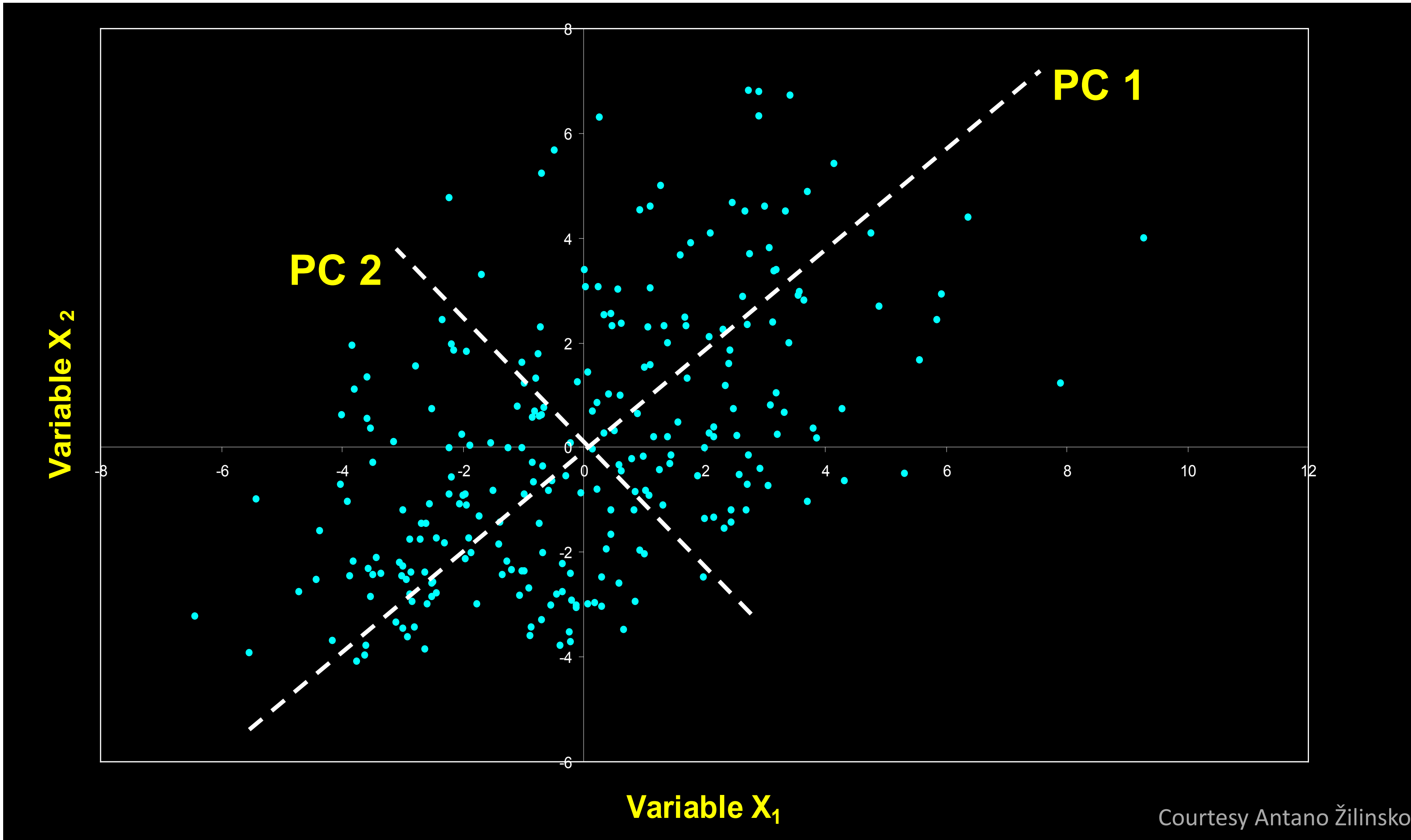
Courtesy Antano Žilinsko

Compute Principal Components

PC 1 has the highest possible variance (9.88)

PC 2 has a variance of 3.03

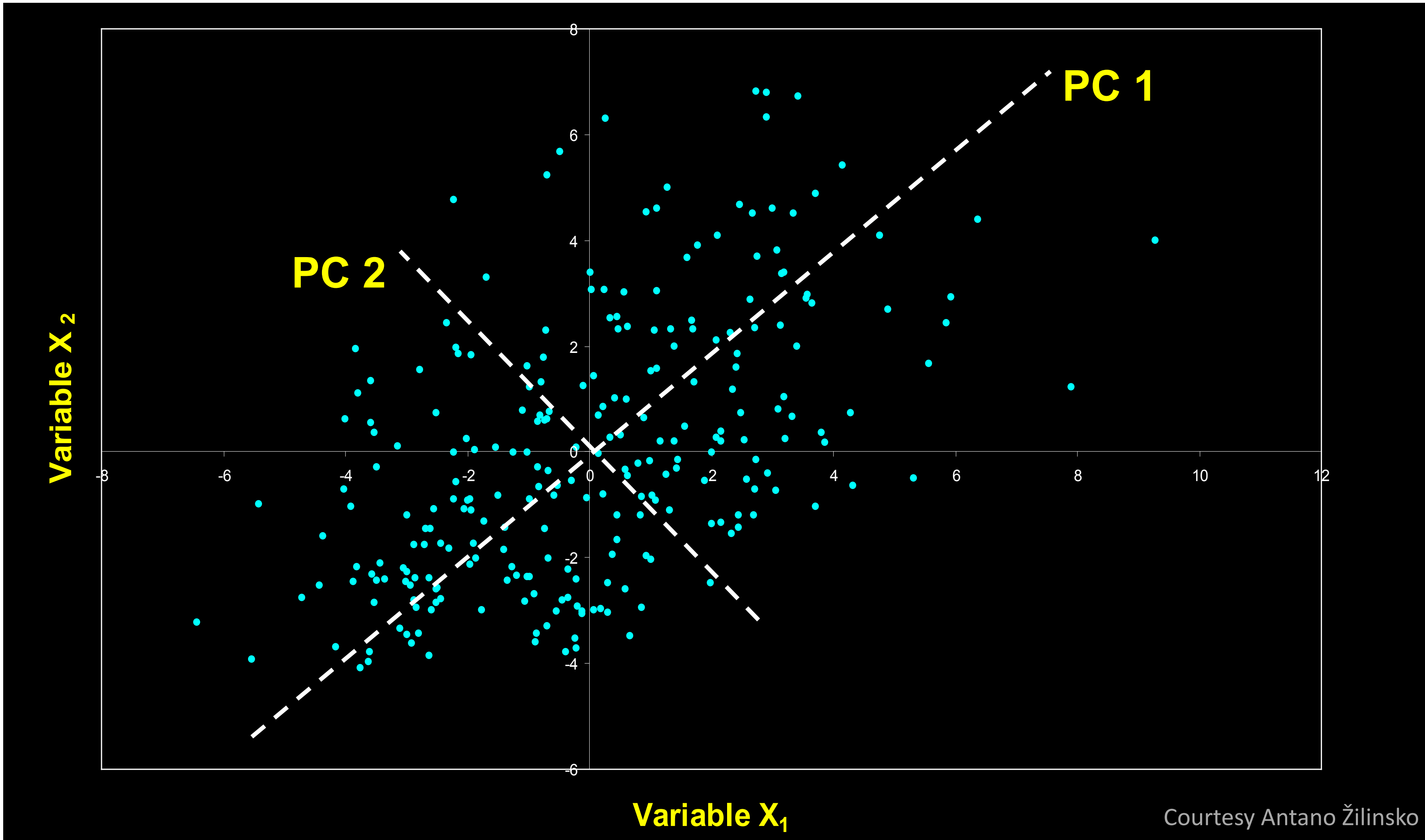
PC 1 and PC 2 have zero covariance.



Courtesy Antano Žilinsko

PC axes are a rigid rotation of the original variables

PC 1 is simultaneously the direction of maximum variance and a least-squares "line of best fit" (squared distances of points away from PC 1 are minimized).



Generalization to *p*-dimensions

if we take the first *k* principal components, they define the *k*-dimensional "hyperplane of best fit" to the point cloud

of the total variance of all *p* variables:

PCs 1 to *k* represent the maximum possible proportion of that variance that can be displayed in *k* dimensions

How many axes are needed?

does the $(k+1)^{th}$ principal axis represent more variance than would be expected by chance?

a common "rule of thumb" when PCA is based on correlations is that axes with eigenvalues > 1 are worth interpreting

Courtesy Antano Žilinsko

PCA as Reconstruction Error

$Z = XU$

NxD DxL

$$\min_{U} |X - ZU^T|^2 =$$

PCA as Reconstruction Error

$$Z = XU$$

NxD DxL

$$\min_U |X - ZU^T|^2 =$$

$$\min_U |X - XUU^T|^2 =$$

PCA as Reconstruction Error

$$Z = XU$$

NxD DxL

$$\min_U |X - ZU^T|^2 =$$

$$\min_U |X - XUU^T|^2 =$$

$$\min_U 2|X|^2 - 2U^T X^T XU =$$

PCA as Reconstruction Error

$Z = XU$

NxD DxL

$$\min_U |X - ZU^T|^2 =$$

$$\min_U |X - XUU^T|^2 =$$

$$\min_U 2|X|^2 - 2U^T X^T X U =$$

$$\min_U C - 2|XU|^2$$

maximizing variance $\leftrightarrow$ minimizing reconstruction error

Slides Credit

https://www.mii.lt/zilinskas/uploads/visualization/lectures/lect4/lect4_pca/PCA1.ppt