Lecture VII: PCA: MSRE and Relative Information Loss

Bernhard Geiger

Graz University of Technology

Winter Term 2017/18

PCA

We have an N-dimensional random vector $X = (X_1, \ldots, X_N)^T$ with PDF f_X and positive semi-definite covariance matrix $\mathbf{C}_X := \mathbb{E}\left(XX^T\right)$.

- Eigenvalue decomposition: $\mathbf{C}_X = \mathbf{W}\mathbf{\Sigma}\mathbf{W}^T$
- PCA: $Y = \mathbf{W}^T X$
- Dimensionality Reduction: $\tilde{Y} := Y_1^M = \mathbf{I}_M Y$, where $\mathbf{I}_M$ is a projection onto the first M coordinates

Question 1

Q1: Can we find a linear transform of X such that the transformed random vector has uncorrelated components?

Question 1

Q1: Can we find a linear transform of X such that the transformed random vector has uncorrelated components?

PCA

Ingredient: Eigenvalue Decomposition

Let $\mathbf{A}$ be a real-valued, symmetric, positive semi-definite $N \times N$-matrix. Then, $\mathbf{A}$ can be written as

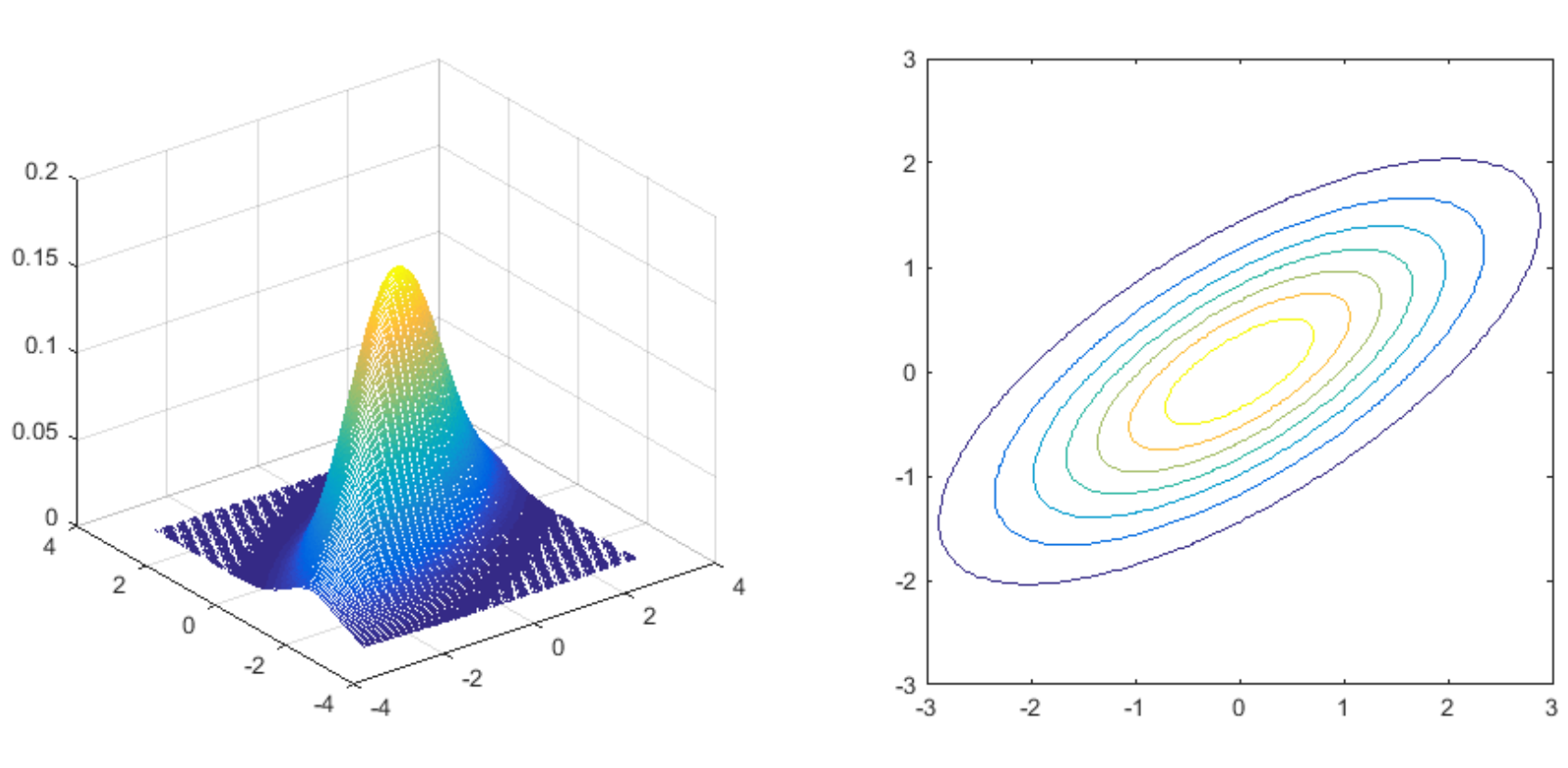
$$\mathbf{A} = \mathbf{W}\mathbf{\Sigma}\mathbf{W}^T$$

where $\mathbf{W}$ is orthogonal and collects the eigenvectors of $\mathbf{A}$ and where

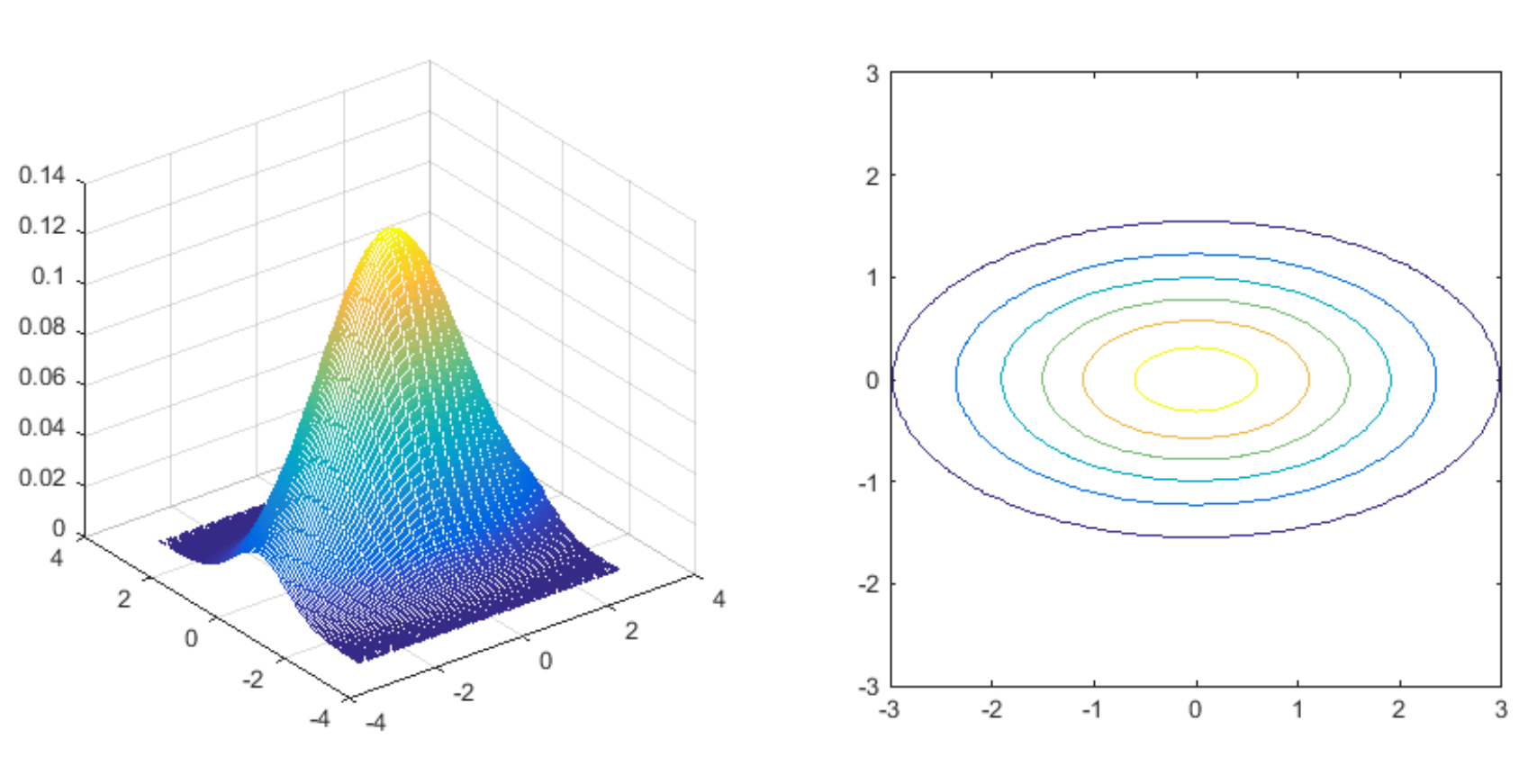
$$\mathbf{\Sigma} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ are the eigenvalues of $\mathbf{A}$.

Decorrelating Transform



Decorrelating Transform



Question 2

We have an N-dimensional random variable $X = (X_1, \ldots, X_N)^T$ with PDF f_X.

Q2: Which linear transform of X to $M < N$ coordinates minimizes the linear mean-squared reconstruction error?

Question 2

We have an N-dimensional random variable $X = (X_1, \ldots, X_N)^T$ with PDF f_X.

Q2: Which linear transform of X to $M < N$ coordinates minimizes the linear mean-squared reconstruction error?

PCA

Ingredient: Trace Bound

Let $\mathbf{A}$ be a real-valued, symmetric, positive semi-definite $N \times N$-matrix, and let $\mathbf{U}$ be a real-valued $M \times N$-matrix ($M < N$) that satisfies $\mathbf{U}\mathbf{U}^T = \mathbf{I}$. Then,

$$\mathrm{tr}(\mathbf{U}\mathbf{A}\mathbf{U}^T) \leq \sum_{i=1}^{M} \lambda_i$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ are the eigenvalues of $\mathbf{A}$.

Question 3

We have an N-dimensional random variable $X = (X_1, \ldots, X_N)^T$ with PDF f_X.

Q2: Which linear transform of X to $M < N$ coordinates minimizes the relative information loss?

Question 3

We have an N-dimensional random variable $X = (X_1, \ldots, X_N)^T$ with PDF f_X.

Q2: Which linear transform of X to $M < N$ coordinates minimizes the relative information loss?

None, any, etc.

Ingredients

- Orthogonal transforms are bi-Lipschitz
- Bi-Lipschitz transforms do not change information dimension
- Relative information loss in projections, given a continuous input:

$$I(X \to Y) = \frac{N - M}{N}$$

It get's worse...

PCA depends on the covariance matrix $\mathbf{C}_X$.

It get's worse...

PCA depends on the covariance matrix $\mathbf{C}_X$.

What happens if we don't know $\mathbf{C}_X$?

It get's worse...

PCA depends on the covariance matrix $\mathbf{C}_X$.

What happens if we don't know $\mathbf{C}_X$?

Let $\mathbf{X}$ be a *matrix* where each of its n columns is an independent copy of the N-dimensional random vector X.

$$\mathbf{C}_X \approx \hat{\mathbf{C}}_X := \frac{1}{n}\mathbf{X}\mathbf{X}^T$$

PCA with Sample Covariance Matrix

- Eigenvalue Decomposition:

$$\hat{\mathbf{C}}_X = \hat{\mathbf{W}}\hat{\mathbf{\Sigma}}\hat{\mathbf{W}}^T$$

where $\hat{\mathbf{W}} = \hat{\mathbf{W}}(\mathbf{X})$.

PCA with Sample Covariance Matrix

- Eigenvalue Decomposition:

$$\hat{\mathbf{C}}_X = \hat{\mathbf{W}}\hat{\mathbf{\Sigma}}\hat{\mathbf{W}}^T$$

  where $\hat{\mathbf{W}} = \hat{\mathbf{W}}(\mathbf{X})$.

- PCA:

$$\mathbf{Y} = \hat{\mathbf{W}}(\mathbf{X})^T\mathbf{X}$$

  becomes *non-linear*!

PCA with Sample Covariance Matrix

- Eigenvalue Decomposition:

$$\hat{\mathbf{C}}_X = \hat{\mathbf{W}}\hat{\mathbf{\Sigma}}\hat{\mathbf{W}}^T$$

  where $\hat{\mathbf{W}} = \hat{\mathbf{W}}(\mathbf{X})$.

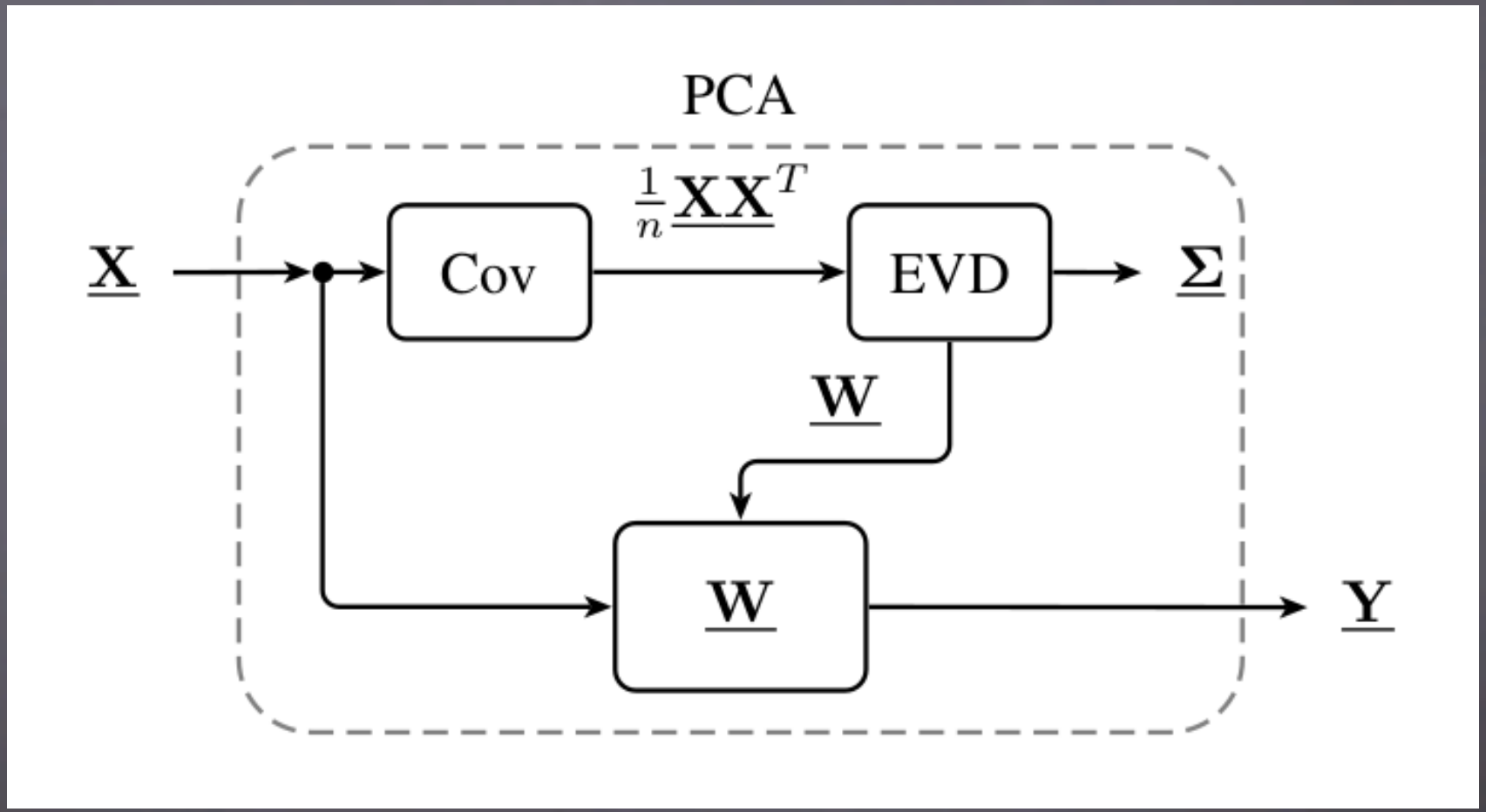
- PCA:

$$\mathbf{Y} = \hat{\mathbf{W}}(\mathbf{X})^T\mathbf{X}$$

  becomes *non-linear*!

- (Let's forget about dimensionality reduction for now.)

PCA with Sample Covariance Matrix



PCA with Sample Covariance Matrix

PCA is lossy even without dimensionality reduction:

$$I(\mathbf{X} \to \mathbf{Y}) = \begin{cases} \frac{N-1}{2n}, & n \geq N \\ \frac{2N-n-1}{2N}, & n < N \end{cases}$$

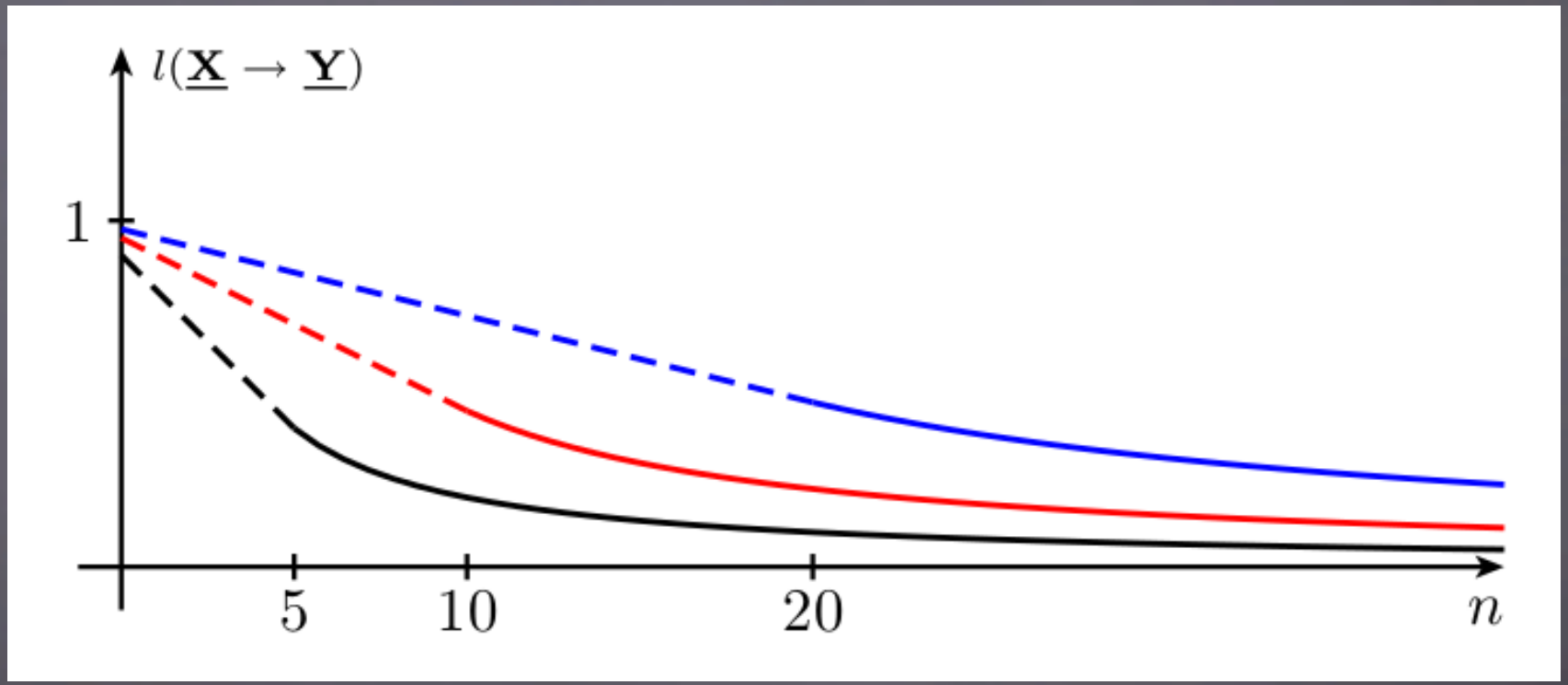
(Follows by simply counting degrees of freedom in $\hat{\mathbf{W}}$.)

PCA with Sample Covariance Matrix

PCA is lossy even without dimensionality reduction:

$$I(\mathbf{X} \to \mathbf{Y}) = \begin{cases} \frac{N-1}{2n}, & n \geq N \\ \frac{2N-n-1}{2N}, & n < N \end{cases}$$

(Follows by simply counting degrees of freedom in $\hat{\mathbf{W}}$.)

We lose information about the direction of X

PCA with Sample Covariance Matrix