Introduction to Machine Learning

Session 2

Dataviz
Dimensionality reduction

Maxime Ossonce

1. Introduction and Dataviz
1.1 Intro (previous lesson)
1.2 Dataviz and dimensionality reduction

Unsupervised learning

- Supervised learning – predictive models:
  - trained on **labeled** training set,
  - expected to generalize on (unseen) **test** samples.
- Unsupervised learning – descriptive models:
  - study data distribution,
  - extract knowledge for data,
  - **clustering,** dataviz...

Notations

- The matrix X is the **data matrix**. If p is the **dimension** of the samples, X is of size $n \times p$: the ith row of X is the ith sample $x^i \in \mathbb{R}^d$.
- The jth variable (or **feature**) of x^i, $x^i_j \in \mathbb{R}$ is the component X_{ij} of the data matrix.

$$X = \begin{array}{cc} \overset{p \text{ variables}}{\longleftrightarrow} & \\ \begin{pmatrix} x_1^1 & \cdots & x_j^1 & \cdots & x_p^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_1^i & \cdots & x_j^i & \cdots & x_p^i \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_j^n & \cdots & x_j^n & \cdots & x_p^n \end{pmatrix} & \updownarrow \, n \text{ samples} \end{array}$$

row i
sample x^i

rcc	wcc	hc	hg	ferr	bmi	ssf	pcBfat	lbm	ht	wt
3.96	7.5	37.5	12.3	60	20.56	109.1	19.75	63.32	195.9	78.9
4.41	8.3	38.2	12.7	68	20.67	102.8	21.3	58.55	189.7	74.4
4.14	5	36.4	11.6	21	21.86	104.6	19.88	55.36	177.8	69.1
4.11	5.3	37.3	12.6	69	21.88	126.4	23.66	57.18	185	74.9
4.45	6.8	41.5	14	29	18.96	80.3	17.64	53.2	184.6	64.6
4.1	4.4	37.4	12.5	42	21.04	75.2	15.58	53.77	174	63.7
4.31	5.3	39.6	12.8	73	21.69	87.2	19.99	60.17	186.2	75.2
4.42	5.7	39.9	13.2	44	20.62	97.9	22.43	48.33	173.8	62.3
4.3	8.9	41.1	13.5	41	22.64	75.1	17.95	54.57	171.4	66.5
4.51	4.4	41.6	12.7	44	19.44	65.1	15.07	53.42	179.9	62.9
4.71	5.3	41.4	**14**	38	25.75	171.1	28.83	68.53	193.4	96.3
4.62	7.3	43.8	14.7	26	21.2	76.8	18.08	61.85	188.7	75.5
4.35	7.8	41.4	14.1	30	22.03	117.8	23.3	48.32	169.1	63
4.26	6.2	41	13.9	48	25.44	90.2	17.71	66.24	177.9	80.5
4.63	6	43.7	14.7	30	22.63	97.2	18.77	57.92	177.5	71.3
4.36	5.8	40.3	13.3	29	21.86	99.9	19.83	56.52	179.6	70.5
3.91	7.3	37.6	12.9	43	22.27	125.9	25.16	54.78	181.3	73.2
4.51	8.3	43.7	14.7	34	21.27	69.9	18.04	56.31	179.7	68.7
4.37	8.1	41.8	14.3	53	23.47	98	21.79	62.96	185.2	80.5
4.9	6.9	44	14.5	59	23.19	96.8	22.25	56.68	177.3	72.9
4.46	5.7	39.2	13.0	43	23.17	80.3	16.25	62.39	179.3	74.5

column j
feature j (hemoglobine)

Bivar plot

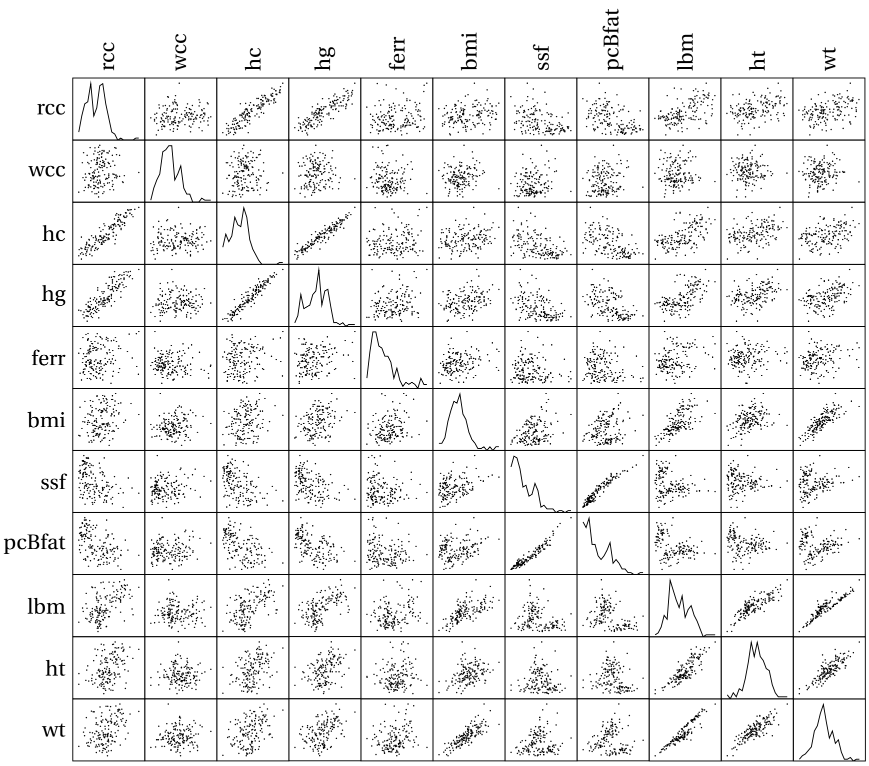


Figure: Bivariate plot of Australian athletes dataset

Dimension reduction

- From $X \in \mathbb{R}^{n\times p}$ to $Z \in R^{n\times q}$.
- Every sample $x^i \in \mathbb{R}^p$ is projected into $\mathbb{R}^q$ with $q < p$.
- $z^i = \mathrm{cod}(x^i) \qquad \tilde{x}^i = \mathrm{dec}(z^i)$
- If x^k is *similar* to x^l then z^k has to be *similar* to z^l.
- $q < p$: features extraction, hidden structure.
- If $q = 2$: visualization.

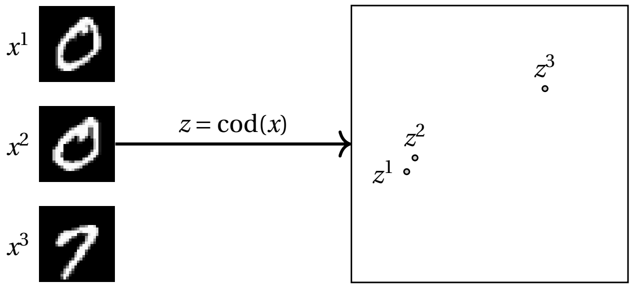
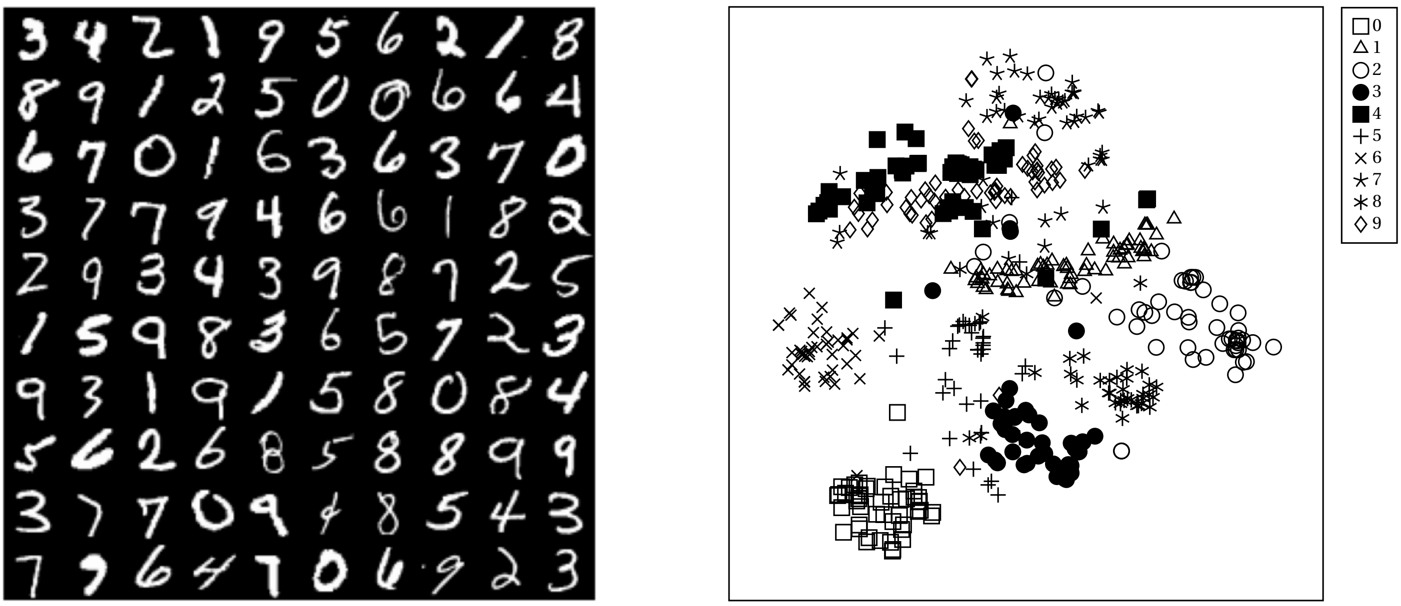


Figure: Illustration $p = 784, q = 2$



(a) Image space: $p = 784$

(b) t-SNE: $q = 2$

Figure: tSNE applied to MNIST dataset

Correlation matrix

- The data **correlation matrix** C of size $p \times p$ is:

$$C = \frac{1}{n} X^\top X.$$

- Hypothesis: X is **centered reduced** ($C_{kk} = 1$ $\frac{1}{n} \sum_i x_j^i = 0$).

$$x_j^i \leftarrow \frac{x_j^i - \bar{x}_j}{\sqrt{\frac{1}{n} \sum_{l=1}^n (x_j^l - \bar{x}_j)^2}}.$$

Correlation coefficients

$$C_{kl} = \frac{1}{n} \sum_{i=1}^n x_k^i x_l^i \quad \forall k, l \in \{1 \dots p\}.$$

Visualization

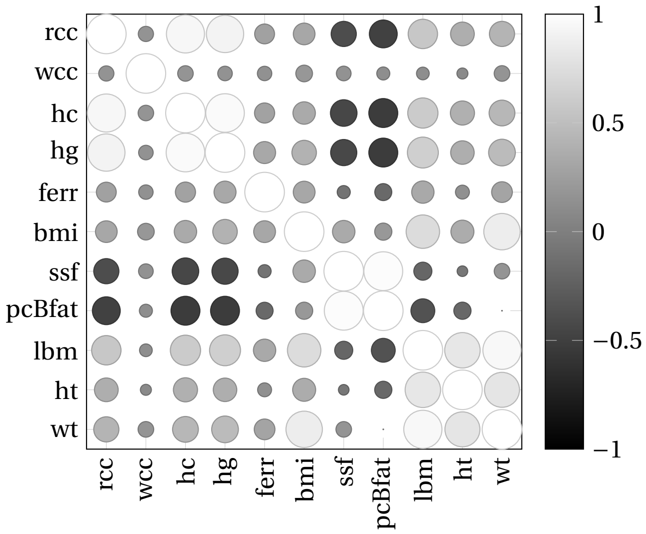


Figure: Australian athletes dataset correlation matrix

Diagonalization I

- C is a **definite positive** matrix.
- Let $\lambda_1, \ldots, \lambda_p$ its (positive) **eigenvalues** in decreasing order.
- Let $u^1, \ldots, u^p$ the orthonormal diagonalization basis:
  - $\sum_{j=1}^{p} u_j^k u_j^l = \delta_{kl}$
  - $Cu^j = \lambda_j u^j$.
- $C = UDU^\top$ where the columns of U are the **eigenvectors** u^j and D is the **diagonal** matrix of eigenvalues.

Diagonalization II

- $V = XU$ is the data matrix in the **eigenbasis** $(u^1, \ldots u^p)$.

$$v = U^\top x \qquad x = Uv.$$

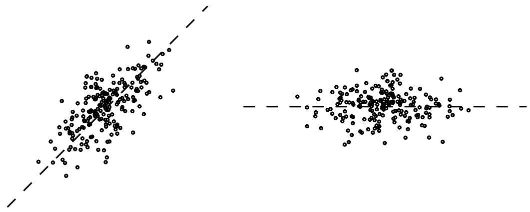
- Its covariance matrix is:

$$\begin{aligned}\frac{1}{n} V^\top V &= \frac{1}{n} U^\top X^\top XU \\ &= U^\top CU \\ &= D.\end{aligned}$$

- In the eigenbasis, the coordinates v of x are **uncorrelated**.

$p = 2$

- $C = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}$
- $\lambda_1 = 1 + a$,
- $\lambda_2 = 1 - a$.
- $u^1 = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})^\top$,
- $u^2 = (-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})^\top$



normalized data

normalized data in eigenbasis

Principal component analysis (PCA)

- **Linear** method.
- W **orthogonal** matrix of size $p \times q$ ($W^\top W = I_q$) such that:
  - $z^i = W^\top x^i$.
  - $\tilde{x}^i = W z^i$.
- $Z = XW$.

$$X = ZW^\top + B.$$

- B is a **matrix** noise orthogonal of W.

$$W = \underset{\mathbb{R}^{p\times q}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \|x^i - \tilde{x}^i\|^2.$$

Variance maximization

- W that minimizes reconstruction error maximizes variance of z:

$$
\begin{aligned}
\|x-\tilde{x}\|^2 &= (x-\tilde{x})^\top (x-\tilde{x}) \in \mathbb{R} \\
&= x^\top x - 2\tilde{x}^\top x + \tilde{x}^\top \tilde{x} \\
&= \|x\|^2 - 2z^\top W^\top x + z^\top W^\top W z \\
&= \|x\|^2 - 2z^\top z + z^\top W^\top W z \\
&= \|x\|^2 - 2z^\top z + z^\top z \\
&= \|x\|^2 - \|z\|^2.
\end{aligned}
$$

- $\arg\min \sum_i \|x^i - \tilde{x}^i\|^2 = \arg\max \sum_i \|z^i\|^2$.
- The PCA maximizes the projection z **variance**.

PCA solution $q = 1$

- We want to find $w^1 \in \mathbb{R}^d$ that maximizes the variance of $z = Xw^1$.
- The variance of z is:

$$
\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} z_j^2 &= \frac{1}{n} w^{1\top} X^\top X w_j^{1\top} \\
&= w^{1\top} C w^1 \\
&= w'^\top D w' && w' = U^\top w^1 \\
&= \sum_{j=1}^{d} \lambda_j w_j'^2 && (1)
\end{aligned}
$$

- The vector s.t. $\|w^1\|^2 = \|w'\|^2 = 1$ that maximizes (1) is $w'^\top = (1, 0 \ldots 0)$: $w^1 = u^1$ the first **eigenvector**.

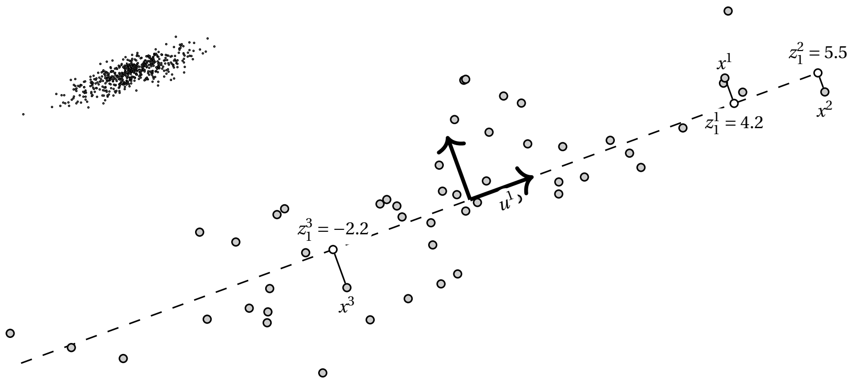


Figure: First component of a PCA

PCA general solution

- $Z = XW$ with $W^\top W = I_q$.
- The columns of W are the q first **eigenvectors**.

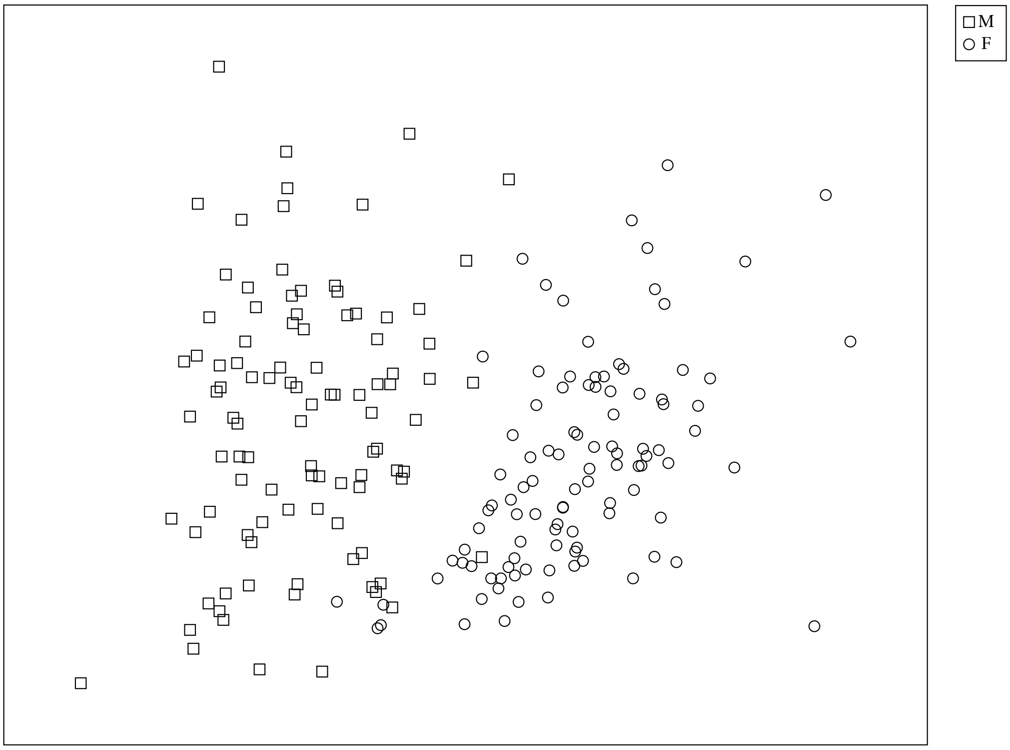


Figure: PCA on the australian athletes dataset

Choice of *q*

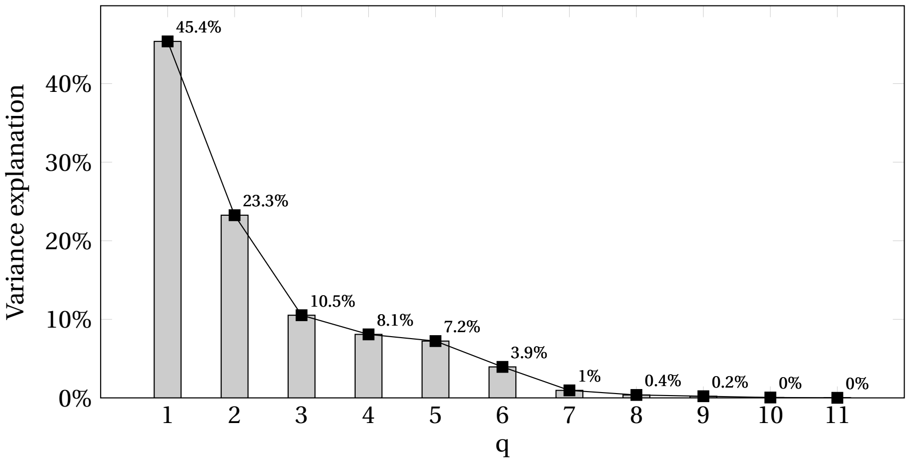


Figure: Explained variance of PCA components

- Cross validation.
- Detection of an **elbow** in the variance explained.
- 95 % of total variance explained.

Drawbacks of PCA

- Liner method.
- Gaussian assumption on the data.
- Other methods: t-distributed stochastic neighbor embedding (tSNE), autoencoders…

Principle of tSNE

- $d(x^i, x^j) \to p_x(x^j|x^i)$ (probability that x^j is close to x^i).
- Assume same probability distribution for the projections $p_z(z^j|z^i)$.
- Find p_z close to p_x.

Details

- Distribution $p_x(x^j|x^i)$:

$$p_x(x^j|x^i) = \frac{e^{-d_{ij}^2}}{\sum_k e^{-d_{ik}^2}}$$
$$d_{ij} = \frac{\|x^i - x^j\|}{\sigma^2}.$$

- Distribution $p_z(z^j|z^i)$:

$$p_z(z^j|z^i) = \frac{(1+\delta_{ij}^2)^{-1}}{\sum_k (1+\delta_{ik}^2)^{-1}}$$
$$\delta_{ij} = \|z^i - z^j\|.$$

- Find $(z^i)_{i=\{1\ldots n\}}$ that minimizes Kullback-Leibler divergence (DKL):

$$\sum_{i,j} p_x(x^j|x^i) \log\left(\frac{p_x(x^j|x^i)}{p_z(z^j|z^i)}\right).$$

Illustration

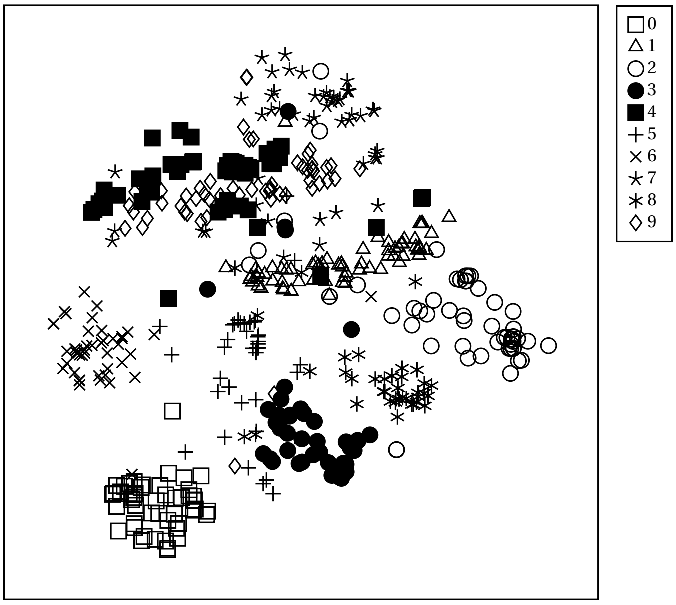


Figure: tSNE on MNIST dataset