Remise à niveau - Statistique
Réduction de dimension

Claire Boyer

université
PARIS-SACLAY

Dimensionality Curse

High Dimension Geometry Curse

- Folks theorem: In high dimension, everyone is alone.
- Theorem: If $X_1, \ldots, X_n$ in the hypercube of dimension d such that their coordinates are i.i.d then

$$d^{-1/p}\left(\max \|X_i - X_j\|_p - \min \|X_i - X_j\|_p\right) = 0 + O\left(\sqrt{\frac{\log n}{d}}\right)$$

$$\frac{\max \|X_i - X_j\|_p}{\min \|X_i - X_j\|_p} = 1 + O\left(\sqrt{\frac{\log n}{d}}\right).$$

- When d is large, all the points are almost equidistant...
- Nearest neighbors are meaningless!

Dimension Reduction

- How to view a high-dimensional dataset?
- High-dimension: dimension larger than 2!
- *Projection* in a 2D space.

Dimension Reduction

- How to view a high-dimensional dataset?
- High-dimension: dimension larger than 2!
- *Projection* in a 2D space.

Dimension Reduction

- How to view a high-dimensional dataset?
- High-dimension: dimension larger than 2!
- *Projection* in a 2D space.

Dimension Reduction

- How to view a high-dimensional dataset?
- High-dimension: dimension larger than 2!
- *Projection* in a 2D space.

Dimension reduction

Dimension reduction

- Training data $\mathcal{D} = \{x_1, \ldots, x_n\} \in \mathcal{X}^n \quad$ (i.i.d. $\sim \mathbb{P}$)
- Space $\mathcal{X}$ of possibly high dimension.

Dimension Reduction Map

Construct a map Φ from the space $\mathcal{X}$ into a space $\mathcal{X}'$ of smaller dimension:

$$\begin{aligned} \Phi : \quad & \mathcal{X} \to \mathcal{X}' \\ & x \mapsto \Phi(x) \end{aligned}$$

Criterion

- Reconstruction error focus here in these slides
- Distance preservation

Reconstruction Error Approach

- Construct a map Φ from the space $\mathcal{X}$ into a space $\mathcal{X}'$ of smaller dimension:

$$\begin{aligned} \Phi : \quad & \mathcal{X} \to \mathcal{X}' \\ & x \mapsto \Phi(x) \end{aligned}$$

- Construct $\widetilde{\Phi}$ from $\mathcal{X}'$ to $\mathcal{X}$
- Control the error between x and its reconstruction $\widetilde{\Phi}(\Phi(x))$
- Canonical example for $x \in \mathbb{R}^d$: find Φ and $\widetilde{\Phi}$ in a parametric family that minimize

$$\frac{1}{n} \sum_{i=1}^{n} \|x_i - \widetilde{\Phi}(\Phi(x_i))\|^2$$

Inertia

- $x_1, \ldots, x_n \in \mathbb{R}^d$
- $m = \frac{1}{n} \sum_{i=1}^{n} x_i$

Two views on inertia

- Inertia:

$$I = \frac{1}{n} \sum_{i=1}^{n} \|x_i - m\|^2$$
$$= \frac{1}{2n^2} \sum_{i,j} \|x_i - x_j\|^2$$

- 2 times the mean squared distance to the mean = Mean squared distance between individual

- Heuristic: a good representation is a representation with a large inertia
- Large dispersion $\sim$ Large average separation!

Inertia and Projection

- What if we replace x by its projection $\widetilde{x} = P(x - m) + m$?

Two views on inertia

- Inertia:

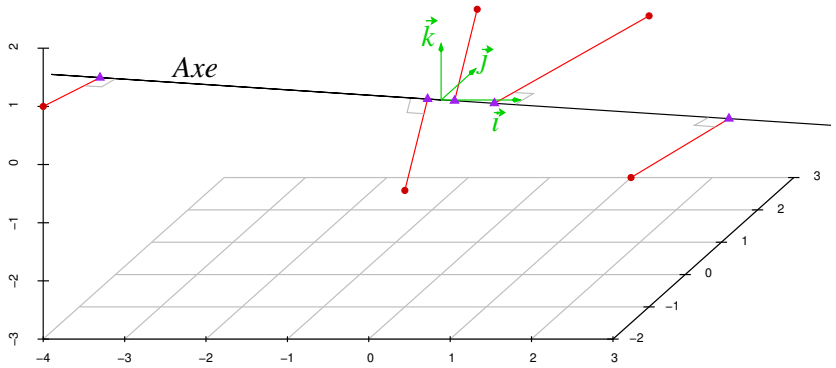
$$\widetilde{I} = \frac{1}{n}\sum_{i=1}^{n} \|\widetilde{x}_i - m\|^2$$
$$= \frac{1}{2n^2}\sum_{i,j} \|\widetilde{x}_i - \widetilde{x}_j\|^2$$

- Inertia:

$$\widetilde{I} = I - \frac{1}{n}\sum_{i=1}^{n} \|\widetilde{x}_i - x_i\|^2$$
$$= I - \frac{1}{2n^2}\sum_{i,j} \left(\|x_i - x_j\|^2 - \|\widetilde{x}_i - \widetilde{x}_j\|^2\right)$$

- Four different way to obtain a large inertia!

First Component of the PCA



- 1D case: $\widetilde{x} = m + a^\top (x - m)a$ with $\|a\| = 1$
- Inertia: $\widetilde{I} = \frac{1}{n} \sum_{i=1}^{n} a^\top (x_i - m)(x_i - m)^\top a$

Principal Component Analysis : optimization of the projection

- Maximization of

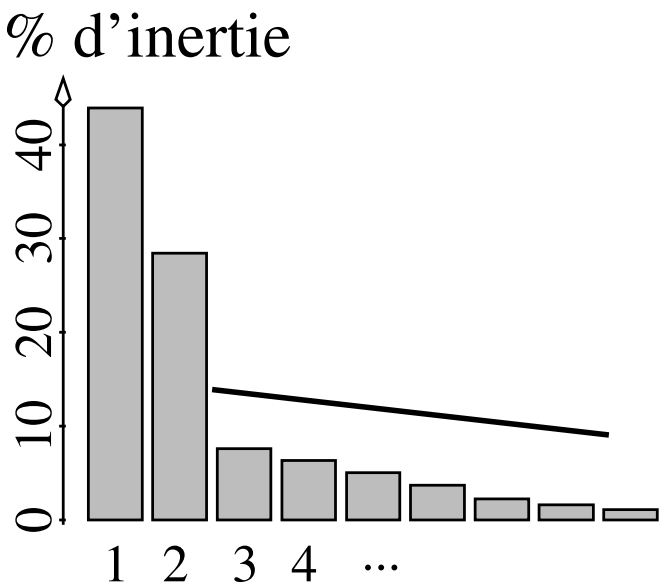
$$\widetilde{I} = \frac{1}{n} \sum_{i=1}^{n} a^\top (x_i - m)(x_i - m)^\top a = a^\top \Sigma a$$

with $\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - m)(x_i - m)^t$ the empirical covariance matrix.

- Explicit optimal choice given by the eigenvector of the largest eigenvalue of Σ.

Principal Component Analysis : optimization of the projection

- Explicit optimal solution obtain by the projection on the eigenvectors of the largest eigenvalues of Σ.
- Projected inertia given by the sum of those eigenvalues.



- Often fast decay of the eigenvalues: some dimensions are much more important than other.
- Not exactly the curse of dimensionality setting...
- Yet a lot of *small* dimension can drive the distance

Take-home message

- Principal components = Eigenvectors of the empirical covariance matrix
- Principal components = "New" variables obtained as linear combinations of the initial variables
  - less interpretable
  - but capture the dataset variance better

PCA as a reconstruction error approach I

- $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{X}' = \mathbb{R}^{d'}$
- Linear map with V orthonormal

$$\Phi(x) = V^\top (x - m) \quad \text{and} \quad \widetilde{\Phi}(x') = m + Vx'$$

- Reconstruction error criterion:

$$\frac{1}{n} \sum_{i=1}^{n} \|x_i - (m + VV^\top (x_i - m)\|^2$$

- *Explicit solution:*
  - m is the empirical mean
  - V is any orthonormal basis of the space spanned by the d' first eigenvectors (the one with largest eigenvalues) of the empirical covariance matrix $\frac{1}{n} \sum_{i=1}^{n} (x_i - m)(x_i - m)^\top$.

PCA as a reconstruction error approach II

PCA Algorithm

- Compute the empirical mean $m = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Compute the empirical covariance matrix $\frac{1}{n}\sum_{i=1}^{n}(x_i - m)(x_i - m)^\top$.
- Compute the d' first eigenvectors of this matrix: $V^{(1)}, \ldots, V^{(d')}$
- Set $\Phi(x) = V^\top(x - m)$

Remarks

- Complexity: $O(n(1 + d^2) + d'd^2)$
- Interpretation:
  $\Phi(x) = V^\top(x - m)$: coordinates in the restricted space
  $V^{(i)}$: influence of each original coordinates in the ith new one.
- Not invariant to a scaling of the variables
  $\rightsquigarrow$ It is custom to normalize the variables before applying PCA.

Multiple Factor Analysis I

- PCA assumes $\mathcal{X} = \mathbb{R}^d$
- How to deal with categorical values?
- MFA = PCA with clever coding strategy for categorical values.

Categorical value code for a single variable

- Classical redundant dummy coding:

$$x \in \{1, \ldots, C\} \mapsto P(x) = (1_{x=1}, \ldots, 1_{x=C})^t$$

- Compute the mean (i.e. the empirical proportions): $\overline{P} = \frac{1}{n} \sum_{i=1}^{n} P(x_i)$
- *Renormalize* $P(x)$ by $1/\sqrt{(C-1)\overline{P}}$:

$$P(x) \mapsto P^r(x)$$

$$(1_{x=1}, \ldots 1_{x=C}) \mapsto \left(\frac{1_{x=1}}{\sqrt{(C-1)\overline{P}_1}}, \ldots, \frac{1_{x=C}}{\sqrt{(C-1)\overline{P}_V}} \right)$$

Multiple Factor Analysis II

- PCA becomes the minimization of

$$\frac{1}{n}\sum_{i=1}^{n}\|P^r(\mathrm{x}_i) - (m + VV^t(P^r(\mathrm{x}_i) - m))\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{v=1}^{V}\frac{\left|1_{\mathrm{x}_i=v} - (m' + \sum_{l=1}^{d'} V^{(l)t}(P(\mathrm{x}_i) - m')V^{(l,v)})\right|^2}{(V-1)\overline{P}_v}$$

- Interpretation:
  - $m' = \overline{P}$
  - $\Phi(\mathrm{x}) = V^t(P^r\mathrm{x} - m)$: coordinates in the restricted space.
  - $V^{(l)}$ can be interpreted s as a probability profile.
- Complexity: $O(n(V + V^2) + d'V^2)$
- Link with Correspondence Analysis (CA)

Multiple Factor Analysis III

MFA Algorithm

- Redundant dummy coding of each categorical variable.
- Renormalization of each block of dummy variable.
- Classical PCA algorithm on the resulting variables

- Interpretation as a reconstruction error with a rescaled/χ^2 metric.
- Interpretation:
  - $\Phi(x) = V^t(P^r(x) - m)$: coordinates in the restricted space.
  - $V^{(l)}$: influence of each modality/variable in the ith new coordinates.
- Scaling: This method is not invariant to a scaling of the continuous variables
  $\rightsquigarrow$ It is custom to normalize the variables (at least within groups) before applying PCA.

Non Linear PCA I

PCA Model

- PCA: Linear model assumption

$$\mathrm{x} \simeq m + \sum_{l=1}^{d'} \mathrm{x}'^{,(l)} V^{(l)} = m + V\mathrm{x}'$$

- with
  - $V^{(l)}$ orthonormal
  - $\mathrm{x}'^{,(l)}$ without constrains.

- Two directions of extension:
  - Other constrains on V (or the coordinates in the restricted space): ICA, NMF, Dictionary approach
  - PCA on a non linear image of x: kernel-PCA
- Much more complex algorithm!

Non Linear PCA II

ICA (Independent Component Analysis)

- Linear model assumption

$$x \simeq m + \sum_{l=1}^{d'} x'^{,(l)} V^{(l)} = m + V x'$$

- with
  - $V^{(l)}$ without constrains.
  - $x'^{,(l)}$ independent

NMF (Non Negative Matrix Factorization)

- (Linear) Model assumption

$$x \simeq m + \sum_{l=1}^{d'} x'^{,(l)} V^{(l)} = m + V x'$$

- with
  - $V^{(l)}$ non negative
  - $x'^{,(l)}$ non negative.

Non Linear PCA III

Dictionary

- (Linear) Model assumption

$$x \simeq m + \sum_{l=1}^{d'} x'^{,(l)} V^{(l)} = m + V x'$$

- with
  - $V^{(l)}$ without constrains
  - x' sparse (with a lot of 0)

kernel PCA

- Linear model assumption

$$\Psi(x - m) \simeq \sum_{l=1}^{d'} x'^{,(l)} V^{(l)} = V x'$$

- with
  - $V^{(l)}$ orthonormal
  - x'_l without constrains.

Auto Encoder

Deep Auto Encoder

- Construct a map Φ with a NN from the space $\mathcal{X}$ into a space $\mathcal{X}'$ of smaller dimension:

$$\begin{aligned} \Phi : \quad & \mathcal{X} \to \mathcal{X}' \\ & x \mapsto \Phi(x) \end{aligned}$$

- Construct $\widetilde{\Phi}$ with a NN from $\mathcal{X}'$ to $\mathcal{X}$
- Control the error between x and its reconstruction $\widetilde{\Phi}(\Phi(x))$:

$$\frac{1}{n} \sum_{i=1}^{n} \|x_i - \widetilde{\Phi}(\Phi(x_i))\|^2$$

- Optimization by gradient descent.
- NN can be replaced by another parametric function...

Another of approach via distance preservation

Pairwise distances

- Different point of view: input is distances between feature vectors
- Use only distances $d(x_i, x_j)$.

Distance Preservation. Construct a map Φ from the space $\mathcal{X}$ into a space $\mathcal{X}'$ of smaller dimension:

$$\begin{aligned} \Phi: \quad & \mathcal{X} \to \mathcal{X}' \\ & x \mapsto \Phi(x) = x' \\ & d(x_i, x_j) \approx d'(x_i', x_j') \end{aligned}$$

such that

A natural criterion is

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| d(x_i, x_j) - d'(x_i', x_j') \right|^2$$

Random Projections

- Distance preserving transform
- Random transform
- Data-independent!
- Dimension-independent!

Theorem (Johnson-Lindenstrauss Lemma)

Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a finite set of vectors with cardinality $n \geq 2$ and $W \in \mathbb{R}^{p \times d}$ be a random matrix such that its entries $\{W_{k\ell}\}_{\substack{1 \leq k \leq p \\ 1 \leq \ell \leq d}}$ are i.i.d. and distributed according to $\mathcal{N}\left(0, \frac{1}{p}\right)$. For any $(\varepsilon, \delta) \in (0,1)^2$, if

$$p \geq 16 \varepsilon^{-2} \log(n/\sqrt{\delta}),$$

then with probability at least $1 - \delta$ on the random matrix W,

$$\forall (x_i, x_j) \in \mathcal{S}^2 : \quad (1-\varepsilon) \left\|x_i - x_j\right\|^2 \leq \left\|Wx_i - Wx_j\right\|^2 \leq (1+\varepsilon) \left\|x_i - x_j\right\|^2 .$$

The mapping $x \in \mathbb{R}^d \mapsto Wx \in \mathbb{R}^p$ is called an ε-isometry on $\mathcal{S}$.

Random projections I

The underlying idea of this approach is that the reduction mapping $x \in \mathbb{R}^d \mapsto Wx \in \mathbb{R}^p$ is an exact isometry "in expectation":

$$\forall x \in \mathbb{R}^d : \quad \mathbb{E}\left(\|Wx\|^2\right) = \|x\|^2 .$$

Indeed, since for all $x \in \mathbb{R}^d$ such that $x \neq 0$, $\frac{p\|Wx\|^2}{\|x\|^2} \sim \chi^2_p$, one has

$$\mathbb{E}\left(\|Wx\|^2\right) = \frac{\|x\|^2}{p}\mathbb{E}\left(\frac{p\,\|Wx\|^2}{\|x\|^2}\right) = \frac{\|x\|^2}{p}p = \|x\|^2 .$$

Random projections II

Let us remark that it is enough for $\{W_{ij}\}_{\substack{1\leq i\leq p\\ 1\leq j\leq d}}$ to be independent with $\mathbb{E}W_{ij} = 0$ and $\mathbb{V}(W_{ij}) = \frac{1}{p}$ (for all $i \in [p]$ and $j \in [d]$) in order to get an exact isometry "in expectation":

$$\mathbb{E}\left(\|Wx\|^2\right) = \sum_{i=1}^{p} \mathbb{E}\left[\left(\sum_{j=1}^{d} W_{ij}x_j\right)^2\right] = \sum_{i=1}^{p}\left(\sum_{j=1}^{d} x_j^2 \mathbb{V}(W_{ij}) + \left(\sum_{j=1}^{d} x_j \mathbb{E}W_{ij}\right)^2\right)$$

Remark. Note that $p \geq 8\varepsilon^{-2}\log(n^2/\delta)$ does not depend on the original dimension d This means that we could consider data in infinite-dimensional Hilbert space...!

Remark. In practice, Multi-Dimensional Scaling (MDS) techniques such as Isomap are more used.