Lecture 2, Gaussian Estimation

DS-GA 1005 Inference and Representation, Fall 2023

Yoav Wald

09/13/2023

Today's Plan

- Linear estimation with PCA
- Gaussian distributions
- Statistical efficiency vs. expressiveness

Compressing Data

- We observe data that is high dimensional (consisting of many features), which we represent as vectors in $\mathbb{R}^d$,

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}, \quad \mathbf{x}_i \in \mathbb{R}^d \quad \forall i \in [n]$$

- How can we summarize the data effectively?
  - *summarize*: compress each vector $\mathbf{x}_i$ to a k-dimensional vector with $k < d$
  - *effectively*: such that we can obtain an optimal approximate reconstruction, $\tilde{\mathbf{x}}_i \in \mathbb{R}^d$, of the original vector $\mathbf{x}_i$

Linear Compression: Principal Component Analysis

- Let us focus on compressing our data with a linear function, $\tilde{\mathbf{x}}_i = UW\mathbf{x}_i$ for $U \in \mathbb{R}^{d\times k}, W \in \mathbb{R}^{k\times d}$
- Furthermore, we measure reconstruction error with $\|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|_2^2$

$$\min_{U\in\mathbb{R}^{d\times k}, W\in\mathbb{R}^{k\times d}} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - UW\mathbf{x}_i\|_2^2 \quad \text{(PCA)}$$

- What does this have to do with inference and probabilities??

From Data to Probabilities

- We treat each $\boldsymbol{x}_i$ as an i.i.d (identically independently distributed) sample from an underlying distribution P
- We will see that PCA estimates the first two moments of P and uses them to "organize" the data in terms of uncorrelated components

First and Second Moments

- Assume our data is centered, $\hat{\boldsymbol{\mu}} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = 0$
  *otherwise reduce $\hat{\boldsymbol{\mu}}$ from each $\mathbf{x}_i$
- Define the matrix $\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$
- These are empirical estimates of the first two moments of a random vector $\mathbf{x} \in \mathbb{R}^d$
  - Mean $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] \in \mathbb{R}^d$, "center of gravity",
  - Covariance matrix $\Sigma = \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\right] \in \mathbb{R}^{d \times d}$, measures spread of data in each direction

Projections of a Random Vector

- Mean $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] \in \mathbb{R}^d$, "center of gravity",
- Covariance matrix $\Sigma = \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\right] \in \mathbb{R}^{d \times d}$, measures spread of data in each direction
- Let $\mathbf{u} \in \mathbb{R}^d$, and define $z = \langle \mathbf{u}, \mathbf{x} \rangle$. The mean and variance of z are:
  - $\mathbb{E}[z] = \langle \mathbf{u}, \mathbb{E}[\mathbf{x}] \rangle = \langle \mathbf{u}, \boldsymbol{\mu} \rangle$
  - $\mathbb{E}\left[(z - \mathbb{E}[z])^2\right] = \mathbf{u}^\top \Sigma \mathbf{u}$

Back to PCA

- *Question*: How can we organize data into uncorrelated components? (Pearson 1901)
  - Two components $z_1 = \langle \mathbf{u}_1, \mathbf{x} \rangle$ and $z_2 = \langle \mathbf{u}_2, \mathbf{x} \rangle$ are uncorrelated if $\mathrm{Cov}(z_1, z_2) = 0$
  - $\Rightarrow$ covariance matrix of these new components will be diagonal
- Let's go back to the (PCA) problem:

$$\min_{U \in \mathbb{R}^{d \times k}, W \in \mathbb{R}^{k \times d}} \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - UW\mathbf{x}_i\|_2^2$$

Lemma

Let (U, W) be a solution to (PCA), then U is orthogonal (i.e. $U^\top U = \mathbf{I}_k$) and $W = U^\top$.

PCA as an Analysis of Second Moment

- Let us rewrite the problem

$$\min_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - UU^\top \mathbf{x}_i\|_2^2 =$$

$$\min_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \frac{1}{n} \left(\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - 2 \sum_{i=1}^{n} \mathbf{x}_i^\top UU^\top \mathbf{x}_i + \sum_{i=1}^{n} \|UU^\top \mathbf{x}_i\|_2^2 \right)$$

PCA as an Analysis of Second Moment

- Let us rewrite the problem

$$\min_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - UU^\top \mathbf{x}_i\|_2^2 =$$

$$\min_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \frac{1}{n} \left(\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - 2 \sum_{i=1}^{n} \mathbf{x}_i^\top UU^\top \mathbf{x}_i + \sum_{i=1}^{n} \|UU^\top \mathbf{x}_i\|_2^2 \right) =$$

$$\min_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \frac{1}{n} \left(\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{i=1}^{n} \mathbf{x}_i^\top UU^\top \mathbf{x}_i \right)$$

PCA as an Analysis of Second Moment

- Let us rewrite the problem

$$\min_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - UU^\top \mathbf{x}_i\|_2^2 =$$

$$\min_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \frac{1}{n} \left(\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - 2 \sum_{i=1}^{n} \mathbf{x}_i^\top UU^\top \mathbf{x}_i + \sum_{i=1}^{n} \|UU^\top \mathbf{x}_i\|_2^2 \right) =$$

$$\min_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \frac{1}{n} \left(\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \sum_{i=1}^{n} \mathbf{x}_i^\top UU^\top \mathbf{x}_i \right) =$$

$$\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - \frac{1}{n} \max_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \sum_{i=1}^{n} \mathbf{x}_i^\top UU^\top \mathbf{x}_i$$

PCA as an Analysis of Second Moment

- We see that to solve PCA we can also solve

$$\max_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top U U^\top \mathbf{x}_i$$

- Furthermore, let $\mathbf{u}_j$ be the j-th column of U, it holds that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top U U^\top \mathbf{x}_i = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \langle \mathbf{x}_i, \mathbf{u}_j \rangle^2 = \sum_{j=1}^{k} \mathbf{u}_j^\top \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{u}_j$$

- Which leaves us with

$$\max_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \sum_{j=1}^{k} \mathbf{u}_j^\top \widehat{\Sigma} \mathbf{u}_j$$

PCA as an Analysis of Second Moment

Theorem (Spectral Theorem)

If $A \in \mathbb{R}^{d \times d}$ is symmetric then it admits a decomposition as $A = U \Lambda U^\top$. The matrix U is orthonormal $U^\top U = \mathbf{I}_d$ and its columns are the eigenvectors of A. $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ contains the eigenvalues.

- According to the characterization of eigenvectors by the Courant-Fischer theorem the solution to,

$$\max_{U \in \mathbb{R}^{d \times k}, U^\top U = \mathbf{I}_k} \sum_{j=1}^{k} \mathbf{u}_j^\top \widehat{\Sigma} \mathbf{u}_j,$$

  is exactly the k eigenvectors corresponding to the largest eigenvalues of $\hat{\Sigma}$

PCA: some conclusions

- PCA demonstrates a nice connection between summarization of data, and estimation of moments, or expected values
- Example [Novembre et al. 2008]: Gene Expression

PCA: some conclusions

- PCA demonstrates a nice connection between summarization of data, and estimation of moments, or expected values
- Example [Novembre et al. 2008]: Gene Expression

PCA: some conclusions

- PCA demonstrates a nice connection between summarization of data, and estimation of moments, or expected values
- Example [Novembre et al. 2008]: Gene Expression
- It induces the "PCA decomposition" of a vector $\mathbf{x} \in \mathbb{R}^d$

$$\mathbf{x} = \boldsymbol{\mu} + \sum_{i=1}^{d} \langle \mathbf{x} - \boldsymbol{\mu}, \mathbf{u}_i \rangle \mathbf{u}_i$$

PCA: some conclusions

- PCA demonstrates a nice connection between summarization of data, and estimation of moments, or expected values
- Example [Novembre et al. 2008]: Gene Expression
- It induces the "PCA decomposition" of a vector $\mathbf{x} \in \mathbb{R}^d$

$$\mathbf{x} = \boldsymbol{\mu} + \sum_{i=1}^{d} \langle \mathbf{x} - \boldsymbol{\mu}, \mathbf{u}_i \rangle \mathbf{u}_i$$

- Example: "Eigenfaces"

PCA: some conclusions

PCA: Questions Left Unanswered

- What probabilistic model have we learned?
  - We estimated the two first moments, what about the rest?
  - What may be a "natural" probabilistic model behind PCA?
- We are interested in the eigenvectors of the *true* covariance matrix of P, $\Sigma := \mathbb{E}_{\mathbf{x}\sim P}\{[\mathbf{x}-\boldsymbol{\mu}][\mathbf{x}-\boldsymbol{\mu}]^\top\}$. However we only have an estimate $\widehat{\Sigma}$ from samples, how good is it?

The Variational Principle

- PCA estimates the first two moments, while specifying a probability distribution requires all the moments
- We will use the **variational principle** to suggest one way of completing these moments
  - The variational principle will appear in several places along the course
  - Today we will be concerned with choosing the most "smooth", or regular distribution

The Variational Principle

- Denote the domain of our features by $\mathcal{X}$
- **Definition.** $\mathcal{P}(\mathcal{X})$ is the space of all probability distributions over $\mathcal{X}$
- *Question*: How can we define a notion of regularity, or smoothness for $p \in \mathcal{P}(\mathcal{X})$?
  - Assume $\mathcal{X} = \{1, \ldots, L\}$ for some integer L. what is the smoothest distribution over $\mathcal{X}$?
  - How about $\mathcal{X} = [0, 1]$?

Maximum Entropy Distributions

- Entropy is a common quantity to measure uncertainty, or smoothness
- For a distribution over $\mathcal{X} = [L]$, it is

$$H(p) = -\sum_{i=1}^{L} p_i \log p_i$$

- Intuition, in a nutshell: the function $-\log(p_i)$ measures the information gained from observing the occurrence of the i-th state out of the possible L states (has axiomatic characterization)
  - Events that have low probability, or occur with less certainty, carry more information
  - $H(p)$ is the expected information, and also a measure of uncertainty

Maximum Entropy Distributions

- Entropy is a common quantity to measure uncertainty, or smoothness
- For a distribution over $\mathcal{X} = [L]$, it is

$$H(p) = -\sum_{i=1}^{L} p_i \log p_i$$

- Under certain assumptions, we can define this for continuous spaces, like $\mathcal{X} = \mathbb{R}^d$

$$H(p) = -\int_{\mathcal{X}} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$$

- The maximum entropy principle: choose the "maximally non-committal" distribution regarding missing information, expressing the most uncertainty regarding the missing information [Jaynes 57]

Maximum Entropy Distributions Under Known First and Second Moments

- The maximum entropy principle: most uncertainty regarding missing information
- Applying this to our case, with known first and second moments, let us find the maximum entropy distribution... For simplicity let us solve the case $\mathcal{X} = \mathbb{R}$, mean μ and variance σ^2

Maximum Entropy Distributions Under Known First and Second Moments

Gaussian Distributions

- *Conclusion*: the maximum-entropy distribution, under known first and second moments is a Gaussian!
  - This extends to $\mathbb{R}^d$ by following a similar derivation with $\boldsymbol{\mu} \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}$
- **Note**: The empirical estimates $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ and $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top$ are the Maximum Likelihood Estimates for the parameters of a Gaussian (not a coincidence)
- Additional important properties:
  - Under mild assumptions, any linear combination of distinct random variables $S = X_1 + \ldots + X_N$ is approximately normally distributed (by the Central Limit Theorem)
  - Gaussians are closed under affine transformations (and conditioning):

$$\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}) \rightarrow A\mathbf{x} + \mathbf{b} \sim \mathcal{N}(\mathbf{b}, AA^\top)$$

Sample Complexity of Estimating Principle Components

- Last week we discussed the sample complexity of the multinomial model
- The number of samples needed to achieve error ε scaled with $2^d \varepsilon^{-1}$ (the curse of dimensionality)
- PCA is implemented as follows:
  - Obtain empirical moments $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$, and $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top$
  - Estimate principal components using eigenvectors of $\widehat{\Sigma} = \hat{U} \hat{\Lambda} \hat{U}^\top$

Sample Complexity of Estimating Principle Components

- Assume for simplicity $\hat{\boldsymbol{\mu}} = 0$, how good is the estimate given by $\widehat{\Sigma}$?
- That is, let $\varepsilon > 0$, how many examples do we need to observe to get $\|\Sigma - \widehat{\Sigma}\| \leq \varepsilon$?

Theorem (Vershynin 10)

Consider a random vector $\mathbf{x}$ *in* $\mathbb{R}^n$ *(*$n \geq 4$*) with bounded* q*-th order moment for some* $q > 4$ *(i.e.* $\mathbb{E}\left[\langle \mathbf{v}, \mathbf{x} \rangle^q\right] \leq \infty$ *for all* $\mathbf{v}$*). Let* $\delta > 0$*, then with probability at least* $1 - \delta$*,*

$$\|\Sigma - \widehat{\Sigma}\| \lesssim O\left((\log\log d)^2\right)\left(\frac{d}{n}\right)^{\frac{1}{2}-\frac{2}{q}}$$

- The inequality is up to some multiplicative factor in δ. So to get error ε we need:

Sample Complexity of Estimating Principle Components

- **Conclusion**: PCA does not suffer from the curse of dimensionality!

A Word on Computational Complexity

- Besides sample complexity, we also want algorithms that calculate the components quickly
- Standard implementations that compute the spectral decomposition take $O(k \cdot n \cdot d)$
- Randomized algorithms can improve to $O(\log k \cdot n \cdot d)$

Are Gaussians all you Need?

- We've seen PCA achieving some impressive results, discussed its optimality for linear compression
- We also learned about the underlying learning problem
  - PCA estimates the first two moments of a distribution
  - It is "compatible" with estimating a Gaussian probabilistic model
  - Sample complexity is linear in d, about the best we can hope for!
  - Can be computed very quickly on large datasets and high dimensions
- Should we just use Gaussian models for all our tasks?

The Pitfalls of Linear Models

- Linear models are often not expressive enough, and fail to appropriately describe many types of data
- Example 1 [Shlens 14]

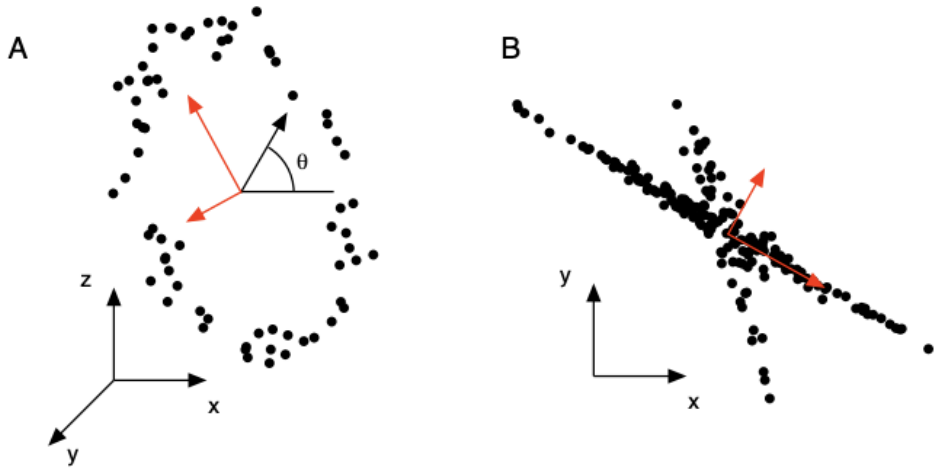

FIG. 6 Example of when PCA fails (red lines). (a) Tracking a person on a ferris wheel (black dots). All dynamics can be described by the phase of the wheel θ, a non-linear combination of the naive basis. (b) In this example data set, non-Gaussian distributed data and non-orthogonal axes causes PCA to fail. The axes with the largest variance do not correspond to the appropriate answer.

The Pitfalls of Linear Models

- Linear models are often not expressive enough, and fail to appropriately describe many types of data
- Example 1 [Shlens 14]
- Example 2, Mixture Models

The Pitfalls of Linear Models

- Linear models are often not expressive enough, and fail to appropriately describe many types of data
- Example 1 [Shlens 14]
- Example 2, Mixture Models
- Which models can be learned and inferred? Next lesson: graphical structures