All you wanted to know about (sparse / high dimensional) PCA and Covariance Estimation

but didn't dare to ask

Boaz Nadler

Department of Computer Science and Applied Mathematics
The Weizmann Institute of Science

Dec. 2021

Part 0: Introduction

Introduction

In many applications we need to analyze multivariate data,

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$$

Throughout talk:

p - dimension, n - number of samples

Introduction

In many applications we need to analyze multivariate data,

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$$

Throughout talk:

p - dimension, n - number of samples

Many different scenarios:

- p small, n large.

- both p and n large.

- p large, n small.

Unsupervised Tasks

- Visualization,

Unsupervised Tasks

- Visualization,

- Dimensionality Reduction

Unsupervised Tasks

- Visualization,

- Dimensionality Reduction

- Compression, Denoising

Unsupervised Tasks

- Visualization,
- Dimensionality Reduction
- Compression, Denoising
- Exploratory Data Analysis - finding structure in data

Unsupervised Tasks

- Visualization,
- Dimensionality Reduction
- Compression, Denoising
- Exploratory Data Analysis - finding structure in data
- Many hypothesis testing problems.

Supervised Tasks

In some cases we observe also a response y_i for each $\mathbf{x}_i$.

Common tasks:

- Classification (Linear Discriminant Analysis)

- Regression (Multivariate Linear Regression).

Density Estimation and Moments

Typical assumption: $\mathbf{x}_i$ are i.i.d. from some r.v. X with unknown density $f(x)$.

Density Estimation and Moments

Typical assumption: $\mathbf{x}_i$ are i.i.d. from some r.v. X with unknown density $f(x)$.

In principle, $f(x)$ describes *everything* about X.

Density Estimation and Moments

Typical assumption: $\mathbf{x}_i$ are i.i.d. from some r.v. X with unknown density $f(x)$.

In principle, $f(x)$ describes *everything* about X.

So, we can try to estimate it.

Density Estimation and Moments

Typical assumption: $\mathbf{x}_i$ are i.i.d. from some r.v. X with unknown density $f(x)$.

In principle, $f(x)$ describes *everything* about X.

So, we can try to estimate it.

Non-parametric (kernel) density estimation:

$$|\hat{f}(x) - f(x)| \sim n^{-2\beta/(2\beta+p)}$$

where β measure of smoothness of $f(x)$.

Density Estimation and Moments

Typical assumption: $\mathbf{x}_i$ are i.i.d. from some r.v. X with unknown density $f(x)$.

In principle, $f(x)$ describes *everything* about X.

So, we can try to estimate it.

Non-parametric (kernel) density estimation:

$$|\hat{f}(x) - f(x)| \sim n^{-2\beta/(2\beta+p)}$$

where β measure of smoothness of $f(x)$.

Curse of dimensionality:
For small error ϵ, need *exponential* in p number of samples, $n = O(\exp(Cp))$.

Density Estimation and Moments

Typical assumption: $\mathbf{x}_i$ are i.i.d. from some r.v. X with unknown density $f(x)$.

In principle, $f(x)$ describes *everything* about X.

So, we can try to estimate it.

Non-parametric (kernel) density estimation:

$$|\hat{f}(x) - f(x)| \sim n^{-2\beta/(2\beta+p)}$$

where β measure of smoothness of $f(x)$.

Curse of dimensionality:
For small error ϵ, need *exponential* in p number of samples, $n = O(\exp(Cp))$.

Typically not practical if $p > 15$

First and Second Moments of X

Since already at moderate dimensions cannot estimate $f(x)$, opt for first and second moments:

Definition: The population mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ is

$$\boldsymbol{\mu} = \mathbb{E}[X] = \int \mathbf{x} f(\mathbf{x}) d\mathbf{x}$$

First and Second Moments of X

Since already at moderate dimensions cannot estimate $f(x)$, opt for first and second moments:

Definition: The population mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ is

$$\boldsymbol{\mu} = \mathbb{E}[X] = \int \mathbf{x} f(\mathbf{x}) d\mathbf{x}$$

Definition: The population covariance matrix $\boldsymbol{\Sigma}_{p \times p}$ is defined as

$$\boldsymbol{\Sigma} = \mathbb{E}[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^T] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T f(\mathbf{x}) d\mathbf{x}$$

In many problems, estimates of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ will suffice to solve a data analysis task.

Example I: Quadratic Discriminant Analysis

Binary (2-class) classification problem:

Example I: Quadratic Discriminant Analysis

Binary (2-class) classification problem:

Observe $\{\mathbf{x}_i\}_{i=1}^{n_+}$ from class $+1$

and

$\{\mathbf{x}'_i\}_{i=1}^{n_-}$ from class -1.

Example I: Quadratic Discriminant Analysis

Binary (2-class) classification problem:

Observe $\{\mathbf{x}_i\}_{i=1}^{n_+}$ from class $+1$

and

$\{\mathbf{x}_i'\}_{i=1}^{n_-}$ from class -1.

Goal: classify label $y \in \{\pm 1\}$ of new $\mathbf{x}$.

Example I: Quadratic Discriminant Analysis

Binary (2-class) classification problem:

Observe $\{\mathbf{x}_i\}_{i=1}^{n_+}$ from class $+1$

and

$\{\mathbf{x}'_i\}_{i=1}^{n_-}$ from class -1.

Goal: classify label $y \in \{\pm 1\}$ of new $\mathbf{x}$.

QDA: Assume each class is multivariate Gaussian,

$$\text{Likellhood Ratio} = C(\Sigma_+, \Sigma_-)\frac{\exp(-(\mathbf{x}-\mu_+)^T\Sigma_+^{-1}(\mathbf{x}-\mu_+)/2)}{\exp(-(\mathbf{x}-\mu_-)^T\Sigma_-^{-1}(\mathbf{x}-\mu_-)/2)}$$

Example I: Quadratic Discriminant Analysis

Binary (2-class) classification problem:

Observe $\{\mathbf{x}_i\}_{i=1}^{n_+}$ from class $+1$

and

$\{\mathbf{x}_i'\}_{i=1}^{n_-}$ from class -1.

Goal: classify label $y \in \{\pm 1\}$ of new $\mathbf{x}$.

QDA: Assume each class is multivariate Gaussian,

$$\text{Likellhood Ratio} = C(\Sigma_+, \Sigma_-) \frac{\exp(-(\mathbf{x} - \mu_+)^T \Sigma_+^{-1} (\mathbf{x} - \mu_+)/2)}{\exp(-(\mathbf{x} - \mu_-)^T \Sigma_-^{-1} (\mathbf{x} - \mu_-)/2)}$$

To apply QDA, need to estimate $\mu_\pm$ and $\Sigma_\pm$ (actually its inverse)

Example II: Markowitz Portfolio Optimization

[Harry Markowitz, Nobel prize 90']

Person can invest in p stocks, with weight vector $\mathbf{w} = (w_1, \ldots, w_p)$, s.t. $\sum w_i = 1$.

Example II: Markowitz Portfolio Optimization

[Harry Markowitz, Nobel prize 90']

Person can invest in p stocks, with weight vector $\mathbf{w} = (w_1, \ldots, w_p)$, s.t. $\sum w_i = 1$.

Stock i has expected return μ_i. All p stocks have joint covariance Σ, where Σ_{ii} is the variance of the i-th stock.

Example II: Markowitz Portfolio Optimization

[Harry Markowitz, Nobel prize 90']

Person can invest in p stocks, with weight vector $\mathbf{w} = (w_1, \ldots, w_p)$, s.t. $\sum w_i = 1$.

Stock i has expected return μ_i. All p stocks have joint covariance Σ, where Σ_{ii} is the variance of the i-th stock.

Goal: Construct a portfolio that achieves an average target return μ_R, namely $\sum_i w_i \mu_i = \mu_R$

Example II: Markowitz Portfolio Optimization

[Harry Markowitz, Nobel prize 90']

Person can invest in p stocks, with weight vector $\mathbf{w} = (w_1, \ldots, w_p)$, s.t. $\sum w_i = 1$.

Stock i has expected return μ_i. All p stocks have joint covariance Σ, where Σ_{ii} is the variance of the i-th stock.

Goal: Construct a portfolio that achieves an average target return μ_R, namely $\sum_i w_i \mu_i = \mu_R$

but

with minimal risk (variance).

$$\min_{\mathbf{w}} \mathbf{w}^T \Sigma \mathbf{w}$$

To construct portfolio, need to estimate μ_i and Σ.

Example III: Dimensionality of Data / Signals in Noise

Analytical Chemistry, Array Signal Processing, ...:

Assume "signal+noise" model

$$\mathbf{x} = \sum_{j=1}^{K} s_j \mathbf{v}_j + noise$$

Given observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, a basic question is: what is K ?

Example III: Dimensionality of Data / Signals in Noise

Analytical Chemistry, Array Signal Processing, ...:

Assume "signal+noise" model

$$\mathbf{x} = \sum_{j=1}^{K} s_j \mathbf{v}_j + \textit{noise}$$

Given observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, a basic question is: what is K ?

The covariance matrix of data has K spikes, eigenvalues larger than noise variance σ^2, and remaining eigenvalues all equal σ^2.

Example III: Dimensionality of Data / Signals in Noise

Analytical Chemistry, Array Signal Processing, ...:

Assume "signal+noise" model

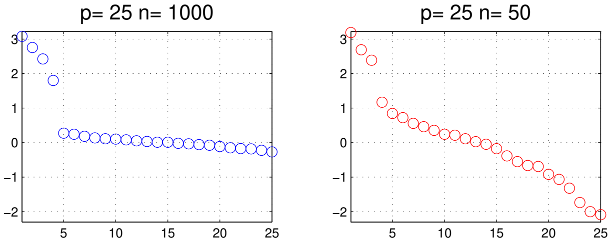
$$\mathbf{x} = \sum_{j=1}^{K} s_j \mathbf{v}_j + \mathit{noise}$$

Given observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, a basic question is: what is K ?

The covariance matrix of data has K spikes, eigenvalues larger than noise variance σ^2, and remaining eigenvalues all equal σ^2.

Question: How to estimate K, which signal strengths can be detected ?

Eigenvalues of S_n (log-scale):



Estimation of Signal Direction

Suppose there is one signal in the data.

$$\mathbf{x} = s\mathbf{v} + \sigma\boldsymbol{\xi}$$

Goal: Estimate signal direction (vector $\mathbf{v}$).

Estimation of Signal Direction

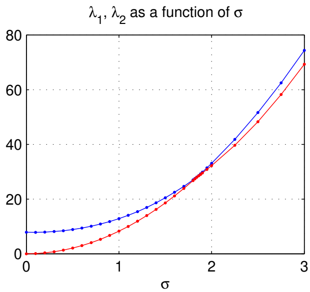
Suppose there is one signal in the data.

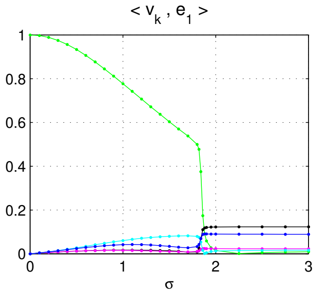
$$\mathbf{x} = s\mathbf{v} + \sigma\boldsymbol{\xi}$$

Goal: Estimate signal direction (vector $\mathbf{v}$).

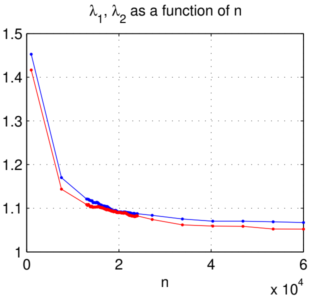
Questions: How should it be estimated ?
How well can it be estimated ?
What happens when dimension p is high ?

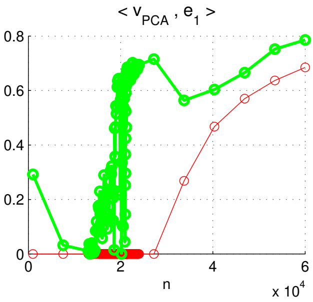
Estimation of Signal Direction





Estimation of Signal Direction





Common Statistical Inference Problems

Given data $\mathbf{x}_1, \ldots, x_n$:

- Estimate population covariance matrix $\mathbf{\Sigma}$

Common Statistical Inference Problems

Given data $\mathbf{x}_1, \ldots, x_n$:

- Estimate population covariance matrix $\mathbf{\Sigma}$

- Estimate the *precision* matrix $\Omega = \mathbf{\Sigma}^{-1}$.

Common Statistical Inference Problems

Given data $\mathbf{x}_1, \ldots, x_n$:

- Estimate population covariance matrix $\mathbf{\Sigma}$
- Estimate the *precision* matrix $\Omega = \mathbf{\Sigma}^{-1}$.
- Estimate the largest eigenvalues and eigenvectors of $\mathbf{\Sigma}$.

Common Statistical Inference Problems

Given data $\mathbf{x}_1, \ldots, x_n$:

- Estimate population covariance matrix $\mathbf{\Sigma}$

- Estimate the *precision* matrix $\Omega = \mathbf{\Sigma}^{-1}$.

- Estimate the largest eigenvalues and eigenvectors of $\mathbf{\Sigma}$.

- Suppose $\mathbf{\Sigma}$ (or $\mathbf{\Sigma}^{-1}$) have many zeros. Find their support.

Common Statistical Inference Problems

Given data $\mathbf{x}_1, \ldots, x_n$:

- Estimate population covariance matrix $\mathbf{\Sigma}$

- Estimate the *precision* matrix $\Omega = \mathbf{\Sigma}^{-1}$.

- Estimate the largest eigenvalues and eigenvectors of $\mathbf{\Sigma}$.

- Suppose $\mathbf{\Sigma}$ (or $\mathbf{\Sigma}^{-1}$) have many zeros. Find their support.

- Many hypothesis testing problems: Is $\Sigma = \Sigma_0$?

- Uncorrelated variables: Is $\Sigma = \textit{diag}$, is Σ banded ?

Inference Problems

Typically, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown and must be estimated from data.

Classical Estimates:

Sample Mean:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}_i$$

Inference Problems

Typically, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown and must be estimated from data.

Classical Estimates:

Sample Mean:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}_i$$

Sample Covariance Matrix:

$$S_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{x})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

Outline

Q-1: How close are S_n and its eigenvalues/vectors to those of Σ ?

Part 1. p small, $n \to \infty$ [classical asymptotic statistics]

Part 2. p, n both large, [modern high dimensional statistics]

Outline

Q-1: How close are S_n and its eigenvalues/vectors to those of Σ ?

Part 1. p small, $n \to \infty$ [classical asymptotic statistics]

Part 2. p, n both large, [modern high dimensional statistics]

Q-2: What if we know additional information: *sparsity*.

Outline

Q-1: How close are S_n and its eigenvalues/vectors to those of Σ ?

Part 1. p small, $n \to \infty$ [classical asymptotic statistics]

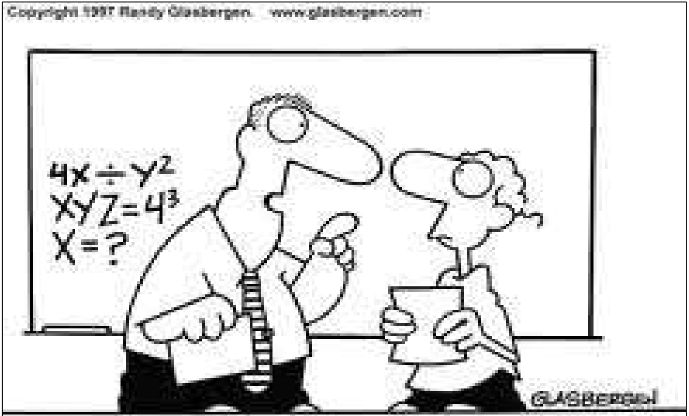
Part 2. p, n both large, [modern high dimensional statistics]

Q-2: What if we know additional information: *sparsity*.

Part 3. Sparse Principal Component Analysis, Sparse Covariance Estimation

The Important Question

Why is this interesting, relevant, important ?



Covariance Estimation will be important to you later in life because there's going to be a test six weeks from now."

References (Partial List)

* Nadler, Finite sample approximation results for PCA, *Annals of Stat.*, 2008.
* Kritchman and Nadler, determining the number of components in a factor model from limited noisy data, 2008.
* Birnbaum et. al., Minimax bounds for sparse PCA with noisy high dimensional data, *Annals of Stat.*, 2013.
* Baik and Silverstein, Eigenvalues of large sample covariance matrices..., *J. Mult. Anal.* 2006.
* Paul, Asymptotics of sample eigenstructure..., *Stat. Sinica*, 2008.
* Bickel and Levina, Covariance regularization by thresholding, *Annals of Stat.*, 2008.
* Cai and Liu, Adaptive thresholding for sparse covariance matrix estimation, J. Am. Stat. Assoc., 2011.
* Cai, Zhang, Zhou, Optimal rates of convergence for covariance matrix estimation, *Ann. of Stat.* 2010.

and many others...