STA 437/2005: Methods for Multivariate Data

Weeks 8: Covariance matrix estimation

Piotr Zwiernik

University of Toronto

Covariance matrix estimation

We started our discussion of PCA on the population level.

- maximizing $\boldsymbol{u}^\top \Sigma \boldsymbol{u}$ gives a direction of the highest variance of $X \in \mathbb{R}^m$.

In practice we have no access to $\Sigma \in \mathbb{S}_+^m$.

The main approach is to estimate Σ using the sample covariance matrix S_n.

Recall that S_n is almost unbiased.

It can be shown that it is a consistent estimator of Σ.

In other words, if n is "very large", S_n should be a good estimator of Σ.

High-dimensional problems

How large n has to be generally depends on m.

This is intuitively clear because Σ has $\binom{m}{2} = \frac{m(m+1)}{2}$ parameters to estimate.

Classical asymptotics lets $n \to \infty$ keeping m fixed.

- Applicable if n is way larger than m.

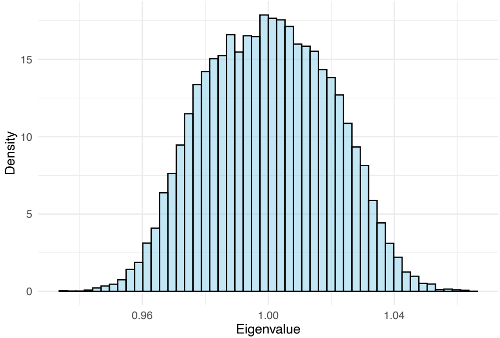
High-dimensional asymptotics studies estimation when both $m, n \to \infty$.

- We assume $m/n \to \gamma \in [0, 1)$.
- May be appplicable in much general contexts.

Why does it matter?

Suppose that $\Sigma = I_m$. If S_n is close to Σ all its eigenvalues should be close to 1.

Consider a simple example: $m = 3$, $n = 1000$. Sample S_n several times and look at the histogram of eigenvalues.



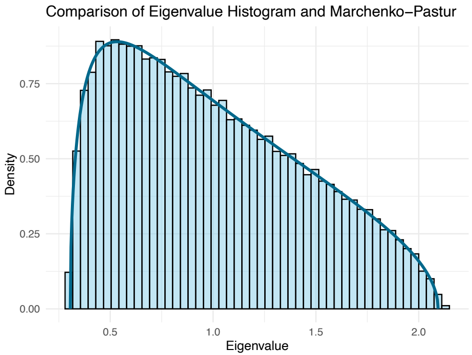
Indeed! A **sharp** concentration around 1.

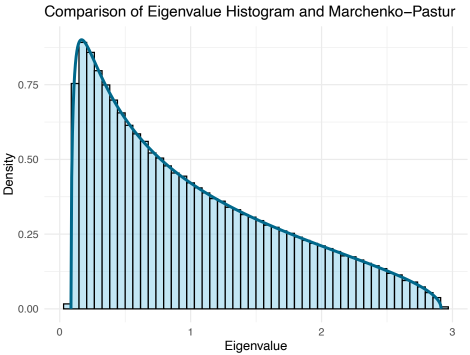
Arguably, this is a very extreme situation.

In typical applications the ration n/m is much smaller.

Consider now the eigenvalue distribution in the same setting but with much higher m.

Take $n = 1000$, $m = 200$ and $m = 500$.





The eigenvalues deviate from 1, following the Marchenko-Pastur law.

Marchenko-Pastur Law

Marchenko-Pastur Law gives the limiting distribution of the eigenvalues of S_n ($\Sigma = I_m$) in the limiting case when $m/n \to \gamma$.

Marchenko-Pastur Law

Let $\lambda_{\text{min}} := (1 - \sqrt{\gamma})^2$, $\lambda_{\text{max}} := (1 + \sqrt{\gamma})^2$. Then MP Law has density

$$f_{\mathrm{MP}}(\lambda) = \frac{1}{2\pi\gamma\lambda}\sqrt{(\lambda_{\text{max}} - \lambda)(\lambda - \lambda_{\text{min}})} \qquad \text{for } \lambda \in [\lambda_{\text{min}}, \lambda_{\text{max}}].$$

Alternative Estimators

This example shows that S_n is not a good estimator of Σ when m/n is too large.

General approach: if there is some additional structure in Σ, exploit it.

- This stabilizes the estimators.

This approach may be problematic if you exploited structure that is not there.

We now review some common approaches that work well in a wide-range of scenarios.

Alternative Estimators Overview

Linear Shrinkage: $\widehat{\Sigma}_{\mathrm{ls}} = (1-\lambda)S_n + \lambda I_m$ for some $\lambda \in (0,1)$.

- Reduces variance by shrinking towards I_m.

Graphical Lasso: We consider penalized Gaussian log-likelihood. Define

$$\widehat{K} \;:=\; \arg\min_{K \in \mathbb{S}_+^m} \{\mathrm{tr}(S_n K) - \log\det(K) + \lambda \|K\|_1\},$$

where $\|K\|_1 = \sum_{i \neq j} |K_{ij}|$ (ℓ_1-penalty). Finally, $\widehat{\Sigma}_{\mathrm{glasso}} = \widehat{K}^{-1}$.

- Promotes sparsity in the precision matrix.

Alternative Estimators Continued

Factor Models: Suppose Σ has the form $\Sigma = WW^\top + \Psi$ where $W \in \mathbb{R}^{m \times r}$ for $r < m$ and Ψ is diagonal.

- We will show how to exploit this in estimation.
- Probabilistic PCA gives one example with $\Psi = \sigma^2 I_m$.

Thresholding-Based Methods: If Σ has zeros, it is natural to estimate

$$\widehat{\Sigma}_{\text{thresh}} = \{(S_n)_{ij} \cdot \mathbb{I}(|(S_n)_{ij}| > \tau)\}_{i,j}.$$

Sets small covariance entries to zero.

Tyler's Scatter Estimator

This is a popular estimator in robust statistics.

If $X \sim E(\mathbf{0}, \Sigma)$ then $Z = \Sigma^{-1/2}X$ is spherical; $Z/\|Z\|$ is uniform on the unit sphere.

Then $\frac{1}{m} I_m = \mathrm{var}(Z/\|Z\|) = \mathbb{E}(\frac{1}{\|Z\|^2} ZZ^\top) = \mathbb{E}(\frac{1}{X^\top \Sigma^{-1} X} \Sigma^{-1/2} XX^\top \Sigma^{-1/2})$.

Equivalently $\mathbb{E}(\frac{1}{X^\top \Sigma^{-1} X} XX^\top) = \frac{1}{m}\Sigma$. Consider a sample version of this equation:

$$\sum_{i=1}^{n} \frac{1}{x^{(i)\top} \Sigma^{-1} x^{(i)}} x^{(i)} x^{(i)\top} = \frac{n}{m} \Sigma.$$

Under mild conditions, there is a unique solution; computed using fixed-point iterations:

$$\hat{\Sigma}^{(k+1)} = \frac{m}{n} \sum_i \frac{x^{(i)} x^{(i)\top}}{x^{(i)\top} (\hat{\Sigma}^{(k)})^{-1} x^{(i)}}.$$

Summary

Estimating the covariance matrix in modern applications raises many challenges.

If Σ satisfies some structure, we could exploit it to stabilize estimation.

We study some structures that can appear in practice.

- Diagonal plus low rank.
- Σ or Σ^{-1} sparse.

This is an active are of research. Links to random matrix theory.