Finance Data Science
Lecture 3: Covariance Matrix Estimation

Laurent El Ghaoui

MFE 230P, Summer 2017
MFE Program
Haas School of Business
UC Berkeley

6/12/2017

Outline

Outline

Motivations and goals

Covariance matrices are widely used in finance:

- Exploratory data analysis (see lecture 4).
- Risk analysis.
- Portfolio optimization.
- Outlier detection.

In practice the number of data points n may be less than the number of dimensions p (assets).

This lecture: examine three estimation methods: one naïve (sample estimate), the other classical (factor model), the last modern.

Eigenvalue decomposition for symmetric matrices

Theorem (EVD of symmetric matrices)

We can decompose any symmetric $p \times p$ matrix S as

$$S = U \Lambda U^T = \sum_{i=1}^{p} \lambda_i u_i u_i^T,$$

where $\Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_p)$, with $\lambda_1 \geq \ldots \geq \lambda_p$ the eigenvalues, and $U = [u_1, \ldots, u_p]$ is a $p \times p$ orthogonal matrix ($U^T U = I_p$) that contains the eigenvectors u_i of S, that is:

$$S u_i = \lambda_i u_i, \quad i = 1, \ldots, p.$$

Corollary: If S is square, symmetric:

$$\lambda_{\max}(S) = \max_{x \,:\, \|x\|_2 = 1} x^T S x. \tag{1}$$

Positive semi-definite (PSD) matrices

A (square) symmetric matrix S is said to be *positive semi-definite* (PSD) if

$$\forall\, x, \quad x^T S x \geq 0.$$

In this case, we write $S \succeq 0$.

From EVD theorem: for any square, symmetric matrix S:

$$S \succeq 0 \iff \text{every eigenvalue of } S \text{ is non-negative.}$$

Hence we can numerically (via EVD) check positive semi-definiteness.

Outline

The sample covariance matrix
Motivation

We can easily define the variance of a collection of numbers $z_1, \ldots, z_m$:

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} (z_i - \hat{z})^2,$$

where $\hat{z} = (1/m)(z_1 + \ldots + z_m)$ is the average of the z_i's.

- How can we extend this notion to higher dimensions (with z_i's as vectors)?
- Why would we want to do that?

Note: for technical reasons the factor $1/m$ is often replaced with $1/(m-1)$, with little effect when m is large.

The sample covariance matrix

Definition

Given a $p \times m$ data matrix $A = [a_1, \ldots, a_m]$ (each row representing say a log-return time-series over m time periods), the *sample covariance matrix* is defined as the $p \times p$ matrix

$$S = \frac{1}{m} \sum_{i=1}^{m} (a_i - \hat{a})(a_i - \hat{a})^T, \quad \hat{a} := \frac{1}{m} \sum_{i=1}^{m} a_i.$$

We can express S as

$$S = \frac{1}{m} A_c A_c^T,$$

where A_c is the *centered data matrix* :

$$A_c = \begin{pmatrix} a_1 - \hat{a} & \ldots & a_m - \hat{a} \end{pmatrix}$$

Note: for technical reasons the factor $1/m$ is often replaced with $1/(m-1)$, with little effect when m is large.

The sample covariance matrix

Link with directional variance

The (sample) variance along direction x is

$$\mathbf{var}(x) = \frac{1}{m}\sum_{i=1}^{m}[x^T(a_i - \hat{a})]^2 = x^T S x = \frac{1}{m}\|A_c x\|_2^2.$$

where A_c is the centered data matrix:

Hence:

- the covariance matrix gives information about variance along any direction, via the quadratic function $x \to x^T Sx$;
- the covariance matrix is always symmetric ($S = S^T$);
- It is also positive-semidefinite (PSD), since $x^T Sx = \mathbf{var}(x) \geq 0$ for every x.

Application: portfolio risk

- *Data:* Consider n assets with returns over one period (*e.g.*, day) r_i, $i = 1, \ldots, n$. In general not known in advance.
- *Portfolio:* described by a vector $x \in \mathbf{R}^n$, with $x_i \geq 0$ the proportion of a total wealth invested in asset i.
- *Portfolio return:* $r^T x$; in general not known.
- *Expected return:* mean value of portfolio return, given by

$$\mathbf{E}\, r^T x = \hat{r}^T x,$$

with $\hat{r} = (\hat{r}_1, \ldots, \hat{r}_n)$ the vector of mean returns.

- *Portfolio risk:* Assuming return vector r is random, with mean $\hat{r}$ and covariance matrix S, the variance of the portfolio is

$$\sigma^2(x) := \mathbf{E}_r (r^T x - \hat{r}^T x)^2 = x^T S x.$$

The sample covariance matrix
Total variance

For a given sample covariance matrix, we define the *total variance* to be the sum of the variances along the unit vectors

$$e_i = (0, \ldots, 1, \ldots, 0) \text{ (with 1 in } i\text{-th position, 0 otherwise).}$$

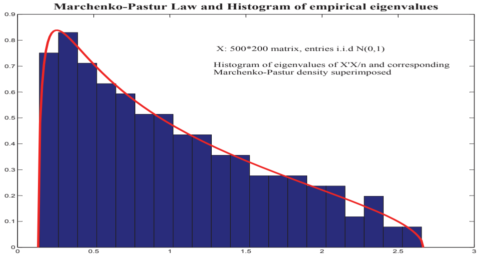
Total variance writes:

$$\sum_{i=1}^{p} \mathbf{var}(e_i) = \sum_{i=1}^{p} e_i^T S e_i = \sum_{i=1}^{p} S_{ii} := \mathbf{Tr}\, S,$$

where the symbol **Tr** (trace) denotes the sum of the diagonal elements of its matrix argument.

What is wrong with the sample covariance?

Assume we draw random data with zero mean and true covariance $S = I_p$, and look at eigenvalues of the sample estimate, when both p, n are large.



Histogram of sample eigenvalues.

- Eigenvalues should be all close to 1!
- This becomes true only when p is fixed and number of samples $n \to +\infty$.
- Red curve shows theoretical result from "random matrix theory" [2], which works for"large p, large n" case (see later).

Estimation problem

In practice, the sample estimate might not work well in high dimensions; so we need to look for better estimates.

Problem: Given data points $x_1, \ldots, x_n \in \mathbf{R}^d$, find an estimate of the covariance $\hat{C}$.

- Many methods start with the sample estimate ...
- ...and remove "noise" from it.

Measuring estimation quality

Cross-validation principle:

- Remove 10 % of data points.
- Record new estimate.
- Measure average "error" between estimates.

How do we measure errors? We need a concept of distance between matrices:

- Frobenius norm (square-root of sum of squares of entries).
- If using a generative model (*e.g.*, Gaussian), we can use Kullback-Leibler divergence (not quite a distance).

Outline

Gaussian assumption

Let us assume that the data points are zero-mean, and follow a multi-variate Gaussian distribution: $x \simeq \mathcal{N}(0, \Sigma)$, with Σ a $p \times p$ covariance matrix. Assume Σ is positive definite.

The Gaussian probability density function is

$$p(\Sigma, x) := \frac{1}{(2\pi \det \Sigma)^{p/2}} \exp((1/2) x^T \Sigma^{-1} x).$$

Maximum-likelihood

How can we find an estimate $\hat{\Sigma}$ of the true Σ, based on data points $x_1, \ldots, x_n$?

Maximum-likelihood principle: maximize the likelihood

$$L(\Sigma) := \prod_{i=1}^{n} p(\Sigma, x_i)$$

over the variable Σ.

Solution

Changing variables ($X := \Sigma^{-1}$), and taking the log of the likelihood, the problem can be written as

$$\max_X \; \log\det X - \mathbf{Tr}\, \hat{C} X$$

where $\hat{C}$ is the sample covariance matrix. In this form, the maximum-likelihood problem is convex.

Solution: $X = \hat{C}^{-1}$, where $\hat{C}$ is the sample covariance matrix!

Caveat: approach fails when $\hat{C}$ is not positive-definite (*e.g.*, when $p > n$!).

Issues

What is wrong with the sample (*i.e.*, ML) estimate?

- Fails in (interesting) case when $p > n$.
- Does not handle missing data.
- High sensitivity to outliers.
- Can come up with better estimates (see next).
- Gaussian assumption is not very good with finance data.

Outline

Data generative model

$$y = Lz + \sigma e$$

- y is the observation (data points).
- e is a noise vector (assume $\mathbf{E}\, e = 0$, $\mathbf{E}\, ee^T = \sigma^2 I$).
- z contains “factors” (assume $\mathbf{E}\, e = 0$, $\mathbf{E}\, ee^T = I$).
- L is a $p \times k$ loading matrix (usually, $k << p$).

This corresponds to a covariance matrix $\Sigma = \sigma^2 I + LL^T$.

Fitting factor models

Given sample covariance matrix $\hat{C} \succeq 0$, we can find L and α buy solving

$$\min_{\alpha \geq 0,\ L} \|\hat{C} - \alpha I - LL^T\|_F.$$

Solution: via EVD of $\hat{C}$.

Scaled version

In practice, we may assume that each random variable has its own noise variance.

Modified problem:

$$\min_{D,L} \|\hat{C} - D - LL^T\|_F \ : \ D \text{ diagonal}, \ D \succeq 0.$$

This time, no obvious solution ...

Can alternate optimization over D (easy) and L (EVD). Results in local optimum.

Computational benefits of factor models

A simple portfolio optimization problem

Risk-return trade-off:

$$\min_x \; f(x) := x^T C x - \lambda r^T x$$

- $r \in \mathbf{R}^p$ (estimate) of returns.
- $x \in \mathbf{R}^p$ portfolio vector (shorting allowed).
- C (estimate of) covariance matrix.
- Parameter $\lambda > 0$ allows to choose trade-off.

The above problem is *convex*.

Assuming $C \succ 0$, optimal point found via $\nabla f(x) = 0$:

$$x^* = \lambda C^{-1} r.$$

Computational benefits of factor models

Direct approach

Assume $C = D + LL^T$, with $D \succ 0$, diagonal, and $F \in \mathbf{R}^{p \times k}$, with $k << p$: we need to solve

$$x^* = (D + LL^T)^{-1} y$$

with $y := \lambda r$.

Direct approach: solve the $p \times p$ linear system

$$(D + LL^T)x = y,$$

without further exploiting structure. Cost: $O(p^3)$.

Computational benefits of factor models

Exploiting structure

Define $z := L^T x$, and rewrite $(D + LL^T)x = y$ as

$$\begin{pmatrix} D & L \\ L^T & -I_k \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}$$

- Eliminate $x = D^{-1}(y - L^T z)$ and get teh $k \times k$ system in z:

$$(I + L^T D^{-1} L)z = D^{-1} Ly.$$

- Then solve for x via $Dx = (y - L^T z)$.

Cost: linear in p!

- Invert diagonal D: $O(p)$.
- Form $I + L^T D^{-1} L$ and solve for z: $O(k^3 + pk^2)$.
- Get x from z: $O(p)$.

The need for shrinkage

- Maximum-likelihood approach fails when $\hat{C} \not\succ 0$ (*e.g.*, $p > n$).
- Well-conditioned estimate is often needed for subsequent use (*e.g.*, portfolio optimization).

  (Condition number of $\hat{C}$ is $\lambda_{\max}(\hat{C})/\lambda_{\min}(\hat{C})$.)

Basic idea: Modify $\hat{C}$ by adding a diagonal, positive-definite term.

Ledoit and Wolf's model [6]

Estimate computed as a convex combination:

$$\hat{\Sigma} = \lambda I + (1 - \lambda)\hat{C},$$

where $\lambda \in (0, 1)$ is a *shrinkage* factor.

- A formula for λ is provided in [7] (has some nice statistical properties).
- Alternatively, choose λ based on cross-validation.
- Can replace the identity with another positive-definite matrix (allows to mix heterogeneous views on markets, such as news-based and price-based).
- Authors show improvements in the context of portfolio optimization.

Outline

What is outlier detection?

Outlier detection problem: Consider a data point $x \in \mathbf{R}^n$. Is it very dissimilar to a data set $X = [x_1, \ldots, x_m]$?

- Arises due to errors in measurement / reporting;
- Also useful prior to running a supervised learning algorithm.
- In practice, we address the problem of ranking possible outliers in a data set (*i.e.*, we solve the above with $x = x_j$, $j = 1, \ldots, m$, and rank the dissimilarity measures.)

- *First idea:* evaluate the distance from the mean $\hat{x}$.
- *Issue:* is the Euclidean norm the "natural" metric to use?
- Many methods are available, including "one-class SVM", more on this later.

In what follows we assume the mean is reset to zero.

Subpsace approach

The (regularized) least-squares objective: ($\lambda > 0$ given)

$$D(x) := \min_{w,b} \|Xw - x\|_2^2 + \lambda \|w\|_2^2.$$

gives an indication of how dissimilar a point x is from the data set X:

- A small value of $D(x)$ indicates that x can almost be expressed as a linear combination of the data points Xw, with small weights w.
- Here $\lambda > 0$ will be a parameter of the outlier detection method.

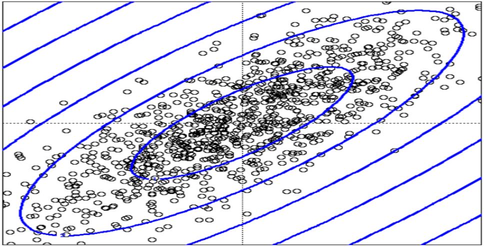
Fact: with X_c the centered data matrix, we have

$$D(x) = x^T (I + (1/\lambda) X_c X_c^T)^{-1} x.$$

Link with Mahalanobis distance

If $C \succ 0$ is a positive-definite covariance matrix, the Mahalanobis distance from a point x and a set of observations with mean $\hat{x}$ is defined as

$$d(x) := (x - \hat{x})^T C^{-1} (x - \hat{x}).$$



The contours of the Mahalanobis distance are ellipsoids.

When $C = \hat{C} + \rho^2 I$ is a regularized estimate, with $\hat{C} = (1/m) X_c X_x^T$ a sample covariance matrix, we recover the previous distance, up to a constant factor (thus, rankings will be the same).

Outline

Motivation

Assume we are given prices corresponding to many assets. We'd like to draw a graph that describes the links between the prices.

- Edges in the graph should exist when some strong, natural metric of similarity exist between assets.
- For better interpretability, a *sparse* graph is desirable.
- Various motivations: portfolio optimization (with sparse risk term), clustering, etc.

Here we focus on exploring *conditional independence* within nodes.

Gaussian assumption

Let us assume that the data points are zero-mean, and follow a multi-variate Gaussian distribution: $x \simeq \mathcal{N}(0, \Sigma)$, with Σ a $p \times p$ covariance matrix. Assume Σ is positive definite.

Gaussian probability density function:

$$p(x) = \frac{1}{(2\pi \det \Sigma)^{p/2}} \exp((1/2)x^T \Sigma^{-1} x).$$

where $X := \Sigma^{-1}$ is the *precision* matrix.

Conditional independence

The pair of random variables x_i, x_j are *conditionally independent* if, for x_k fixed $(k \neq i, j)$, the density can be factored:

$$p(x) = p_i(x_i) p_j(x_j)$$

where p_i, p_j depend also on the other variables.

Conditional independence

The pair of random variables x_i, x_j are *conditionally independent* if, for x_k fixed ($k \neq i, j$), the density can be factored:

$$p(x) = p_i(x_i) p_j(x_j)$$

where p_i, p_j depend also on the other variables.

Interpretation: if all the other variables are fixed then x_i, x_j are independent.

Conditional independence

The pair of random variables x_i, x_j are *conditionally independent* if, for x_k fixed ($k \neq i, j$), the density can be factored:

$$p(x) = p_i(x_i)p_j(x_j)$$

where p_i, p_j depend also on the other variables.

Example: Gray hair and shoe size are independent, conditioned on age.

Conditional independence

C.I. and the precision matrix

Theorem (C.I. for Gaussian RVs)

The variables x_i, x_j are conditionally independent if and only if the i, j element of the precision matrix is zero:

$$(\Sigma^{-1})_{ij} = 0.$$

Proof.

The coefficient of $x_i x_j$ in $\log p(x)$ is $(\Sigma^{-1})_{ij}$. $\blacksquare$

Sparse precision matrix estimation

Let us encourage sparsity of the precision matrix in the maximum-likelihood problem:

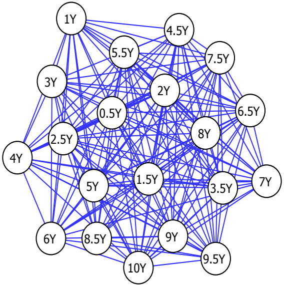
$$\max_X \log\det X - \mathbf{Tr}\,\hat{C}X - \lambda\|X\|_1,$$

with $\|X\|_1 := \sum_{i,j}|X_{ij}|$, and $\lambda > 0$ a parameter.

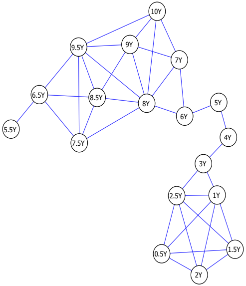
- The above provides an invertible result, even if $\hat{C}$ is not positive-definite.
- The problem is convex.
- The result allows to discover a sparse graph revealing conditional independencies: look pairs (i,j) for which $X_{ij} = 0$.
- Motivations for the use of the l_1-norm: encourages sparsity.

Example

Data: Interest rates



Using covariance matrix ($\lambda = 0$).

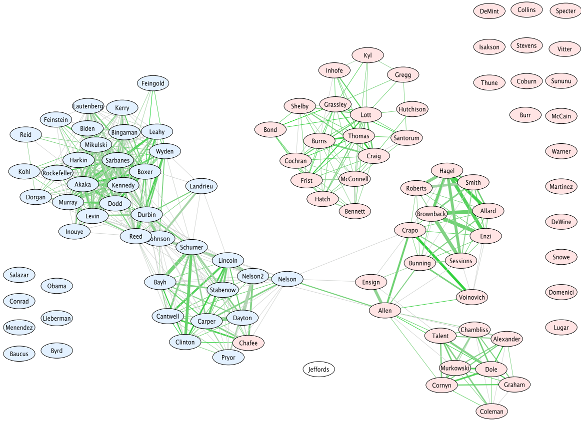


Using $\lambda = 0.1$.

The original precision matrix is dense, but the sparse version reveals the maturity structure (an information that was not given to the algorithm).

Example

Data: US Senate voting, 2002-2004



Again the sparse version reveals information, here political blocks within each party.

Outline

Code

- Python:
  http://scikit-learn.org/stable/modules/covariance.html
  Implements a few methods for covariance estimation, including the sparse inverse covariance estimator.
- R: http://strimmerlab.org/software/corpcor/
  Focuses on a special type of shrinkage estimator (James-Stein)

References I

N. El Karoui.

Spectrum estimation for large dimensional covariance matrices using random matrix theory.

The Annals of Statistics, 36(6):2757–2790, 2008.

N. El Karoui.

High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: Risk underestimation.

The Annals of Statistics, 38(6):3487–3566, 2010.

Jianqing Fan, Yuan Liao, and Han Liu.

An overview of the estimation of large covariance and precision matrices.

The Econometrics Journal, 19(1):C1–C32, 2016.

Online version at `https://arxiv.org/pdf/1504.02995.pdf`.

Jürgen Franke, Wolfgang K. Härdle, and Christian M. Hafner.

Statistics of Financial Markets: An Introduction.

2008.

T. L. Lai and H. Xing.

Statistical models and methods for financial markets.

Springer, 2008.

Olivier Ledoit and Michael Wolf.

Honey, i shrunk the sample covariance matrix.

The Journal of Portfolio Management, 30(4):110–119, 2004.

References II

Olivier Ledoit and Michael Wolf.

A well-conditioned estimator for large-dimensional covariance matrices.

Journal of Multivariate Analysis, 88:365–411, February 2004.

O.Banerjee, L. El Ghaoui, and A. d'Aspremont.

Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data.

Journal of Machine Learning Research, 9:485–516, March 2008.

Adam J Rothman, Elizaveta Levina, and Ji Zhu.

A new approach to cholesky-based covariance regularization in high dimensions.

Biometrika, pages 539–550, 2010.