Title: PCA: Theory, Applications, Extensions

Chapter: PCA: Theory

Section: Covariance matrices

Notation and Definitions

Let X_i, $i = 1, 2, \dots, p$ be square integrable random variables.
Denote: $E(X_i) = \mu_i$, Cov$(X_i, X_j) = \sigma_{ij}$. Note that $\sigma_{ii} = \sigma_i^2$.

We define the expectation vector

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}, \qquad E(\boldsymbol{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} =: \boldsymbol{\mu}$$

and the covariance matrix

$$\mathsf{V}(\boldsymbol{X}) = E((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^t) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}$$

Important note: $\boldsymbol{X}$ is not a vector consisting of sampling replications !

Expectation

The expectation $E(\boldsymbol{X})$ has the following properties:

- $E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y})$
- $E(\lambda \boldsymbol{X}) = \lambda E(\boldsymbol{X})$
- $E(\boldsymbol{AX}) = \boldsymbol{A} E(\boldsymbol{X})$

The third property means:

$$E(a_{i1}X_1 + a_{i2}X_2 + \dots + a_{ip}X_p) = a_{i1}E(X_1) + a_{i2}E(X_2) + \dots + a_{ip}E(X_p)$$

Covariance matrix

The covariance matrix $\mathsf{V}(\boldsymbol{X})$ has the following properties:

- $\mathsf{V}(\lambda \boldsymbol{X}) = \lambda^2 \mathsf{V}(\boldsymbol{X})$
- $\mathsf{V}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\mathsf{V}(\boldsymbol{X})\boldsymbol{A}^t$ for any $m \times p$-matrix
- $\mathsf{V}(\boldsymbol{a}^t\boldsymbol{X}) = \boldsymbol{a}^t\mathsf{V}(\boldsymbol{X})\boldsymbol{a}$ for any $\boldsymbol{a} \in \mathbb{R}^p$

Proof of the second property:

$$\mathsf{V}(\boldsymbol{A}\boldsymbol{X}) = E(\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t\boldsymbol{A}^t) = \boldsymbol{A}E((\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t)\boldsymbol{A}^t = \boldsymbol{A}\mathsf{V}(\boldsymbol{X})\boldsymbol{A}^t$$

The third property means:

$$\mathsf{V}(a_1X_1 + a_2X_2 + \cdots + a_pX_p) = \sum_{i,j=1}^{p} a_i a_j \sigma_{ij}$$

Recall from Linear Algebra:

Lemma: Every matrix of the form $\boldsymbol{M}\boldsymbol{M}^t$ is symmetric and positive semidefinite.

Proof:

$$(\boldsymbol{M}\boldsymbol{M}^t)^t = (\boldsymbol{M}^t)^t\boldsymbol{M}^t = \boldsymbol{M}\boldsymbol{M}^t$$

$$\boldsymbol{a}^t\boldsymbol{M}\boldsymbol{M}^t\boldsymbol{a} = (\boldsymbol{M}^t\boldsymbol{a})^t(\boldsymbol{M}^t\boldsymbol{a}) = ||\boldsymbol{M}^t\boldsymbol{a}||^2 \geq 0$$

Lemma: Every covariance matrix is symmetric and positive semidefinite.

Proof 1: $(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t$ is symmetric and positive definite which carries over to the expectation.

Proof 2: Symmetry follows from $\sigma_{ij} = \mathsf{Cov}(X_i, X_j) = \mathsf{Cov}(X_j, X_i) = \sigma_{ji}$.
Since $\boldsymbol{a}^t\mathsf{V}(\boldsymbol{X})\boldsymbol{a} = \mathsf{V}(\boldsymbol{a}^t\boldsymbol{X}) \geq 0$ for every $\boldsymbol{a} \in \mathbb{R}^p$ it follows that $\mathsf{V}(\boldsymbol{X})$ is positive semidefinite.

Question: What about invertibility of covariance matrices ?
Recall from Linear Algebra:

Lemma: Let Σ be any symmetric positive semidefinite matrix. Then Σ is invertible iff it is positive definite.

Proof:
If Σ is not invertible then there is some $\boldsymbol{a} \neq \boldsymbol{o}$ such that $\Sigma\boldsymbol{a} = \boldsymbol{o}$. Then $\boldsymbol{a}^t\Sigma\boldsymbol{a} = 0$ and thus Σ is not positive definite.
If Σ is not positive definite then there is some eigenvalue $\lambda = 0$. Let $\boldsymbol{a} \neq \boldsymbol{o}$ be a corresponding eigenvector. Then $\Sigma\boldsymbol{a} = \lambda\boldsymbol{a} = \boldsymbol{o}$. Hence Σ is not invertible.

Invertibility of the covariance matrix

A linear relation between $X_1, X_2, \dots, X_p$ is an equation

$$P(a_1X_1 + a_2X_2 + \cdots + a_pX_p = b) = 1 \quad \Leftrightarrow \quad \mathsf{V}(\boldsymbol{a}^t\boldsymbol{X}) = 0 \quad \Leftrightarrow \quad \boldsymbol{a}^t\mathsf{V}(\boldsymbol{X})\boldsymbol{a} = 0$$

where $\boldsymbol{a} \neq \boldsymbol{o}$.

Theorem:
If the covariance matrix $\mathsf{V}(\boldsymbol{X})$ is invertible/positive definite then there is no linear relation between $X_1, X_2, \dots, X_p$.
If the covariance matrix $\mathsf{V}(\boldsymbol{X})$ is not invertible/positive definite then there exists a linear relation between $X_1, X_2, \dots, X_p$.

Proof: Immediate consequence of the lemma.

Exercise: Let $\boldsymbol{X} = (X_1, X_2)^t$ be square integrable. Show that $\mathsf{V}(\boldsymbol{X})$ is invertible iff $|\rho| < 1$. Give a linear relation between X_1 and X_2 in case $|\rho| = 1$.

Solution:

$$\det \mathsf{V}(\boldsymbol{X}) = \det \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = \sigma_1^2\sigma_2^2 - \sigma_{12}^2 = \sigma_1^2\sigma_2^2(1 - \rho^2)$$

Assume that $\sigma_1^2 \neq 0$. If $|\rho| = 1$ then

$$X_2 = \hat{X}_2 = \frac{\sigma_2}{\sigma_1}X_1 + \left(\mu_2 - \frac{\sigma_2}{\sigma_1}\mu_1\right)$$

This implies

$$\sigma_1X_2 - \sigma_2X_1 = \sigma_1\mu_2 - \sigma_2\mu_1$$

Standardized vectors

A square integrable random vector $\boldsymbol{X}$ is called centered if $E(\boldsymbol{X}) = \boldsymbol{o}$.

Centering: Let $E(\boldsymbol{X}) = \boldsymbol{\mu}$. Then $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{\mu}$ is centered.

A square integrable random vector $\boldsymbol{X}$ with invertible covariance matrix is called standardized if $E(\boldsymbol{X}) = \boldsymbol{o}$ and $\mathsf{V}(\boldsymbol{X}) = \boldsymbol{E}$.

Standardization: Let $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathsf{V}(\boldsymbol{X}) = \Sigma$. If there is some matrix $\boldsymbol{B}$ such that $\boldsymbol{B}\Sigma\boldsymbol{B}^t = \boldsymbol{E}$ then then $\boldsymbol{Y} = \boldsymbol{B}(\boldsymbol{X} - \boldsymbol{\mu})$ is standardized.

How to obtain a matrix $\boldsymbol{B}$ such that $\boldsymbol{B}\Sigma\boldsymbol{B}^t = \boldsymbol{E}$?

Theorem: Let Σ be a symmetric positive definite matrix. Then there exist invertible matrices $\boldsymbol{A}$ such that $\Sigma = \boldsymbol{A}\boldsymbol{A}^t$.

Consequence: Let $\boldsymbol{B} = \boldsymbol{A}^{-1}$. Then $\boldsymbol{B}\Sigma\boldsymbol{B}^t = \boldsymbol{B}(\boldsymbol{A}\boldsymbol{A}^t)\boldsymbol{B}^t = (\boldsymbol{B}\boldsymbol{A})(\boldsymbol{A}^t\boldsymbol{B}^t) = \boldsymbol{E}$

The representation $\Sigma = \boldsymbol{A}\boldsymbol{A}^t$ is not unique !
There are infinitely many transformations $\boldsymbol{B}$ leading to standardizations.

However: (details later)

Cholesky decomposition: There exists a uniquely determined lower triangle matrix $\boldsymbol{L}$ with positive diagonal such that $\Sigma = \boldsymbol{L}\boldsymbol{L}^t$.

Principal components: Diagonalization along eigenvectors and eigenvalues of Σ.

Square root: There exists a uniquely determined symmetric positive definite matrix $\boldsymbol{A}$ such that $\Sigma = \boldsymbol{A}\boldsymbol{A}^t$.

Correlation matrix

Let $\boldsymbol{X}$ be a square integrable random vector with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and Cov$(\boldsymbol{X}) = \Sigma$. Assume that $\sigma_i > 0$ for all $i = 1, 2, \ldots, n$.

Let $Y_i := \dfrac{X_i - \mu_i}{\sigma_i}$ the standardized components of the vector $\boldsymbol{X}$.
Then the vector $\boldsymbol{Y}$ is not necessarily standardized but may have non-vanishing covariances:

$$\text{Cov}(Y_i, Y_j) = \text{Cov}\Big(\frac{X_i - \mu_i}{\sigma_i}\frac{X_j - \mu_j}{\sigma_j}\Big) = \rho_{ij} = \text{Cor}(X_i, X_j)$$

Correlation matrix:

$$\boldsymbol{\Sigma}_0 := \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1n} \\ \rho_{12} & 1 & \ldots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1n} & \rho_{2n} & \ldots & 1 \end{pmatrix}$$

Chapter: PCA: Theory

Section: Multiple linear regression

Best linear predictors

How to explain X_p by a linear function of $X_1, X_2, \ldots, X_{p-1}$?

Formal statement: Find $\beta_0, \beta_1, \ldots, \beta_{p-1}$ such that

$$E([X_p - \beta_0 - \beta_1 X_1 - \cdots - \beta_{p-1} X_{p-1}]^2) = \text{Min !}$$

The optimal solution is called the best linear predictor of X_p
with respect to $X_1, X_2, \ldots, X_{p-1}$:

$$L(X_p | X_1, X_2, \ldots \;, X_{p-1}) = \beta_0 + \beta_1 X_1 + \cdots \; + \beta_{p-1} X_{p\text{-}1}$$

Calculation of the best linear predictor

Let $\hat{X}_p := \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$. W.l.g we may assume that

$$\hat{X}_p = \mu_p + \beta_1 (X_1 - \mu_1) + \cdots + \beta_{p-1}(X_{p-1} - \mu_{p-1})$$

Then

$$E([X_p - \hat{X}_p]^2) = \sigma_p^2 - 2\text{Cov}(X_p, \hat{X}_p) + \text{V}(\hat{X}_p)$$

Denoting

$$\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{p-1} & \boldsymbol{\gamma} \\ \hline \boldsymbol{\gamma} & \sigma_p^2 \end{array} \right)$$

we have

$$\text{Cov}(X_p, \hat{X}_p) = \beta_1 \sigma_{1p} + \cdots + \beta_{p-1} \sigma_{p-1,p} = \boldsymbol{\gamma}^t \boldsymbol{\beta}$$

and

$$\text{V}(\hat{X}_p) = \sum_{i,j=1}^{p-1} \beta_i \beta_j \sigma_{ij} = \boldsymbol{\beta}^t \boldsymbol{\Sigma}_{p-1} \boldsymbol{\beta}$$

It follows that

$$E([X_p - \hat{X}_p]^2) = \sigma_p^2 - 2\boldsymbol{\gamma}^t\boldsymbol{\beta} + \boldsymbol{\beta}^t\boldsymbol{\Sigma}_{p-1}\boldsymbol{\beta}$$

This is a convex quadratic function having a global minimum at any solution of

$$\boldsymbol{\Sigma}_{p-1}\boldsymbol{\beta} = \boldsymbol{\gamma}$$

This is the system of normal equations for multiple linear regression.

The system of normal equations has a unique solution if $\boldsymbol{\Sigma}_{p-1}$ is invertible, i.e. if there is no linear relation between $X_1, X_2, \ldots, X_{p-1}$.

But: The system of normal equations has always a solution.
Proof: We apply the nullation lemma.
If $\boldsymbol{a}$ is such that $\boldsymbol{a}^t\boldsymbol{\Sigma}_{p-1} = \boldsymbol{o}$ then $\boldsymbol{a}^t\boldsymbol{\Sigma}_{p-1}\boldsymbol{a} = \boldsymbol{o}$. This implies

$$\mathsf{V}\Big(\sum_{i=1}^{p-1} a_i X_i\Big) = 0 \;\Rightarrow\; \boldsymbol{a}^t\gamma = \mathsf{Cov}\Big(\sum_{i=1}^{p-1} a_i X_i, X_p\Big) = 0$$

Theorem:
Let $Z = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1}X_{p-1}$ be some linear function of $X_1, X_2, \ldots, X_{p-1}$.
Then $Z = L(X_p | X_1, X_2, \ldots, X_{p-1})$ iff the residual $\epsilon := X_p - Z$ satisfies the following conditions:
1. $E(\epsilon) = 0$
2. Cov$(\epsilon, X_i) = 0$ for $i = 1, 2, \ldots, p-1$

Proof: If condition 1 is not the case then $E(\epsilon^2)$ can be decreased by centering, which would contradict the minimality of $E(\epsilon^2)$.
For condition 2 note that

$$\boldsymbol{\Sigma}_{p-1}\boldsymbol{\beta} = \begin{pmatrix} \mathsf{Cov}(Z, X_1) \\ \vdots \\ \mathsf{Cov}(Z, X_{p-1}) \end{pmatrix} \quad \text{and} \quad \gamma = \begin{pmatrix} \mathsf{Cov}(X_p, X_1) \\ \vdots \\ \mathsf{Cov}(X_p, X_{p-1}) \end{pmatrix}$$

Therefore we have

$$\boldsymbol{\Sigma}_{p-1}\boldsymbol{\beta} = \gamma \;\Leftrightarrow\; \mathsf{Cov}(\epsilon, X_i) = 0 \text{ for } i = 1, 2, \ldots, p-1$$

Exercise: Let $\hat{X}_p$ be the best linear predictor of X_p w.r.t. $X_1, \ldots, X_{p-1}$ and let $\rho = \mathsf{Cor}(X_p, \hat{X}_p)$. Show that

$$\mathsf{V}(\epsilon) = \mathsf{V}(X_p - \hat{X}_p) = \mathsf{V}(X_p)(1 - \rho^2)$$

Solution: First, note that

$$X_p = \hat{X}_p + (X_p - \hat{X}_p) \ \Rightarrow \ \mathsf{Cov}(\hat{X}_p, X_p) = \mathsf{V}(\hat{X}_p)$$

which implies

$$\rho^2 = \frac{\mathsf{Cov}(\hat{X}_p, X_p)^2}{\mathsf{V}(X_p)\mathsf{V}(\hat{X}_p)} = \frac{\mathsf{V}(\hat{X}_p)}{\mathsf{V}(X_p)} \ \Rightarrow \ \mathsf{V}(\hat{X}_p) = \mathsf{V}(X_p)\rho^2$$

Now, we get the ANOVA-equation (variance decomposition):

$$\mathsf{V}(X_p) = \mathsf{V}(\hat{X}_p) + \mathsf{V}(X_p - \hat{X}_p) \ \Rightarrow \ \mathsf{V}(X_p - \hat{X}_p) = \mathsf{V}(X_p)(1 - \rho^2)$$

Interpretation: ρ^2 is called "R-square". It measures the percentage of the variance of X_p which can be "explained" by the linear predictor.

Sampling the multiple regression model

Let $X_1, X_2, \ldots, X_p$ be square integrable random variables with usual notations and let $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_p$ be the sampling replications.

LSQ-principle: Find $b_0, b_1, \ldots, b_{p-1}$ and such that

$$\sum_{i=1}^{n}(X_{ip} - b_0 - b_1 X_{i1} - \cdots - b_{p-1} X_{i,p-1})^2 = \text{Min !}$$

Solution: Let $\boldsymbol{D} = \begin{pmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \ldots & X_{2,p-1} \\ \ldots & & & & \\ 1 & X_{n1} & X_{n2} & \ldots & X_{n,p-1} \end{pmatrix}$ (design matrix).

Then the LSQ-principle means that

$$||\underline{X}_p - \boldsymbol{Db}||^2 = \text{Min !}$$

This is the definition of an orthogonal projection with the solution

$$\boldsymbol{b}^* = (\boldsymbol{D}^t \boldsymbol{D})^{-1} \boldsymbol{D}^t \underline{X}_p$$

Chapter: PCA: Theory

Section: Cholesky decomposition

The 2×2-case

Let X_1 and X_2 be square integrable with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and V$(\boldsymbol{X}) = \boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ is positive definite.

We want to find numbers a, b, c such that

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix} \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} = \begin{pmatrix} a^2 & ab \\ ab & b^2 + c^2 \end{pmatrix}$$

By easy computations . . .

$$\begin{aligned} \sigma_{11} &= a^2 & &\Rightarrow a = \sqrt{\sigma_{11}} = \sigma_1 > 0 \\ \sigma_{12} &= ab & &\Rightarrow b = \sigma_{12}/\sigma_1 = \rho\sigma_2 \\ \sigma_{22} &= b^2 + c^2 & &\Rightarrow c = \sqrt{\sigma_2^2 - \sigma_{12}^2/\sigma_1^2} = \sigma_2\sqrt{1-\rho^2} \end{aligned}$$

Important note: c is a real number iff $\boldsymbol{\Sigma}$ is positive semidefinite.
$c > 0$ iff $\boldsymbol{\Sigma}$ is positive definite.

We have

$$\boldsymbol{L} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix} \quad \text{and} \quad \boldsymbol{L}\boldsymbol{L}^t = \boldsymbol{\Sigma}$$

and

$$\boldsymbol{L}^{-1} = \frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}} \begin{pmatrix} \sigma_2\sqrt{1-\rho^2} & 0 \\ -\rho\sigma_2 & \sigma_1 \end{pmatrix}$$

Note, that the diagonals of $\boldsymbol{L}$ and $\boldsymbol{L}^{-1}$ are positive and that $\boldsymbol{L}^{-1}$ is a lower triangle matrix, too.

Cholesky decomposition

Theorem: Let Σ be any symmetric matrix.

If Σ is positive semidefinite then there exists a uniquely determined lower triangle matrix $\boldsymbol{L}$ with nonnegative diagonal such that $\Sigma = \boldsymbol{L}\boldsymbol{L}^t$.

If Σ is positive definite then there exists a uniquely determined lower triangle matrix $\boldsymbol{L}$ with positive diagonal such that $\Sigma = \boldsymbol{L}\boldsymbol{L}^t$.

There is an efficient algorithm for obtaining the Cholesky decomposition which is part of every good software package dealing with linear algebra.

Corollary: An upper (lower) triangle matrix is invertible iff its diagonal contains no zeros, and the inverse is an upper (lower) triangle matrix, too.

Proof: Try to invert an upper triangle matrix by Gauss-Jordan elimination.

Standardization

Let $\boldsymbol{X}$ be a random vector with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathsf{V}(\boldsymbol{X}) = \Sigma$ where Σ is positive definite. Let $\Sigma = \boldsymbol{L}\boldsymbol{L}^t$ be the Cholesky decomposition of the covariance matrix.

Then $\boldsymbol{L}$ is invertible and we obtain a standardization of $\boldsymbol{X}$ by

$$\boldsymbol{Y} := \boldsymbol{L}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$$

This special standardization has a very peculiar structure:

Theorem: The components of $\boldsymbol{Y} := \boldsymbol{L}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ are proportional to the residuals

$$\epsilon_i = X_i - L(X_i | X_1, X_2, \ldots, X_{i-1})$$

of best linear predictors, i.e.

$$Y_i = \frac{\epsilon_i}{\sqrt{\mathsf{V}(\epsilon_i)}}$$

Actually, calculating the Cholesky decomposition is the most efficient way to calculate best linear predictors.

Illustration for the 2×2-case:

$$\boldsymbol{L}^{-1} = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix}$$

We multiply by constants such that the diagonal coefficients become 1:

$$\begin{aligned} Y_1 &= a(X_1 - \mu_1) && \Big| \quad \cdot 1/a \\ Y_2 &= b(X_1 - \mu_1) + c(X_2 - \mu_2) && \Big| \quad \cdot 1/c \end{aligned}$$

and obtain

$$\begin{aligned} \epsilon_1 &:= X_1 - \mu_1 \\ \epsilon_2 &:= \frac{b}{c}(X_1 - \mu_1) + (X_2 - \mu_2) = X_2 - \Big(\mu_2 + \frac{b}{c}(X_1 - \mu_1)\Big) \end{aligned}$$

Since Y_1 and Y_2 are uncorrelated, ϵ_1 and ϵ_2 are uncorrelated, too. Therefore, ϵ_2 is uncorrelated to X_1 !

The term within the bracket is necessarily $L(X_2|X_1)$!

Illustration for the 2×2-case: (making things explicit for those who don't believe)

$$\boldsymbol{L}^{-1} = \frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}} \begin{pmatrix} \sigma_2\sqrt{1-\rho^2} & 0 \\ -\rho\sigma_2 & \sigma_1 \end{pmatrix} = \begin{pmatrix} 1/\sigma_1 & 0 \\ -\rho/(\sigma_1\sqrt{1-\rho^2}) & 1/(\sigma_2\sqrt{1-\rho^2}) \end{pmatrix}$$

We multiply by constants such that the diagonal coefficients become 1:

$$\begin{aligned} Y_1 &= \frac{X_1 - \mu_1}{\sigma_1} && \Big| \quad \cdot \sigma_1 \\ Y_2 &= \frac{1}{\sqrt{1-\rho^2}}\Big(-\rho\frac{X_1 - \mu_1}{\sigma_1} + \frac{X_2 - \mu_2}{\sigma_2}\Big) && \Big| \quad \cdot \sigma_2\sqrt{1-\rho^2} \end{aligned}$$

and obtain

$$\begin{aligned} \epsilon_1 &:= X_1 - \mu_1 \\ \epsilon_2 &:= -\rho\frac{\sigma_2}{\sigma_1}(X_1 - \mu_1) + (X_2 - \mu_2) = X_2 - \Big(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X_1 - \mu_1)\Big) \end{aligned}$$

Since Y_1 and Y_2 are uncorrelated, ϵ_1 and ϵ_2 are uncorrelated, too. Therefore, ϵ_2 is uncorrelated to X_1 !

The term within the bracket is necessarily $L(X_2|X_1)$!

Exercise: Let $\boldsymbol{X} = (X_1, X_2)^t$ be such that $E(\boldsymbol{X}) = \boldsymbol{o}$ and $\sigma_1^2 = 2$, $\sigma_2^2 = 3$, $\sigma_{12} = 1$. Use the Cholesky decomposition to find the linear predictor of X_2 w.r.t. X_1.

Solution: We want to find numbers a, b, c such that

$$\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix} \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} = \begin{pmatrix} a^2 & ab \\ ab & b^2 + c^2 \end{pmatrix}$$

By easy computations we get $a^2 = 2 \;\Rightarrow\; a = \sqrt{2}$, $ab = 1 \;\Rightarrow\; b = 1/\sqrt{2}$, $b^2 + c^2 = 3 \;\Rightarrow\; c = \sqrt{5/2}$. Thus we obtain

$$\boldsymbol{L} = \begin{pmatrix} \sqrt{2} & 0 \\ 1/\sqrt{2} & \sqrt{5/2} \end{pmatrix} \quad \text{and} \quad \boldsymbol{L}^{-1} = \frac{1}{\sqrt{5}} \begin{pmatrix} \sqrt{5/2} & 0 \\ -1/\sqrt{2} & \sqrt{2} \end{pmatrix}$$

We arrive at

$$\begin{aligned} Y_1 &= \frac{1}{\sqrt{2}} X_1 & \qquad \epsilon_1 &= X_1 \\ Y_2 &= \frac{1}{\sqrt{5}} \Big(-\frac{1}{\sqrt{2}} X_1 + \sqrt{2} X_2 \Big) & \qquad \epsilon_2 &= -\frac{1}{2} X_1 + X_2 \end{aligned}$$

The linear predictor is $\hat{X}_2 = \dfrac{1}{2} X_1$.

Exercise: The random vector $\boldsymbol{X} = (X_1, X_2, X_3)$ has mean $\boldsymbol{\mu}^t = (3, -1, 2)$. For the covariance matrix it is known that

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1/3 & 0 \\ 1/3 & 2/9 & 1/6 \\ 0 & 1/6 & 1/2 \end{pmatrix} = \boldsymbol{L}\boldsymbol{L}^t \quad \text{and} \quad \boldsymbol{L}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 3 & 0 \\ 1 & -3 & 2 \end{pmatrix}$$

Find $L(X_2|X_1)$ and $L(X_3|X_1, X_2)$ and the corresponding values of R-square.

Solution: Standardization by $\boldsymbol{L}^{-1}$ gives

$$\begin{aligned} Y_1 &= (X_1 - 3) & &= X_1 - 3 \\ Y_2 &= -(X_1 - 3) + 3(X_2 + 1) & &= -X_1 + 3X_2 + 6 \\ Y_3 &= (X_1 - 3) - 3(X_2 + 1) + 2(X_3 - 2) & &= X_1 - 3X_2 + 2X_3 - 10 \end{aligned}$$

Dividing the second line by 3 and the third line by 2 we get

$$\begin{aligned} \epsilon_2 &= -\tfrac{1}{3} X_1 + X_2 + 2 & \Rightarrow X_2 &= \tfrac{1}{3} X_1 - 2 + \epsilon_2 \\ \epsilon_3 &= \tfrac{1}{2} X_1 - \tfrac{3}{2} X_2 + X_3 - 5 & \Rightarrow X_3 &= -\tfrac{1}{2} X_1 + \tfrac{3}{2} X_2 + 5 + \epsilon_3 \end{aligned}$$

It follows that $L(X_2|X_1) = \frac{1}{3} X_1 - 2$ and $L(X_3|X_1, X_2) = -\frac{1}{2} X_1 + \frac{3}{2} X_2 + 5$.

Finding R-square for $L(X_2|X_1) = \hat{X}_2$:
We have

$$\epsilon_2 = \frac{1}{3}Y_2, \quad \mathsf{V}(Y_2) = 1 \;\Rightarrow\; \mathsf{V}(\epsilon_2) = \frac{1}{9}$$

which implies

$$\rho^2 = 1 - \frac{\mathsf{V}(\epsilon_2)}{\mathsf{V}(X_2)} = \frac{1}{2}$$

Finding R-square for $L(X_3|X_1, X_2) = \hat{X}_3$:
We have

$$\epsilon_3 = \frac{1}{2}Y_3, \quad \mathsf{V}(Y_3) = 1 \;\Rightarrow\; \mathsf{V}(\epsilon_3) = \frac{1}{4}$$

which implies

$$\rho^2 = 1 - \frac{\mathsf{V}(\epsilon_3)}{\mathsf{V}(X_3)} = \frac{1}{2}$$

The general case

There is a lower triangle matrix $\boldsymbol{L}$ with positive diagonal such that $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^t$.

It is easy to see that the inverse $\boldsymbol{L}^{-1}$ is also a lower triangle matrix with positive diagonal:

$$\boldsymbol{L}^{-1} = \begin{pmatrix} \gamma_{11} & 0 & 0 & \dots & 0 \\ \gamma_{21} & \gamma_{22} & 0 & \dots & 0 \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{p1} & \gamma_{p2} & \gamma_{p3} & \dots & \gamma_{pp} \end{pmatrix}$$

We define $\boldsymbol{Y} = \boldsymbol{L}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$, i.e.

$$\begin{aligned} Y_1 =& \gamma_{11}(X_1 - \mu_1) \\ Y_2 =& \gamma_{21}(X_1 - \mu_1) + \gamma_{22}(X_2 - \mu_2) \\ &\dots\dots \\ Y_n =& \gamma_{p1}(X_1 - \mu_1) + \gamma_{p2}(X_2 - \mu_2) + \dots + \gamma_{pp}(X_p - \mu_p) \end{aligned}$$

The variables $Y_1, Y_2, \dots, Y_p$ are uncorrelated and standardized.

We divide row 1 by γ_{11}, row 2 by γ_{22} and so on:

$$\begin{aligned}\epsilon_1 =&(X_1-\mu_1)\\ \epsilon_2 =&(X_2-\mu_2)-\beta_{21}(X_1-\mu_1)\\ &\ldots\ldots\\ \epsilon_p =&(X_p-\mu_p)-\beta_{p1}(X_1-\mu_1)-\beta_{p2}(X_2-\mu_2)-\cdots-\beta_{p,p-1}(X_{p-1}-\mu_{p-1})\end{aligned}$$

This gives

$$\begin{aligned}X_1 =&\mu_1+\epsilon_1\\ X_2 =&\mu_2+\beta_{21}(X_1-\mu_1)+\epsilon_2\\ X_3 =&\mu_3+\beta_{31}(X_1-\mu_1)+\beta_{32}(X_2-\mu_2)+\epsilon_3\\ &\ldots\ldots\\ X_p =&\mu_p+\beta_{p1}(X_1-\mu_1)+\beta_{p2}(X_2-\mu_2)+\cdots+\beta_{p,p-1}(X_{p-1}-\mu_{p-1})+\epsilon_p\end{aligned}$$

ϵ_2 is centered and uncorrelated to ϵ_1, hence to X_1:

$$X_2 = \underbrace{\mu_2+\beta_{21}(X_1-\mu_1)}_{L(X_2|X_1)}+\epsilon_2$$

ϵ_3 is centered and uncorrelated to ϵ_1 and ϵ_2, hence to X_1 and X_2:

$$X_3 = \underbrace{\mu_3+\beta_{31}(X_1-\mu_1)+\beta_{32}(X_2-\mu_2)}_{L(X_3|X_1,X_2)}+\epsilon_3$$

...

ϵ_p is centered and uncorrelated to $\epsilon_1, \ldots \epsilon_{p-1}$, hence to $X_1, \ldots, X_{p-1}$:

$$X_p = \underbrace{\mu_p+\beta_{p1}(X_1-\mu_1)+\beta_{p2}(X_2-\mu_2)+\cdots+\beta_{p,p-1}(X_{p-1}-\mu_{p-1})}_{L(X_p|X_1,X_2,\ldots,X_{p-1})}+\epsilon_p$$

Exercise: Calculate the Cholesky decomposition of a 2×2-correlation matrix.

Solution: Let $\mathbf{\Sigma}_0 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Then with the notation used before we have

$$\begin{aligned} 1 &= a^2 & &\Rightarrow a = 1 \\ \rho &= ab & &\Rightarrow b = \rho \\ 1 &= b^2 + c^2 & &\Rightarrow c = \sqrt{1-\rho^2} \end{aligned}$$

This gives

$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \quad \text{and} \quad \boldsymbol{L}^{-1} = \begin{pmatrix} 1 & 0 \\ -\frac{\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} \end{pmatrix}$$

Chapter: PCA: Theory

Section: Principal components

Diagonalization and standardization

Let $\boldsymbol{X}$ be a square integrable random vector with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and Cov$(\boldsymbol{X}) = \Sigma$.
Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ be the eigenvalues of Σ and $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_p$ the corresponding normed eigenvectors.
Let $\boldsymbol{T}$ be the orthogonal matrix with columns $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_p$ and let $\boldsymbol{D}$ be the diagonal matrix with diagonal $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

We know: $\boldsymbol{T}^t \Sigma \boldsymbol{T} = \boldsymbol{D}$ and $\boldsymbol{TDT}^t = \Sigma$. Defining $\boldsymbol{A} := \boldsymbol{TD}^{1/2}$ we get $\Sigma = \boldsymbol{AA}^t$.
If Σ is invertible then we have $\boldsymbol{A}^{-1} = \boldsymbol{D}^{-1/2}\boldsymbol{T}^t$ and we may use $\boldsymbol{A}^{-1}$ for standardization of $\boldsymbol{X}$.

Let us have a closer look at the structure of $\boldsymbol{A}^{-1}$!

Principal components

Let $\boldsymbol{Z} = \boldsymbol{A}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{D}^{-1/2}\boldsymbol{T}^t(\boldsymbol{X} - \boldsymbol{\mu})$. Then

$$Z_k = \frac{1}{\sqrt{\lambda_k}} \boldsymbol{b}_k^t (\boldsymbol{X} - \boldsymbol{\mu}) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{p} b_{ik}(X_i - \mu_i)$$

Definition: The variables Z_k, $k = 1, 2, \ldots, p$ are called the principal components of $\boldsymbol{X}$.

The principal components $\boldsymbol{Z}$ are a special standardization of $\boldsymbol{X}$.
Where does the name come from ? Why are the eigenvectors of the covariance matrix called principal ?

Dimensionality reduction

Let $\boldsymbol{X}$ be a square integrable random vector with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and Cov($\boldsymbol{X}$) $= \Sigma$.

Goal: We want to find a linear combination of the components of $\boldsymbol{X}$ that tells us as much as possible about the vector $\boldsymbol{X}$.

Direction (axis) in $\mathbb{R}^p$: $\boldsymbol{a} \in \mathbb{R}^p$, $||\boldsymbol{a}|| = 1$
Linear combination with coefficients of a direction $(\boldsymbol{a}^t\boldsymbol{X})\boldsymbol{a}$: orthogonal projection along a line

Formal statement: We want to find that direction (axis) such that
V$(\boldsymbol{a}^t\boldsymbol{X}) = \boldsymbol{a}^t\Sigma\boldsymbol{a} =$ Max !: principal axis of Σ

Principal axis

Let $\boldsymbol{X}$ be a square integrable random vector with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and Cov($\boldsymbol{X}$) $= \Sigma$.
Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ be the eigenvalues of Σ and $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_p$ the corresponding normed eigenvectors.

Theorem: The principal axis of Σ is $\boldsymbol{b}_1$, i.e. the eigenvector corresponding to the maximal eigenvalue.
Proof: Each direction $\boldsymbol{a}$ can be written as

$$\boldsymbol{a} = \sum_{i=1}^{p} x_i \boldsymbol{b}_i \text{ where } \sum_{i=1}^{p} x_i^2 = 1.$$

(Recall that Pythagoras' theorem implies $1 = ||\boldsymbol{a}||^2 = \sum x_i^2 ||\boldsymbol{b}_i||^2 = \sum x_i^2$.)
Then the assertion follows from

$$\text{V}(\boldsymbol{a}^t\boldsymbol{X}) = \sum_{i=1}^{p} x_i^2 \text{V}(\boldsymbol{b}_i^t\boldsymbol{X}) = \sum_{i=1}^{p} x_i^2 \lambda_i$$

Principal component decomposition

We have $\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{AZ} = \boldsymbol{TD}^{1/2}\boldsymbol{Z}$ which means $X_i - \mu_i = \sum_{k=1}^{p} b_{ik}\sqrt{\lambda_k} Z_k$

Terminology: The variables Z_k are principal components or principal factors ”explaining” the variables X_i.

Questions: Dimensionality reduction

- It is possible to simplify the factor model by omitting some factors without spoiling the covariance structure too much ?
- How many and which factors can be omitted ?

Basic idea of the dimensionality reduction

We know that

$$X_i = \mu_i + \sum_{k=1}^{p} b_{ik}\sqrt{\lambda_k}Z_k$$

For small eigenvalues λ_k the contribution of Z_k could be viewed as negligible. Therefore we would like to omit terms with small λ_k.

We define for some K:

$$\widetilde{X}_i := \mu_i + \sum_{k=1}^{K} b_{ik}\sqrt{\lambda_k}Z_k$$

How to choose K ?

Reduction argument

Let $\widetilde{X}_i := \mu_i + \displaystyle\sum_{k=1}^{K} b_{ik}\sqrt{\lambda_k}Z_k$. How to choose K ?

We note that

$$\mathsf{V}(X_i) = \sum_{k=1}^{p} b_{ik}^2\lambda_k \;\Rightarrow\; \sum_{i=1}^{p}\mathsf{V}(X_i) = \sum_{k=1}^{p}\lambda_k$$

and

$$\mathsf{V}(\widetilde{X}_i) = \sum_{k=1}^{K} b_{ik}^2\lambda_k \;\Rightarrow\; \sum_{i=1}^{p}\mathsf{V}(\widetilde{X}_i) = \sum_{k=1}^{K}\lambda_k$$

Reduction rule: Choose K such that

$$\sum_{k=1}^{K}\lambda_k \geq \alpha\sum_{k=1}^{k}\lambda_k$$

for some given $\alpha \in (0,1)$.

Correlation matrices:

- A correlation matrix is a covariance matrix: symmetric, positive semidefinite.
- A correlation matrix has a Cholesky decomposition and principal components.

Important:

The Cholesky decompositions of a covariance matrix and of the corresponding correlation matrix are equal up to standardization.
It does not matter on which matrix the Cholesky decomposition is based.

The principal components of a covariance matrix and of the corresponding correlation matrix are completely different and not related by standardization.
Therefore it does matter for which matrix principal components are calculated !

Exercise: Calculate the principal components of a 2×2-correlation matrix.

Solution: The characteritic equation is $(1-\lambda)^2 - \rho^2$ giving eigenvalues $\lambda_1 = 1 + |\rho|$ and $\lambda_2 = 1 - |\rho|$. Let $\epsilon = \text{sgn}(\rho)$. Then the eigenvectors are the columns of

$$\boldsymbol{T} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{\epsilon}{\sqrt{2}} & -\frac{\epsilon}{\sqrt{2}} \end{pmatrix}$$

The standardizing transformation $\underline{Z} = \boldsymbol{D}^{-1/2}\boldsymbol{T}^t\underline{X}$ is then given by

$$\begin{aligned} Z_1 &= \frac{1}{\sqrt{2(1+|\rho|)}}(X_1 + \epsilon X_2) \\ Z_2 &= \frac{1}{\sqrt{2(1-|\rho|)}}(X_1 - \epsilon X_2) \end{aligned}$$

Chapter: PCA: Applications

Section: Equity

-> https://github.com/BorisVelichkov/fmi-irkd-kdd-2021/PCA_Theory_Applications_Extensions_vilimiry26402/blob/main/PCA_ex1.ipynb

Chapter: PCA: Applications

Section: Fixed Income

-> Results from an Author's research project (summary)

Figure 1. Credit spread evolution

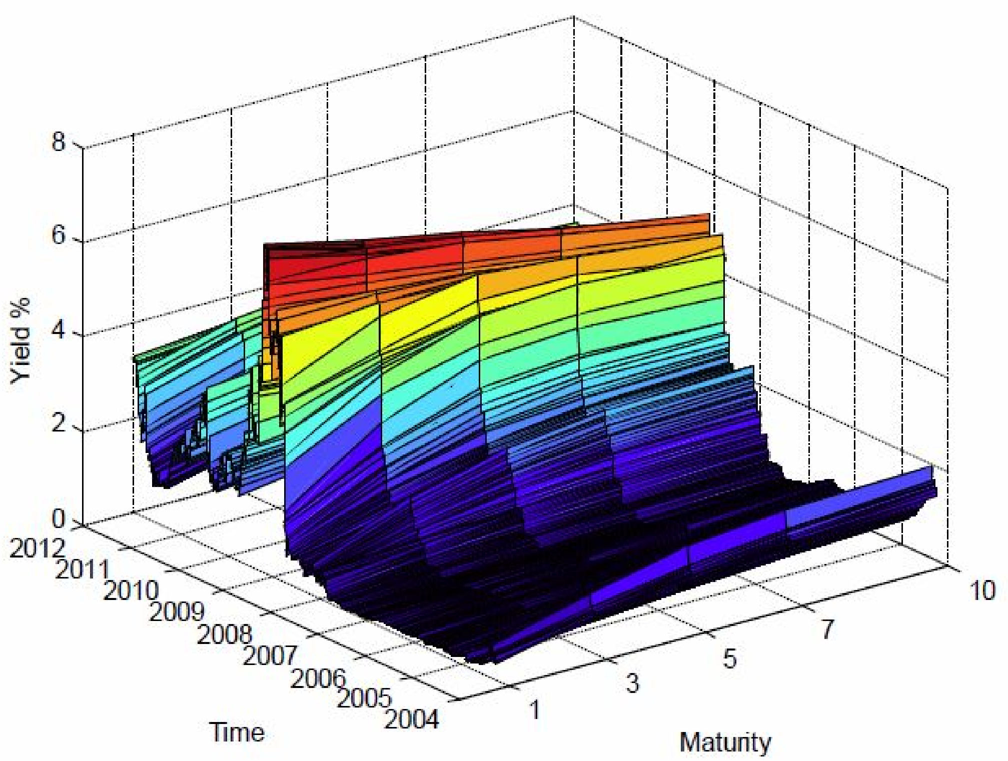


Figure 2. Currency spread evolution

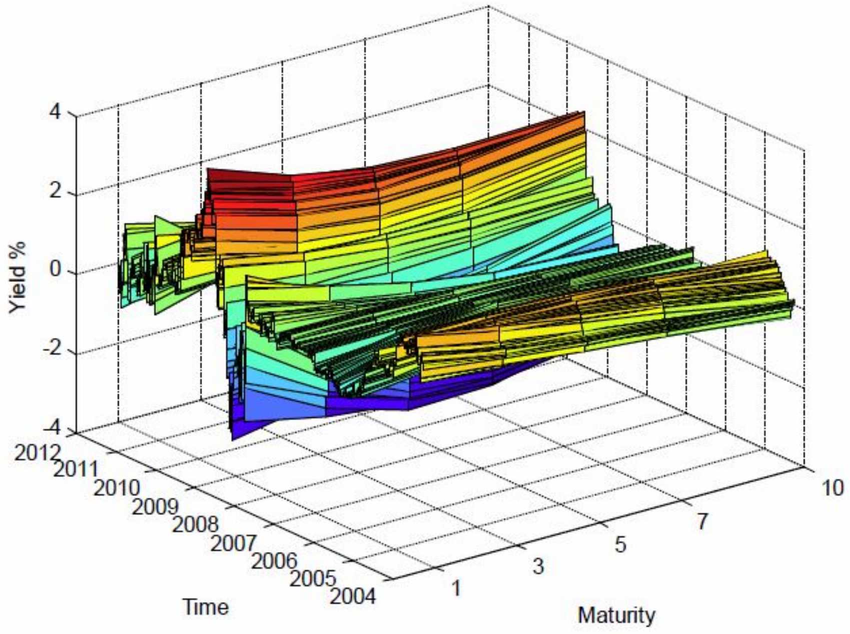


Figure 3. General currency spread evolution

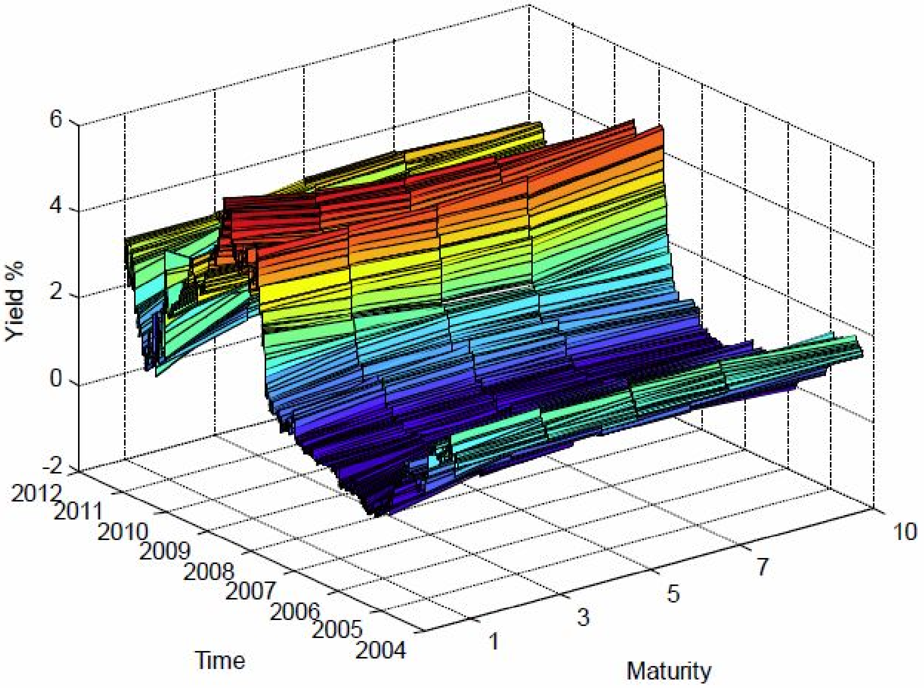


Figure 4. Risky spreads by maturity spectrum

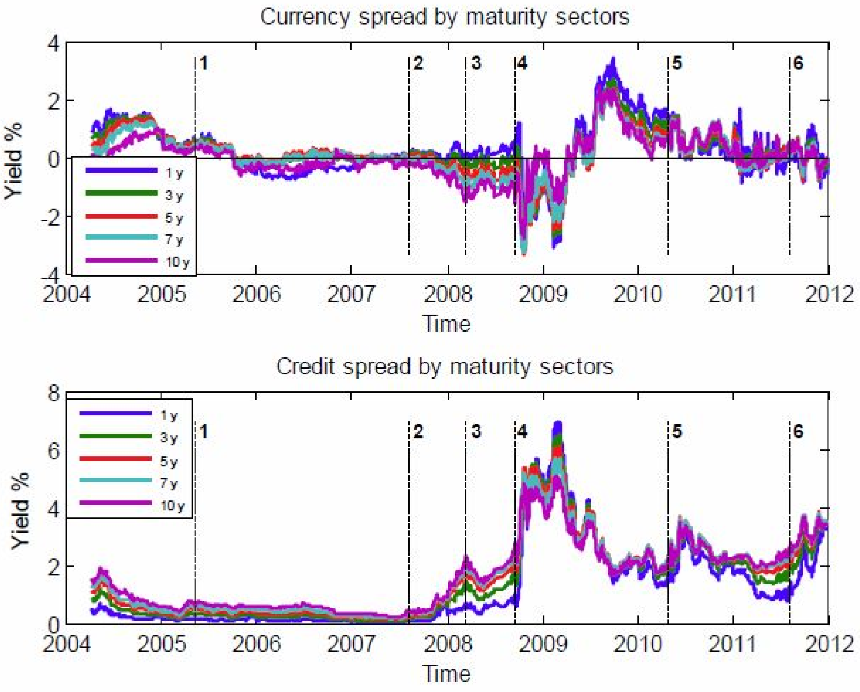
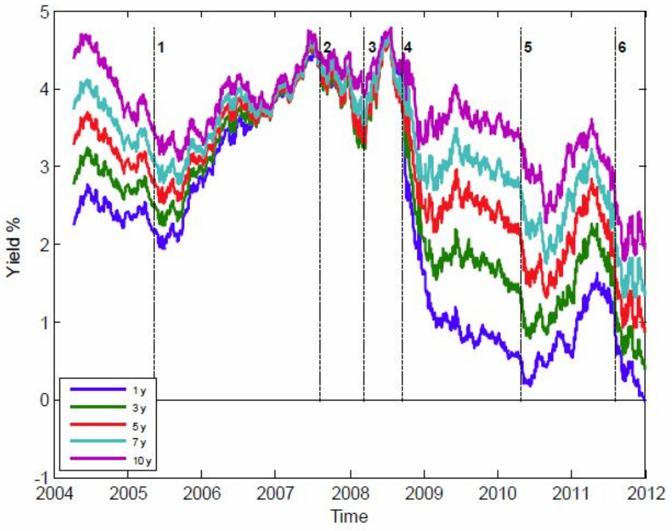


Figure 5. German yields by maturity spectrum



In all the figures, the events: 1 - GM and Ford ratings downgrade of May 09, 2005, 2 - Liquidity crisis of August 09, 2007, 3 - Bear Sterns default of March 14, 2008, 4 - Lehman default of September 15, 2008, 5 - Greek turmoil start of April 23, 2010, 6 – the US rating downgrade of August 05, 2011 are marked by the vertical dashed lines.

Table 1. PCA factors

	Bulgaria		Germany
Factor % / Object	Credit spread	Currency spread	Yield
Shift	98.60	91.39	95.76
Slope	1.36	6.38	4.03
Rotation	0.03	2.20	0.21
4	0.00	0.03	0.00
5	0.00	0.00	0.00

Table 2. PCA and Kalman Filter factors' correlations

Bulgaria			PCA						Kalman			
			Shift			Slope			Ger.		Spr. EUR	Spr. BGN
			Ger.	Spr. EUR	Spr. BGN	Ger.	Spr. EUR	Spr. BGN	fact. 1	fact. 2		
PCA	Shift	Ger.	1,00	0,24	0,34	0,00	-0,20	0,36	0,99	-0,38	-0,06	-0,20
		Spr. EUR	0,24	1,00	-0,25	-0,64	0,00	-0,02	0,25	-0,69	0,94	-0,41
		Spr. LC.	0,34	-0,25	1,00	-0,28	0,06	0,00	0,35	-0,39	-0,35	0,82
	Slope	Ger.	0,00	-0,64	-0,28	1,00	-0,17	0,29	0,00	0,92	-0,66	-0,30
		Spr. EUR	-0,20	0,00	0,06	-0,17	1,00	0,47	-0,20	-0,08	0,07	0,19
		Spr. LC.	0,36	-0,02	0,00	0,29	0,47	1,00	0,36	0,13	-0,14	-0,19
Kalman	Ger.	fact. 1	0,99	0,25	0,35	0,00	-0,20	0,36	1,00	-0,39	-0,06	-0,21
		fact. 2	-0,38	-0,69	-0,39	0,92	-0,08	0,13	-0,39	1,00	-0,59	-0,20
	Spr. EUR		-0,06	0,94	-0,35	-0,66	0,07	-0,14	-0,06	-0,59	1,00	-0,35
	Spr. LC.		-0,20	-0,41	0,82	-0,30	0,19	-0,19	-0,21	-0,20	-0,35	1,00

Figure 6. Credit spread factor dynamics

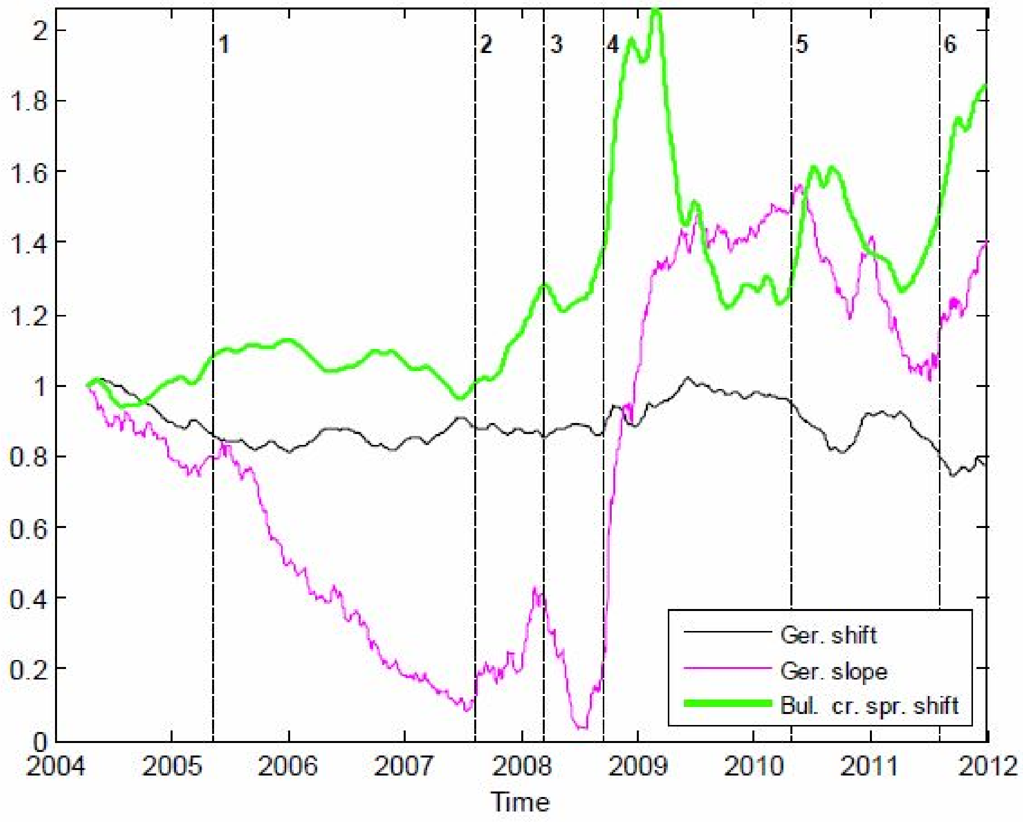
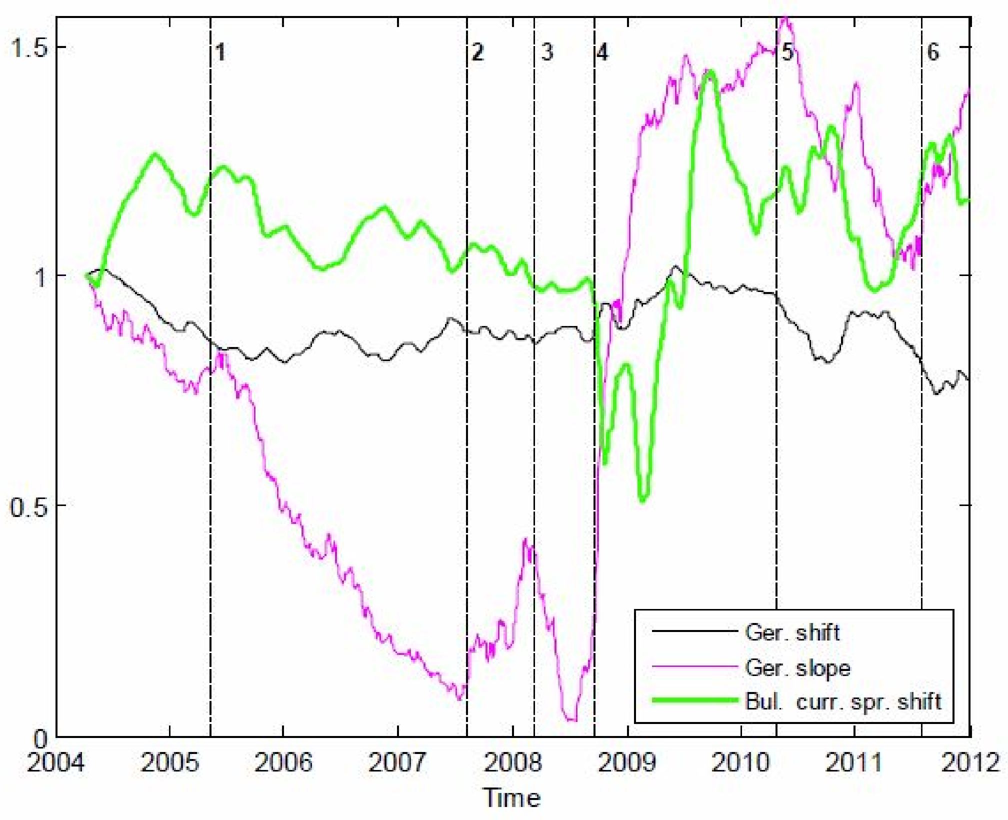


Figure 7. Currency spread factor dynamics



The sections so far are based on the VY research & practice

Chapter: PCA: Extensions

Section: Misc.

Other linear dimensional reduction techniques:

- Independent component analysis (ICA))

  Decomposing a multivariate signal into independent non-Gaussian signals. Details: Lee, T.-W, Independent component analysis: Theory and applications, Boston, Mass: Kluwer Academic Publishers, 1998

- Singular value decomposition (SVD)

  See the previous sections. More details: https://public.lanl.gov/mewall/kluwer2002.html

- Factor analysis

  More general treatment than the PCA. E.g. Dynamic PCA, etc. Details: Gorsuch, Richard, Factor Analysis: Classic Second Edition, Taylor & Francis, 2015

Chapter: PCA: Extensions

Section: Misc.

Non-linear dimensional reduction techniques:

- Sammon's mapping
- Self-organizing map
- Kernel principal component analysis
- Maximum variance unfolding
- Many other algorithms

Details: John A. Lee, Michel Verleysen, Nonlinear Dimensionality Reduction, Springer, 2007

THANK YOU FOR YOUR ATTENTION!

Q&A