An introduction to Principal Component Analysis

Ramon van den Akker (Tilburg University)

Agenda

Introduction and outline PCA

PCA - derivation

PCA - standard derivation

Implementation and remarks

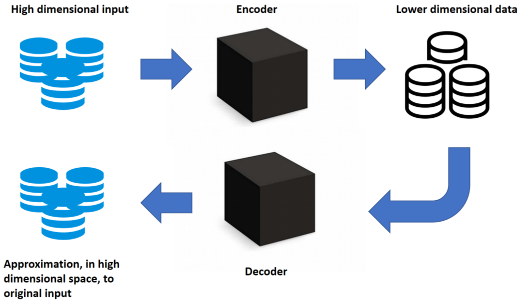
Demo

References

Section 1

Introduction and outline PCA

Dimension reduction



Applications:

- data compression
- reduction dimension features
- noise removal
- visualisation
- anomaly detection

Agenda

- we discuss Principal Component Analysis (PCA)
  - dates back to Pearson (1901) and Hotelling (1933)
- see, for example, Van der Maaten et al. (2009) for review of dimension reduction techniques

Outline - PCA

Heuristic description:

- encoding: find ‘small’ number of directions in input space that explain variation in data as well as possible
- decoding: represent data in original dimension by projecting along those directions

Outline - PCA

Training:

- given p-dimensional observations $X_1, \ldots, X_n$ with mean $\mu = \mathbb{E}X_i$
- choice for dimension encoder is made ($d < \min\{p, n\}$)
- p-dimensional vectors $w_1, \ldots, w_d$, are constructed (*principal components*) and stored

Encoding of observations:

For p-dimensional observation x (can also be new observation):

- calculate *principal scores* $s_1 = w_1'(x - \hat{\mu}), \ldots, s_d = w_d'(x - \hat{\mu})$
- store d-dimensional $(s_1, \ldots, s_d)$ and throw x itself away

Decoding of observations:

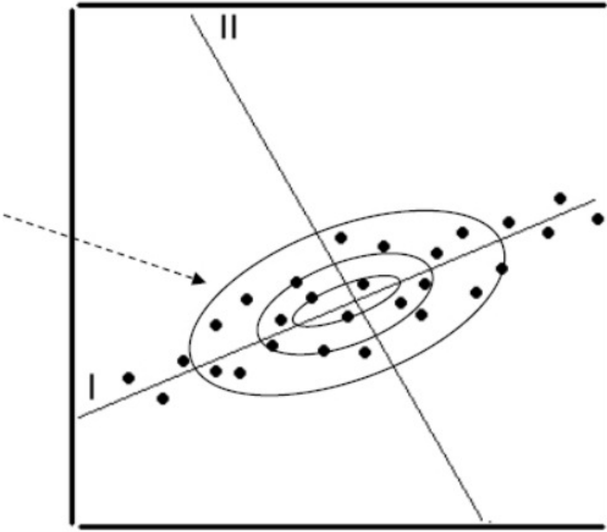
Approximation of x in $\mathbb{R}^p$ by:

$$x_d = \mu + \sum_{j=1}^{d} s_j w_j \in \mathbb{R}^p$$

Need to store $d \times p + n \times d$ numbers instead of $n \times p$.

Intuition

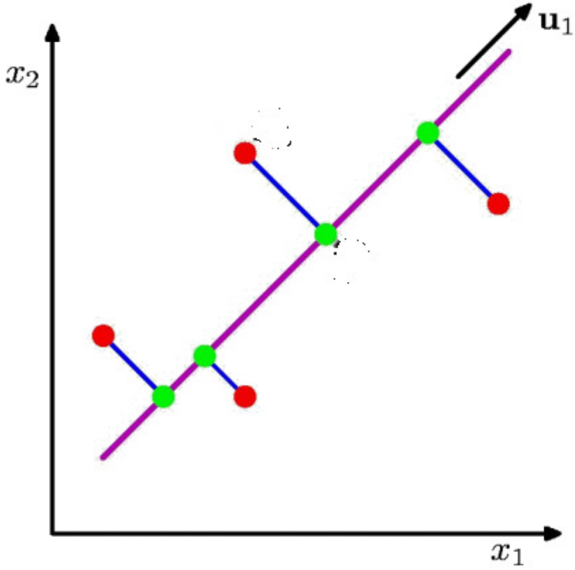
"Best" reduction to dimension 1 of 2-dimensional data?



Such 'directions' in data are described by covariance matrix data

Intuition

Approximate observation $\tilde{x}$ by $s\tilde{x} \in \mathbb{R}^2$, where $s = \tilde{x}' u_1$:



Setup

Setup:

Consider p-dimensional random vector X with mean 0_p and *known* $p \times p$ positive definite covariance matrix Σ

- later on we will consider situation in which Σ is unknown and have i.i.d. observations $X_1, \ldots, X_n$ available

Goal:

Construct, for $d = 1, \ldots, p-1$, linear subspace of dimension d that explains "as much as possible variation" in X

Remarks:

- **First we will consider $\mu = 0$ and Σ to be known**. Afterwards, we will discuss how to proceed in case $\mu \neq 0$ and Σ are unknown.

Section 2

PCA - derivation

PCA - derivation

We will exploit **spectral theorem:**

As Σ is real, symmetric and positive definite matrix we have:

- there are p real, positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and corresponding (real) eigenvectors $u_1, \ldots, u_p \in \mathbb{R}^p$ with
  - $\|u_j\|^2 = u_j' u_j = 1$
  - $u_j' u_i = 0$ for $i \neq j$, i.e. eigenvectors are orthogonal
- Σ can be written as:

$$\Sigma = U \Lambda U' = \sum_{j=1}^{p} \lambda_j u_j u_j'$$

  where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and $U = [u_1 \ \cdots \ u_p]$
- U is orthogonal:
  - $U'U = UU' = I_p$
  - $U^{-1} = U'$
- the eigenvectors u_j are called *principal components*

PCA - derivation

Spectral decomposition yields (please note that X is random vector)

$$X = I_p X = (UU')X = U(U'X) = \sum_{j=1}^{p} (X' u_j) u_j.$$

Note that this represents X in the coordinate system determined by the eigenvectors $u_1, \ldots, u_p$. Approximate X by

$$X_d = \sum_{j=1}^{d} (X' u_j) u_j.$$

We have

$$\text{var}(X' u_j) = u_j' \text{var}(X) u_j = u_j' U \Lambda U' u_j = \lambda_j.$$

And, for $k \neq j$,

$$\text{cov}\left(X' u_j, X' u_k\right) = u_j' \text{var}(X) u_k = u_j' U \Lambda U' u_k = 0.$$

PCA - derivation

For approximation error, $\varepsilon_d = X - X_d$, we obtain

$$\mathbb{E}\|\varepsilon_d\|^2 = \mathbb{E}\left\|\sum_{j=d+1}^{p}(X'u_j)u_j\right\|^2 = \sum_{j=d+1}^{p}\mathbb{E}(X'u_j)^2 = \sum_{j=d+1}^{p}\lambda_j.$$

And, similarly,

$$\mathbb{E}\|X\|^2 = \sum_{j=1}^{p}\lambda_j \text{ and } \mathbb{E}\|X_d\|^2 = \sum_{j=1}^{d}\lambda_j.$$

Measure for variation captured by first d principal components:

$$\frac{\sum_{j=1}^{d}\lambda_j}{\sum_{j=1}^{p}\lambda_j}\,(\times 100\%)$$

PCA

- p-dimensional vectors $w_1 = u_1, \ldots, w_d = u_d$ are called the first d *principal components*
- dimension reduction by using first d principal components:
  - replace p-dimensional observation $\mathbf{x}$ by *principal scores*

$$s_j = w_j'(\mathbf{x} - \mu) \in \mathbb{R}, \quad j = 1, \ldots, d$$

- approximation/reconstruction of $\mathbf{x}$ by

$$x_{PCA} = \mu + \sum_{j=1}^{d} s_j w_j \in \mathbb{R}^p$$

- applying PCA to observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$: instead of storing np numbers, we need to store $p + dp + nd$ numbers

Section 3

PCA - standard derivation

PCA - alternative derivation

Procedure (with $d \leq p$):

- determine $w_1'X$ with $\|w_1\| = 1$ such that $\text{var}(w_1'X)$ is maximal
- determine $w_2'X$ with $\|w_2\| = 1$ and $\text{cov}(w_1'X, w_2'X) = 0$ such that $\text{var}(w_2'X)$ is maximal

  $\vdots$

- determine $w_d'X$ with $\|w_d\| = 1$ and $\text{cov}(w_j'X, w_d'X) = 0$ for $j = 1, \ldots, d-1$ such that $\text{var}(w_d'X)$ is maximal

PCA - alternative derivation

First principal component w_1 solves:

$$\max_{\alpha \in \mathbb{R}^p :\, \|\alpha\|=1} \mathrm{var}(\alpha' X) = \alpha' \Sigma \alpha$$

Use method of Lagrange mulipliers:

$$\max_{\alpha \in \mathbb{R},\, \lambda \in \mathbb{R}} \mathcal{L}(\alpha, \lambda) = \max_{\alpha \in \mathbb{R},\, \lambda \in \mathbb{R}} \alpha' \Sigma \alpha - \lambda(\alpha' \alpha - 1)$$

Stationary point follows from solving:

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\alpha} \mathcal{L}(\alpha, \lambda) = 2\Sigma\alpha - 2\lambda\alpha \\ 1 &= \alpha' \alpha \end{aligned}$$

which yields $\Sigma \alpha = \lambda \alpha$ i.e. α is eigenvector of Σ corresponding to eigenvalue λ

PCA - alternative derivation

From F.O.C. we obtained:

$$\Sigma\alpha = \lambda\alpha$$

As we want to maximize, use constraint $\|\alpha\| = 1$,

$$\alpha'\Sigma\alpha = \alpha'(\Sigma\alpha) = \alpha'\lambda\alpha = \lambda$$

it follows that $w_1 = u_1$ and $\lambda = \lambda_1$

PCA - alternative derivation

- suppose we have already shown $w_j = u_j$ for $j = 1, \dots, d-1$

Note that

$$0 = \text{cov}(w_j' X, w_d' X) = w_d' \Sigma w_j = \lambda_j w_d' w_j \text{ for } j = 1, \dots, d-1$$

To determine w_d we need to solve:

$$\max_{\substack{\alpha \in \mathbb{R}^p:\, \|\alpha\|=1 \\ \text{cov}(w_d' X, w_j' X)=0,\, j=1,\dots,d-1}} \quad \text{var}(\alpha' X) = \alpha' \Sigma \alpha$$

PCA - alternative derivation

Use method of Lagrange mulipliers:

$$\max_{\substack{\alpha \in \mathbb{R}^p \\ \lambda \in \mathbb{R}, \kappa \in \mathbb{R}^{d-1}}} \alpha' \Sigma \alpha - \lambda(\alpha' \alpha - 1) - 2 \sum_{j=1}^{d-1} \kappa_j (w_j' \Sigma \alpha - 0)$$

Stationary point follows from solving:

$$0 = \frac{\mathrm{d}}{\mathrm{d}\alpha} \mathcal{L}(\alpha, \lambda) = 2\Sigma\alpha - 2\lambda\alpha - 2 \sum_{j=1}^{d-1} \kappa_j \Sigma w_j$$

$$1 = \alpha' \alpha$$

$$0 = \alpha' \Sigma w_j \text{ for } j = 1, \ldots, d-1$$

PCA - alternative derivation

Multiplying first equation by w_j, with $j \in \{1, \ldots, d-1\}$, yields

$$w_j' \Sigma \alpha = \lambda w_j' \alpha + \sum_{j=1}^{d-1} \kappa_j \lambda_j$$

Inserting $0 = w_j' \Sigma \alpha = \lambda_j \alpha' w_j = 0$ we obtain $\kappa_j = 0$. Hence

$$\Sigma \alpha = \lambda \alpha$$

As eigenvectors $u_1, \ldots, u_{d-1}$ cannot be used: $w_d = u_d$, the eigenvector corresponding to eigenvalue λ_d

Section 4

Implementation and remarks

Implementation

- we have an algorithm for the case $\mu = 0$ and Σ is known
- now consider situation in which Σ, of rank p, is unknown and $\mu = \mathbb{E}X \neq 0_p$, but have i.i.d. observations $Y_1, \ldots, Y_n$ available
- just use 'anology principle':
  - estimate Σ by sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{\mu})(Y_i - \hat{\mu})'$$

    with $\hat{\mu} = n^{-1} \sum_{i=1}^{n} Y_i$
  - apply PCA using $\hat{\Sigma}$ and centered observations $X_i = Y_i - \hat{\mu}$.
  - rank of $\hat{\Sigma}$ is at most $n - 1$
  - if $n > p$ (and true Σ has full column rank) then you we will typically have $\text{rank}(\hat{\Sigma}) = p$
  - if $\hat{\Sigma}$ is not of rank p, then $d = \text{rank}(\hat{\Sigma})$ is the maximal number of principal components you can use
  - estimators $\hat{w}_1, \ldots, \hat{w}_d$ of of principal components $w_1, \ldots, w_d$

Implementation - to scale or not to scale?

- from the theory it is clear that PCA is not scale invariant (see notebook for numerical illustration), so the results depend on the scale of the variables
  - for example, using 'expenditures in euro' can yield different results compared to using 'expenditures in cents'
- often variables are scaled by their (estimated) standard deviation before applying PCA
- not always a good idea to preprocess variables: if variables have been measured in same units

Remarks - Statistical Properties

Statistical properties?

- not trivial
- consistency and asymptotic normality (for principal components) have been studied for various settings:
  - p is fixed and $n \to \infty$
  - n fixed and $p \to \infty$ (useful for "high-dimensional, but small data")
  - both $p \to \infty$ and $n \to \infty$ (sometimes with restrictions on relative speed, like $n/d \to c \in (0, \infty)$)
  - references: Anderson, T.W. (1963), Jung et al. (2009), and Shen et al. (2016)

Remarks - selected actuarial applications

- use of PCA in yield curve modelling
  - see, for example, Diebold and Li (2006), Barber and Copper (2012)
- use in mortality and longevity modelling
  - see, for example, Yanga et al. (2010)
- use in car insurance
  - see, for example, Segovia-Gonzalez (2009) and Zhu and Wüthrich (2020)

Section 5

Demo

Demo

See notebook

Section 6

References

References

- Anderson, T.W. (1963). Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics* 34, pp.122-148.
- Barber, J.R. and M.L. Copper (2012). Principal component analysis of yield curve movements. *Journal of Economics and Finance* 36, pp.750–765.
- Diebold, F.X. and C. Li (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics* 130, pp.337–364.
- Jung, S. and J. S. Marron (2009). PCA consistency in high dimension low sample size context. *The Annals of Statistics* 37, pp.4104–4130.
- Shen, D. H. Shen, and J.S. Marron (2016). *Journal of Machine Learning Research* 17, pp.1-34
- van der Maaten, L.J.P., E.O. Postma, and H.J. van den Herik. Dimensionality Reduction: A Comparative Review. Tilburg University Technical Report, TiCC-TR 2009-005, 2009.
- Segovia-Gonzalez, M.M., F.M. Guerrero, and P.Herranz (2009). Explaining functional principal component analysis to actuarial science with an example on vehicle insurance. *Insurance: Mathematics and Economics* 45, pp.278–285.
- Turk, M. and A. Pentland (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience* 3(1), 71–86.
- Yanga, S.S., J.C. Yue, and H.-C. Huang (2010). Modeling longevity risks using a principal component approach: A comparison with existing stochastic mortality models. *Insurance: Mathematics and Economics* 46, pp.254–270.
- Zhu, R. and Wüthrich, M. V. (2020). Clustering driving styles via image processing. *Annals of Actuarial Science* 2, pp.276–290.