Week 6 - Unsupervised learning 1.

PCA (PCR and PLS)

Anikó Lovik

a.lovik@fsw.leidenuniv.nl

Statistical learning
2024-03-13

Unsupervised learning

Topics for Weeks 6-7-8:

- Week 6: PCA (and maybe PCR, PLS)
- Week 7: Clustering (k-means and hierarchical)
- Week 8: Gaussian mixture modelling, maybe PCR, PLS

Materials:

- ISL2, chapter 12
- ISL2, chapter 6,: section 6.3 (pp. 251-261)
- Separate articles as examples

Topics for Week 6

1. Supervised and unsupervised learning
2. Principal component analysis
3. Supervised dimension reduction methods
   - Principal components regression
   - Partial least squares
   - (Principal covariates regression)

ISL2 book:
PCA: chapter 12, sections 12.1-12.3 (pp. 497-516)
PCR&PLS: chapter 6, section 6.3 (pp. 251-261)

Unsupervised vs. supervised methods

Unsupervised learning . . .

- has no criterion/label (Y) to supervise the learning
- instead of prediction $\rightarrow$ searching for structure in the data
  - groups of similar observations or variables
  - finds directions that explain most variance
- more exploratory in nature
- more difficult to assess the performance of the method
- interesting alternatives for high-dimensional problems
- sometimes used as pre-processing for supervised methods $\rightarrow$ identifying important variables when having many predictors

Different types of techniques

- **Dimension reduction techniques**
  - **P**rincipal **C**omponent **A**nalysis (PCA)
  - **E**xploratory **F**actor **A**nalysis (EFA)
  - Correspondence analysis
  - **C**anonical **C**orrelation **A**nalysis (CCA)
  - **I**ndependent **C**omponent **A**nalysis (ICA)
  - **N**on-negative **M**atrix **F**actorisation (NMF)
  - **t**-distributed **S**tochastic **N**eighbor **E**mbedding (t-SNE)
  - **M**ulti**D**imensional **S**caling (MDS)
- **Clustering techniques**
  - One-mode clustering
    - k-means clustering
    - Hierarchical clustering
    - Gaussian mixture analysis
    - Latent class analysis
  - Two-mode clustering (bi-clustering)

Principal component analysis

$X_1, \ldots, X_p$ are correlated variables, z_{i1} (the first principal component for observation i, $i = 1, \ldots, n$) is the normalised linear combination of all these variables with the highest variance:

$$z_{i1} = \phi_{11} x_{i1} + \ldots + \phi_{p1} x_{ip} \tag{1}$$

with $\sum_{j=1}^{p} \phi_{j1}^2 = 1$ $\quad$ ($\leftarrow$ this being the normalised part)

For the first component Z_1 then $\phi_{11}, \ldots, \phi_{p1}$ are obtained by solving:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximise}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \tag{2}$$

After finding Z_1, Z_2 again is the linear combination of $X_1, \ldots, X_p$ that has maximal variance and is uncorrelated with Z_1.

Different views of principal component analysis

first view: find (uncorrelated) linear combinations of the (correlated) variables with largest variance across the samples

- summarizes (the variance in) the data into a small number of components (i.e., main directions in the data)
- low-dimensional representation of the dataset

second view: line (1D) or subspace (2D/3D) closest to the data in terms of squared distances (i.e., least squares approximation)

- (with centered variables): approximate x_{ij} with $\sum_{m=1}^{M} z_{im}\phi_{jm}$
- find z_{im} and ϕ_{jm}'s such that

$$\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \sum_{m=1}^{M} z_{im}\phi_{jm})^2 \tag{3}$$

is minimal $\rightarrow$ with $\phi_1, ..., \phi_m$ of length one and orthogonal to each other

A side note on the different views of PCA and its history

These analogies suggest that, in choosing among the infinity of possible modes of resolution of our variables into components, we begin with a component γ_1 whose contributions to the variances of the x_i have as great a total as possible; that we next take a component γ_2, independent of γ_1, whose contribution to the residual variance is as great as possible; and that we proceed in this way to determine the components, not exceeding n in number, and perhaps neglecting those whose contributions to the total variance are small. This we shall call *the method of principal components*. Its technique will be considered in the subsequent sections.

[Hotelling, 1933]

Proportion of variance explained

i.e., how much of the information is not contained in the first few PC's?

The PVE of the m-th component is

$$\frac{\sum_{i=1}^{n}(\phi_{1m}x_{i1}+\phi_{2m}x_{i2}+\cdots+\phi_{pm}x_{ip})^2}{\sum_{i=1}^{n} x_{i1}^2+x_{i2}^2+\cdots+x_{ip}^2}$$

$$\frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n}x_{ij}^2} \tag{4}$$

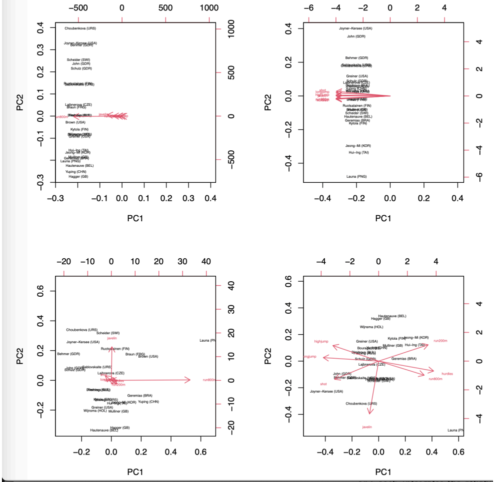
The PVE 's of the $min(n-1,p)$ components sum to one, and the PVE of the first m components can be interpreted similarly to R^2.

Important concepts

- **component loadings** (ϕ_{jm}): weight of each variable in the components (for interpretation of the components)
- **component scores** (z_{im}): score of each case on the components (to see structure among the cases)
- **proportion explained variance of each component**: the variance of each component denotes the importance of that component

Scaling of variables

- Centering - covariance matrix
- Centering + normalisation - correlation matrix



How many components?

p components have advantages, e.g. solve

Maximum number of components is $min(n, p)$ (or $min(n-1, p)$ for scaled variables)

collinearity

Several ways to decide:

- Kaiser's rule (1960): still a default setting in some software, stop to think before using it!
- Horn's parallel analysis (1965) Monte-Carlo
- Cattell's scree plot (1966)
- Velicer's minimum average partial test (MAP) (1976, 2000)
- Revelle & Rocklin's Very Simple Structure (VSS) (1979)
- ... and many more ...

In sum, no simple (or even single) answer possible!

Rotations

make C_1, C_2, C_3 contains diff types of information,

- In PCA, most variables load on the first factors
  $\rightarrow$ rotation to simple structure: easier interpretation
- Rotation redistributes the PVE among the PC's
  $\rightarrow$ successive maximisation of the of the unrotated components is lost
  $\rightarrow$ the total variance of the m components is more evenly distributed
- Two main types:
  - Orthogonal: e.g., varimax, quartimax 正交
  - Oblique: e.g., oblimin 倾斜
- Analyses (with different rotations) are hard to compare

On the importance of checking the input data

Preliminary Eigenvalues: Total = 12101,4962
Average = 310,294776

	Eigenvalue	Difference	Proportion	Cumulative
1	4676,24009	1460,08130	0,3864	0,3864
2	3216,15879	1815,97618	0,2658	0,6522
3	1400,18261	427,54074	0,1157	0,7679
4	972,64188	249,03252	0,0804	0,8483
5	723,60936	347,17931	0,0598	0,9081
6	376,43005	139,10572	0,0311	0,9392
7	237,32433	72,70745	0,0196	0,9588
8	164,61688	46,18434	0,0136	0,9724
9	118,43254	37,33710	0,0098	0,9822
10	81,09545	28,89558	0,0067	0,9889
11	52,19986	10,99219	0,0043	0,9932
12	41,20767	31,51905	0,0034	0,9966
13	9,68862	1,83732	0,0008	0,9974
14	7,85130	0,45058	0,0006	0,9980
15	7,40072	0,84302	0,0006	0,9986
16	6,55769	2,03149	0,0005	0,9992
17	4,52621	1,32161	0,0004	0,9996
18	3,20460	0,57495	0,0003	0,9998
19	2,62964	0,22822	0,0002	1,0000
20	2,40143	1,13838	0,0002	1,0002
21	1,26304	0,35406	0,0001	1,0003

Preliminary Eigenvalues: Total = 12101,4962
Average = 310,294776

	Eigenvalue	Difference	Proportion	Cumulative
22	0,90898	0,15866	0,0001	1,0004
23	0,75032	0,17589	0,0001	1,0005
24	0,57443	0,22816	0,0000	1,0005
25	0,34627	0,26690	0,0000	1,0006
26	0,07937	0,06851	0,0000	1,0006
27	0,01087	0,14854	0,0000	1,0006
28	-0,13767	0,08902	-0,0000	1,0006
29	-0,22669	0,04667	-0,0000	1,0005
30	-0,27336	0,10848	-0,0000	1,0005
31	-0,38184	0,07889	-0,0000	1,0005
32	-0,46072	0,05735	-0,0000	1,0004
33	-0,51807	0,05473	-0,0000	1,0004
34	-0,57280	0,12223	-0,0000	1,0004
35	-0,69503	0,08012	-0,0001	1,0003
36	-0,77515	0,03624	-0,0001	1,0002
37	-0,81139	0,17538	-0,0001	1,0002
38	-0,98677	0,01049	-0,0001	1,0001
39	-0,99726		-0,0001	1,0000

Missing data

- Complete case analysis a.k.a. listwise deletion (dropping incomplete observations): wasteful, can cause bias
- Mean imputation: artificially reduces the variance
- Matrix completion (if MAR): finds PC's and imputes the missing values iteratively (still a single-value imputation and M needs to be set) - used in recommender systems

Other issues to consider with PCA

We have discussed the number of components, scaling, missing data and rotations.

PCA can be impacted by

- outliers
- number of variables loading on a component
- extremely skewed variables
- rounding

Many of these issues are more important with smaller samples.

Group Exercise

Please form groups of 2-4 students.
Look at the paper on Brightspace.

Dietary patterns and risk of nasopharyngeal carcinoma: a population-based case-control study in southern China

Tingting Huang,[1,2] Alexander Ploner,[1] Ellen T Chang,[3,4] Qing Liu,[5,6] Yonglin Cai,[7,8] Zhe Zhang,[9,10] Guomin Chen,[11] Qihong Huang,[12] Shanghang Xie,[5,6] Sumei Cao,[5,6] Weihua Jia,[6] Yuming Zheng,[7,8] Jian Liao,[13] Yufeng Chen,[1] Longde Lin,[10] Ingemar Ernberg,[14] Guangwu Huang,[9,10] Yi Zeng,[11] Yixin Zeng,[6,15] Hans-Olov Adami,[1,16,17] and Weimin Ye[1,18]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm, Sweden; [2]Department of Radiation Oncology, The First Affiliated Hospital of Guangxi Medical University, Nanning, Guangxi, PR China; [3]Center for Health Sciences, Exponent, Inc., Menlo Park, CA, USA; [4]Stanford Cancer

Discuss the following questions with your group:

- What field is this application from?
- What is the aim of the study?
- How was PCA used in this study?
- How is PCA described in the study?
- How are PCA results presented in the study?

Supervised dimension reduction methods

Three methods to improve least squares linear regression model (see Week 5)

- Selection of predictors and fit least squares
  - (best) subset selection
- Shrinkage of regression coefficients (fit least squares with a constraint)
  - Reduces the variance and can perform variable selection
- Dimension reduction methods (fit least squares on derived predictors/features)
  - Based on forming linear combinations of the original variables
  - No explicit selection of variables
  - Not always easy to interpret the linear combinations
  - Reduces the variance because some constraint on the coefficients is imposed (but may lead to bias)
    $\rightarrow$ penalty methods also constrain the coefficients
    $\rightarrow$ adding constraints is the only option when $n \ll p$!

Supervised dimension reduction methods

Two-step procedure

- Step 1: Compute new variables as linear combinations of the original predictors (e.g., z_m's in PCA)
- Step 2: Perform least squares regression with the new variables

Bias-variance trade off

- The constraint increases the bias: it's a simpler model, less flexible
- However, it may reduce the variance

Principal Components Regression (PCR)

- Step 1: perform PCA (on standardized data) and take the first M components
  - Principal components are linear combinations of the original variables that have the largest variance
  - When predictors are correlated: a few principal components will capture most of the data
  - Later principal components are uncorrelated to former ones (no issue of multicollinearity)
  - When $M = P$: original least squares regression is obtained
  - Larger M gives a smaller bias but a larger variance
- Step 2: perform least squares regression with these M components
- Use cross-validation to determine M

PCR assumes that the direction of variation of the predictors is also the direction where the response is varying (i.e., the linear combinations are related to the response)

Partial Least Squares (PLS) regression

- Supervised way of selecting the linear combinations:
  $\rightarrow$ first identifies the $Z_1, \ldots, Z_m$ components that explain a lot of variance in the predictors and that are strongly related with the response Y
- Coefficients are obtained from univariate regressions: directions are strongly determined by variables having the largest correlation with the response
  $\rightarrow$ use standardized predictors and response
- Procedure:
  - Compute loadings for Z_1 by regressing Y onto X_j $\rightarrow$ highest weights for X_j's strongly related to Y.
  - To identify Z_2: first adjust each variable by Z_1 and use the residuals for computing Z_2 same way as Z_1.
  - When all $Z_1, \ldots, Z_m$ have been obtained: fit a least squares regression to predict Y.

Principal Covariates Regression (PCovR)

- Not in ISLR2, but good to know it exists
- Like PLS, it is popular in chemometrics
- Not a two-step procedure $\rightarrow$ simultaneously look for components that explain a lot of variance in the predictors and that are strongly related with the response
- Emphasis on dimension reduction or prediction can be adjusted through weights
  $\rightarrow$ very flexible
- Implemented in R packages: PCovR and PCovR2.

Preparation for workgroup

Please read in ISLR2 book:

- PCA: chapter 12, sections 12.1-12.3 (pp. 497-516)
- PCR&PLS: chapter 6, section 6.3 (pp. 251-261)
  + relevant R labs (parts of sections 12.5, 6.5)

R Labs are also collected into one html file on Brightspace
$\rightarrow$ (look under "Between Lecture + Workgroup")

References

Harold Hotelling (1933)
Analysis of a complex statistical variables into principal components
J Edu Psych 417–441