Module 19 Multivariate Analysis for Genetic data
Session 03: Principal component analysis

Jan Graffelman[1,2]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

[2]Department of Biostatistics
University of Washington
Seattle, WA, USA

28th Summer Institute in Statistical Genetics (SISG 2023)

Contents

A bit of history

- Pearson, K. (1901) On lines and planes of closest fit to systems of points in space *Philosophical Magazine* **6**(2): 559-572.
- Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, **24**: 417-441,498-520.
- Gabriel, K. R. (1971) The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58**(3): 453-467.

PCA objectives

Main goals:

- Reduce the number of variables
- A picture of the data matrix (the biplot)

Demographic example: poverty data set

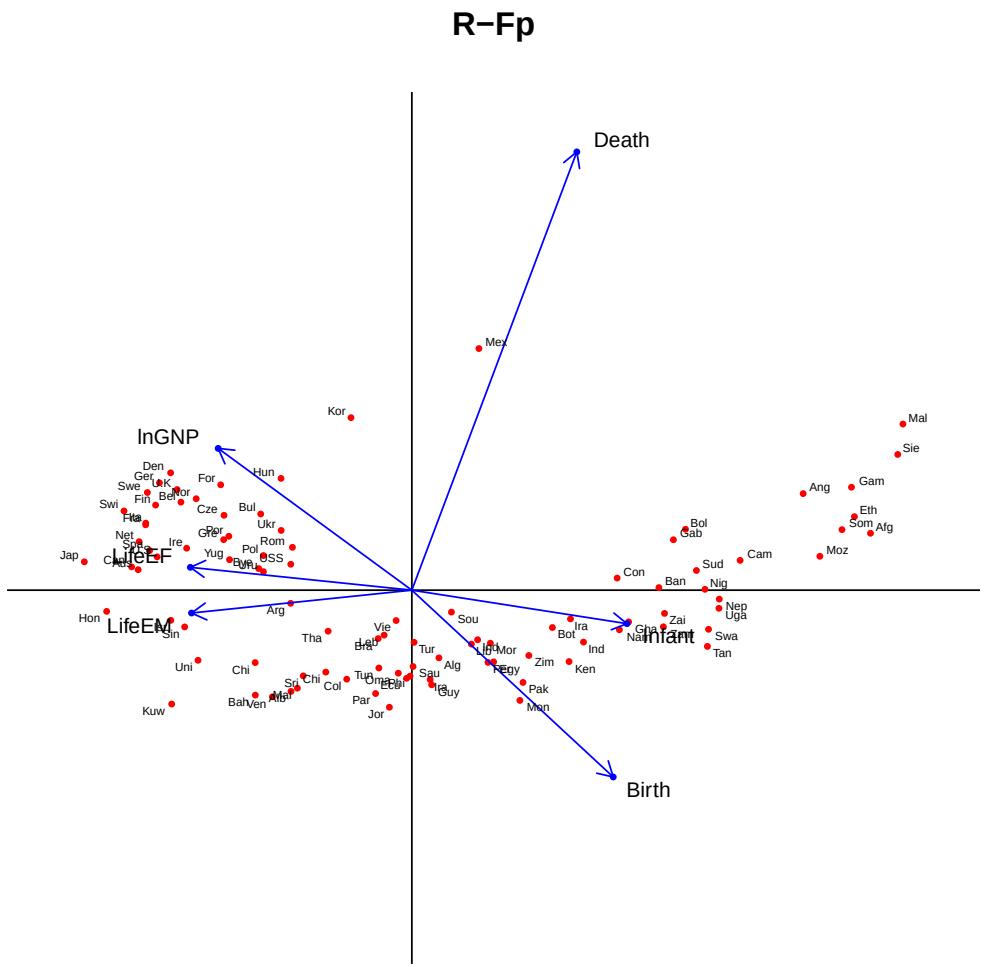
For 97 countries in the world the following variables are registered

- Birth: Live birth rate per 1,000 of population
- Death: Death rate per 1,000 of population
- Infant: Infant deaths per 1,000 of population under 1 year old
- LifeEM: Life expectancy at birth for males
- LifeEF: Life expectancy at birth for females
- GNP: Gross National Product per capita in U.S. dollars
- Country: Name of the country

Country	Birth	Death	Infant	LifeEM	LifeEF	GNP
Albania	24.70	5.70	30.80	69.60	75.50	600
Bulgaria	12.50	11.90	14.40	68.30	74.70	2250
Czechoslovakia	13.40	11.70	11.30	71.80	77.70	2980
Former_E._Germany	12.00	12.40	7.60	69.80	75.90	NA
Hungary	11.60	13.40	14.80	65.40	73.80	2780
Poland	14.30	10.20	16.00	67.20	75.70	1690
Romania	13.60	10.70	26.90	66.50	72.40	1640
Yugoslavia	14.00	9.00	20.20	68.60	74.50	NA
USSR	17.70	10.00	23.00	64.60	74.00	2242
⋮	⋮	⋮	⋮	⋮	⋮	⋮

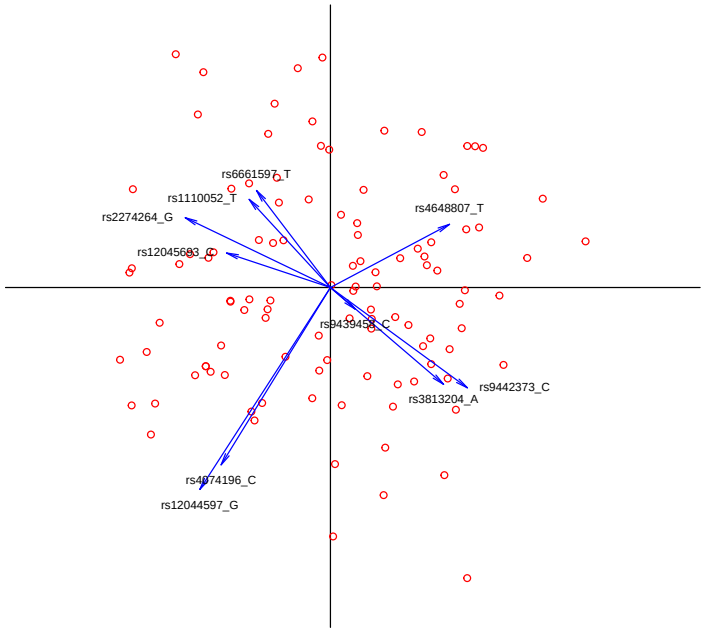
Demographic example: poverty data set

	Birth	Death	Infant	LifeEM	LifeEF	GNP
Alb	24.70	5.70	30.80	69.60	75.50	600
Bul	12.50	11.90	14.40	68.30	74.70	2250
Cze	13.40	11.70	11.30	71.80	77.70	2980
E.G	12.00	12.40	7.60	69.80	75.90	NA
Hun	11.60	13.40	14.80	65.40	73.80	2780
Pol	14.30	10.20	16.00	67.20	75.70	1690
Rom	13.60	10.70	26.90	66.50	72.40	1640
Yug	14.00	9.00	20.20	68.60	74.50	NA
USSR	17.70	10.00	23.00	64.60	74.00	2242
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.



Genetic example: (10 SNPs of the CHD sample of the 1,000G project)

ID	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
NA17962	2	0	0	2	2	0	0	1	0	2
NA17965	1	0	1	2	1	2	2	0	0	1
NA17966	1	1	1	1	0	1	1	2	0	2
NA17967	0	2	1	1	2	1	0	0	0	1
NA17968	1	0	1	1	1	2	1	2	1	2
NA17969	2	2	0	1	1	2	1	1	1	2
NA17970	1	1	2	0	2	0	2	0	0	1
NA17972	1	2	2	2	2	0	1	0	2	0
NA17974	0	2	2	0	2	1	2	1	1	1
NA17975	2	0	0	0	2	1	2	2	2	0
NA17976	0	2	2	1	1	0	0	1	2	0
NA17977	1	0	1	2	1	2	2	1	1	1
NA17978	2	1	1	1	1	1	1	2	2	1
NA17979	1	0	2	1	0	2	0	0	1	0
NA17980	0	0	0	2	0	0	0	2	1	0
NA17981	1	0	0	0	1	2	2	1	1	1
NA17982	1	1	2	1	2	2	0	1	1	0
NA17983	1	0	1	0	2	0	0	2	1	0
NA17986	1	0	0	0	1	2	2	1	1	1
NA17987	0	1	2	0	2	1	2	0	0	1
NA17988	2	1	0	1	0	1	0	0	2	2
NA17989	1	1	1	1	2	1	2	1	1	1
NA17990	1	2	2	1	1	1	2	0	1	1
NA17993	0	2	1	1	1	2	2	0	0	0
NA17995	0	1	1	1	0	1	0	1	1	0
NA17996	1	1	1	1	1	0	1	1	1	1
NA17997	0	1	1	2	1	2	2	2	0	1
NA17998	1	0	0	2	2	1	1	2	0	1
NA17999	2	2	1	0	1	1	0	1	1	1
NA18101	1	1	2	0	1	2	0	2	0	1
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.



Theory PCA (1)

We search for linear combinations of the original variables

$$\begin{aligned} F_{i1} &= a_{11}X_{i1} + a_{12}X_{i2} + \cdots + a_{1p}X_{ip} \\ F_{i2} &= a_{21}X_{i1} + a_{22}X_{i2} + \cdots + a_{2p}X_{ip} \\ &\vdots \\ F_{ip} &= a_{p1}X_{i1} + a_{p2}X_{i2} + \cdots + a_{pp}X_{ip} \end{aligned}$$

Subject to:

- $F_1, F_2, \ldots F_p$ uncorrelated
- $\text{Var}(F_1)$ maximal
- $\text{Var}(F_1) \geq \text{Var}(F_2) \geq \cdots \geq \text{Var}(F_p)$
- $a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1 \qquad (-1 \leq a_{ij} \leq 1)$

Theory PCA

With some algebra, it can be shown that

- The coefficients of the linear combinations are the eigenvectors of the covariance matrix.
- The eigenvalues of the covariance matrix are the variances of the principal components.
- All coefficients and eigenvalues can efficiently be obtained by the spectral decomposition of the covariance matrix:

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{D}_\lambda\mathbf{A}'.$$

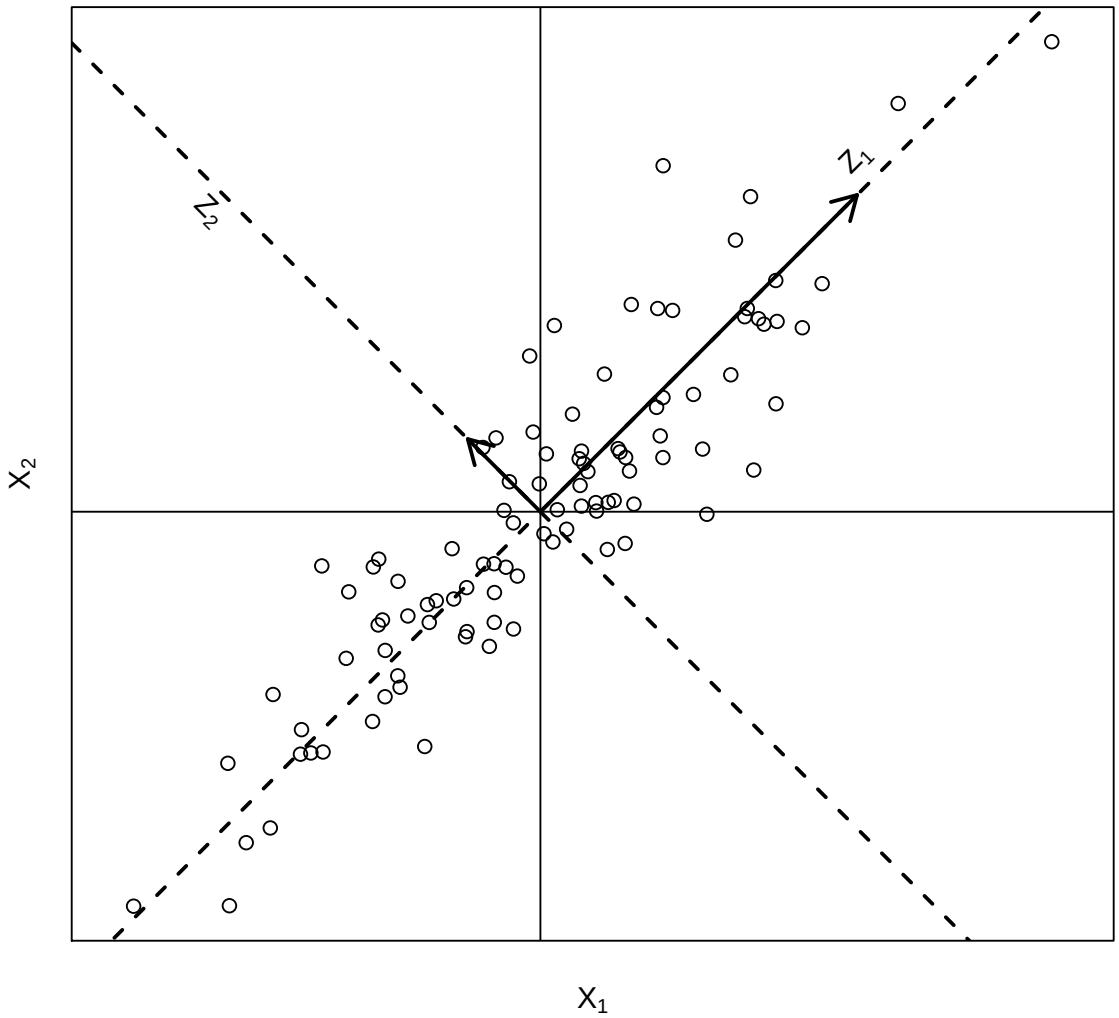
- We will use the sample covariance matrix $\mathbf{S}$ to estimate $\mathbf{\Sigma}$

$$\mathbf{S} = \mathbf{A}\mathbf{D}_\lambda\mathbf{A}'.$$

- and estimate the components by sample components:

$$\underset{(n\times p)}{\mathbf{F}} = \underset{(n\times p)}{\mathbf{X}_c}\ \underset{(p\times p)}{\mathbf{A}}$$

Geometric Interpretation ($p = 2$)



Computing the PCA solution by the SVD

$$\mathbf{X}_t = (1/\sqrt{n})\mathbf{X}_c = \mathbf{U}\mathbf{D}_s\mathbf{A}' \quad (\mathbf{X}_t{}'\mathbf{X}_t = \mathbf{S})$$

$$\mathbf{F}_p = \mathbf{X}_c\mathbf{A} = \sqrt{n}\mathbf{U}\mathbf{D}_s\mathbf{A}'\mathbf{A} = \sqrt{n}\mathbf{U}\mathbf{D}_s$$

$$\frac{1}{n}\mathbf{F}_p{}'\mathbf{F}_p = \frac{1}{n}(\sqrt{n}\mathbf{U}\mathbf{D}_s)'\mathbf{U}\mathbf{D}_s\sqrt{n} = \mathbf{D}_s^2 = \mathbf{D}_\lambda \qquad \lambda_i = d_i^2$$

$$\mathbf{F}_s = \mathbf{F}_p\mathbf{D}_\lambda^{-\frac{1}{2}} = \sqrt{n}\mathbf{U}\mathbf{D}_s\mathbf{D}_s^{-1} = \sqrt{n}\mathbf{U}$$

$$\frac{1}{n}\mathbf{F}_s{}'\mathbf{F}_s = \mathbf{U}'\mathbf{U} = \mathbf{I}.$$

and we obtain factorizations for making biplots

$$\mathbf{X}_c = \sqrt{n}\mathbf{U}\mathbf{D}_s\mathbf{A}' = \mathbf{F}_p\mathbf{A}' \text{ or } \mathbf{X}_c = \mathbf{F}_s(\mathbf{A}\mathbf{D}_s)'$$

Notes:

- The SVD paves the way for biplot construction
- The variables are also LC of the components
- PCA provides a low rank approximation to the centred data matrix that is optimal in the least squares sense
- PCA also provides a low rank approximation to the covariance matrix that is optimal in the least squares sense

PCA biplots – some notation

- PCA allows the construction of many different biplots. Depending on whether the analysis is based on **S** or on **R** and on how the singular values are distributed over column and/or row coordinates, the corresponding biplots will have different properties.
- We will use the following notation for row and column coordinates in biplots. **F** refers to the row coordinates, **G** refers to the column coordinates. A suffix p or s is used to indicate principal or standard coordinates respectively. Ignoring weights and scaling factors, these are defined as:

$$\mathbf{F}_p = \mathbf{UD} \qquad \mathbf{F}_s = \mathbf{U}$$
$$\mathbf{G}_s = \mathbf{V} \qquad \mathbf{G}_p = \mathbf{VD}$$

- The principal coordinates carry the weight of the singular values, whereas the standard coordinates do not.
- Thus we can make biplots by the joint use of $\mathbf{F}_s$ and $\mathbf{G}_p$ or the joint use of $\mathbf{F}_p$ and $\mathbf{G}_s$. Other scalings are certainly possible, but the ones proposed here are relatively common and have many nice properties.

The "form" biplot

- Also called row metric preserving (RMP) biplot.
- We use biplot factorization $\mathbf{X}_c = \mathbf{F}_p \mathbf{G}_s{}'$
- Take $\mathbf{F}_p = \mathbf{U}\mathbf{D}_s$ and $\mathbf{G}_s = \mathbf{A}$
- This biplot of the centred data matrix is obtained by plotting the first two principal components (row points) and the rows of the first two eigenvectors (as arrows).
- Note that $\mathbf{X}_c \mathbf{X}_c{}' = n\mathbf{U}\mathbf{D}_\lambda \mathbf{U}' = \mathbf{F}_p \mathbf{F}_p{}'$, implying:
- $d_E^2(\mathbf{x}_i, \mathbf{x}_{i'}) = (\mathbf{x}_i - \mathbf{x}_{i'})'(\mathbf{x}_i - \mathbf{x}_{i'}) = (\mathbf{f}_i - \mathbf{f}_{i'})'(\mathbf{f}_i - \mathbf{f}_{i'}) = d_E^2(\mathbf{f}_i, \mathbf{f}_{i'})$.
- This biplot preserves the Euclidean distance

The "covariance" biplot

- Also called column-metric preserving (CMP) biplot.
- We use biplot factorization $\mathbf{X}_c = \mathbf{F}_s \mathbf{G}_p{}'$
- Take $\mathbf{F}_s = \sqrt{n}\mathbf{U}$ and $\mathbf{G}_p = \mathbf{A}\mathbf{D}_s$.
- $\mathbf{G}_p$: contains the covariances between variables and components.

$$\frac{1}{n}\mathbf{X}_c{}'\mathbf{F}_s = \frac{1}{n}\sqrt{n}\mathbf{A}\mathbf{D}_s\mathbf{U}'\mathbf{U}\sqrt{n} = \mathbf{A}\mathbf{D}_s = \mathbf{G}_p$$

- This biplot of the centred data matrix is obtained by plotting the first two standardized principal components (row points) and the covariances of the first two components with the variables (as arrows).
- $\mathbf{X}_c\mathbf{S}^{-1}\mathbf{X}_c{}' = n\mathbf{U}\mathbf{U}' = \mathbf{F}_s\mathbf{F}_s{}'$
- $d_M^2(\mathbf{x}_i, \mathbf{x}_{i'}) = (\mathbf{x}_i - \mathbf{x}_{i'})'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_{i'}) = (\mathbf{f}_i - \mathbf{f}_{i'})'(\mathbf{f}_i - \mathbf{f}_{i'}) = d_E^2(\mathbf{f}_i, \mathbf{f}_{i'})$.
- Euclidean distances in the biplot represent Mahalanobis distances.

More properties of the covariance biplot

- $\mathbf{G}_p\mathbf{G}_p{}' = \mathbf{A}\mathbf{D}_\lambda\mathbf{A}' = \mathbf{S} = \frac{1}{n}\mathbf{X}_c{}'\mathbf{X}_c$
- $\cos(\mathbf{g}_i, \mathbf{g}_j) = \frac{\mathbf{g}_i{}'\mathbf{g}_j}{\|\mathbf{g}_i\|\,\|\mathbf{g}_j\|} = \frac{\frac{1}{n}\mathbf{x}_i{}'\mathbf{x}_j}{\frac{1}{\sqrt{n}}\|\mathbf{x}_i\|\,\frac{1}{\sqrt{n}}\|\mathbf{x}_j\|} = \frac{\frac{1}{n}\mathbf{x}_i{}'\mathbf{x}_j}{\sqrt{\frac{1}{n}\mathbf{x}_i{}'\mathbf{x}_i}\sqrt{\frac{1}{n}\mathbf{x}_j{}'\mathbf{x}_j}} = r(\mathbf{x}_i, \mathbf{x}_j)$
- $\|\,\mathbf{g}_i\,\| = \sqrt{(1/n)\mathbf{x}_i{}'\mathbf{x}_i} = s_{x_i}$
- For standardized data, $\|\,\mathbf{g}_i\,\| = \sqrt{r^2(\mathbf{x}_i, \mathbf{F}_1) + r^2(\mathbf{x}_i, \mathbf{F}_2) + \cdots + r^2(\mathbf{x}_i, \mathbf{F}_p)} = \sqrt{R^2}$
- Note that these are all full space results.

How many components ?

Criteria:

- Percentage of explained variance ($> 80\%$).
- Size of the eigenvalue ($> \overline{\lambda}$).
- The scree plot.
- Significance tests with the eigenvalues.

How many components ?

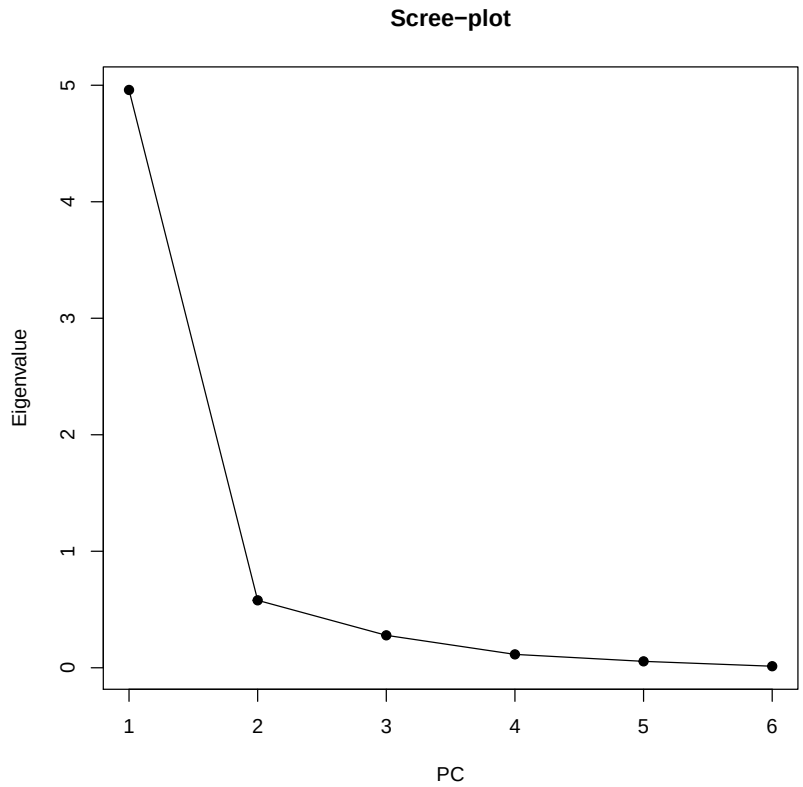
We have:

$$\text{tr}(\mathbf{S}) = \text{tr}(\mathbf{A}\mathbf{D}_\lambda\mathbf{A}') = \text{tr}(\mathbf{D}_\lambda).$$

$$\sum_{i=1}^{p} V(X_i) = \sum_{i=1}^{p} V(F_i) = \sum_{i=1}^{p} \lambda_i.$$

Component	F_1	F_2	$\cdots$	F_p
Variance	λ_1	λ_2	$\cdots$	λ_p
Fraction	$\lambda_1/\sum\lambda_i$	$\lambda_2/\sum\lambda_i$	$\cdots$	$\lambda_p/\sum\lambda_i$
Cum. Fraction	$\lambda_1/\sum\lambda_i$	$(\lambda_1+\lambda_2)/\sum\lambda_i$	$\cdots$	$\sum\lambda_i/\sum\lambda_i$

The scree plot (poverty data set; correlation-based)



Types of PCA

There are two types of PCA. Computations can be based on

- the covariance matrix (**S**)
  - Not invariant w.r.t. the scale of measurement
  - The variable with the largest variance dominates
  - Some authors focus on components with $\lambda_i > \overline{\lambda}$
- the correlation matrix (**R**)
  - Invariant w.r.t. the scale of measurement
  - All variables have equal weight
  - Some authors focus on components with $\lambda_i > 1$

Interpretation

Components can be interpreted with the aid of:

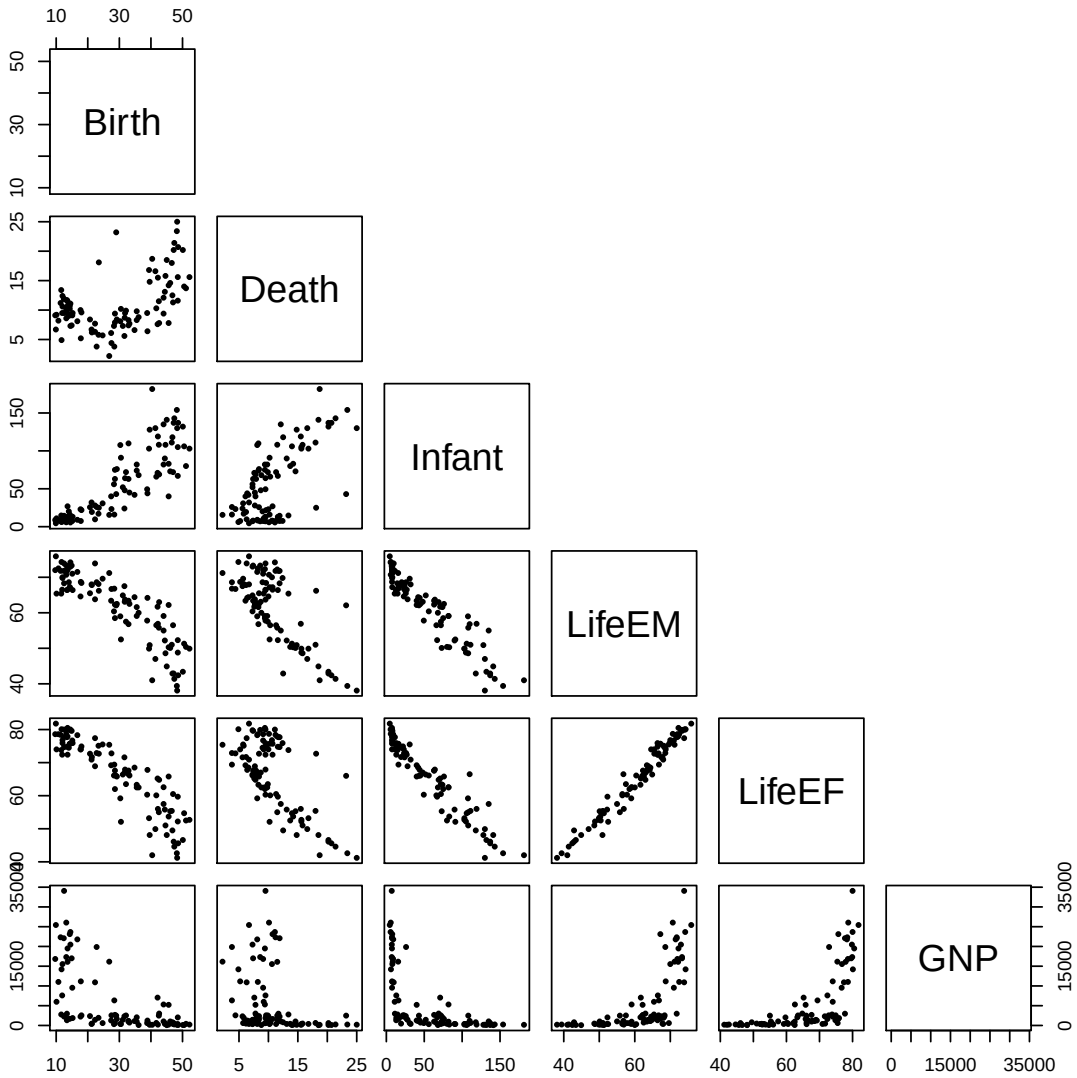
- the coefficients
- the correlations between variables and components
- the biplot

If the aim is to get a picture of the data matrix, then interpretation of the components may not be needed.

Two examples

- UN poverty data ($n = 97$ countries, $p = 6$ variables)
- 1,000 Genomes project CHD sample ($n = 109$, $p = 28,158$ SNPs)

Scatterplot matrix poverty data



Correlation matrix and variance decomposition

	Birth	Death	Infant	LifeEM	LifeEF	lnGNP
Birth	1.00	0.49	0.86	-0.87	-0.89	-0.74
Death	0.49	1.00	0.65	-0.73	-0.69	-0.51
Infant	0.86	0.65	1.00	-0.94	-0.96	-0.79
LifeEM	-0.87	-0.73	-0.94	1.00	0.98	0.81
LifeEF	-0.89	-0.69	-0.96	0.98	1.00	0.83
lnGNP	-0.74	-0.51	-0.79	0.81	0.83	1.00

	PC1	PC2	PC3	PC4	PC5	PC6
λ	4.96	0.58	0.28	0.11	0.06	0.01
fraction	0.83	0.10	0.05	0.02	0.01	0.00
cumulative	0.83	0.92	0.97	0.99	1.00	1.00

Coefficients (eigenvectors) and correlations

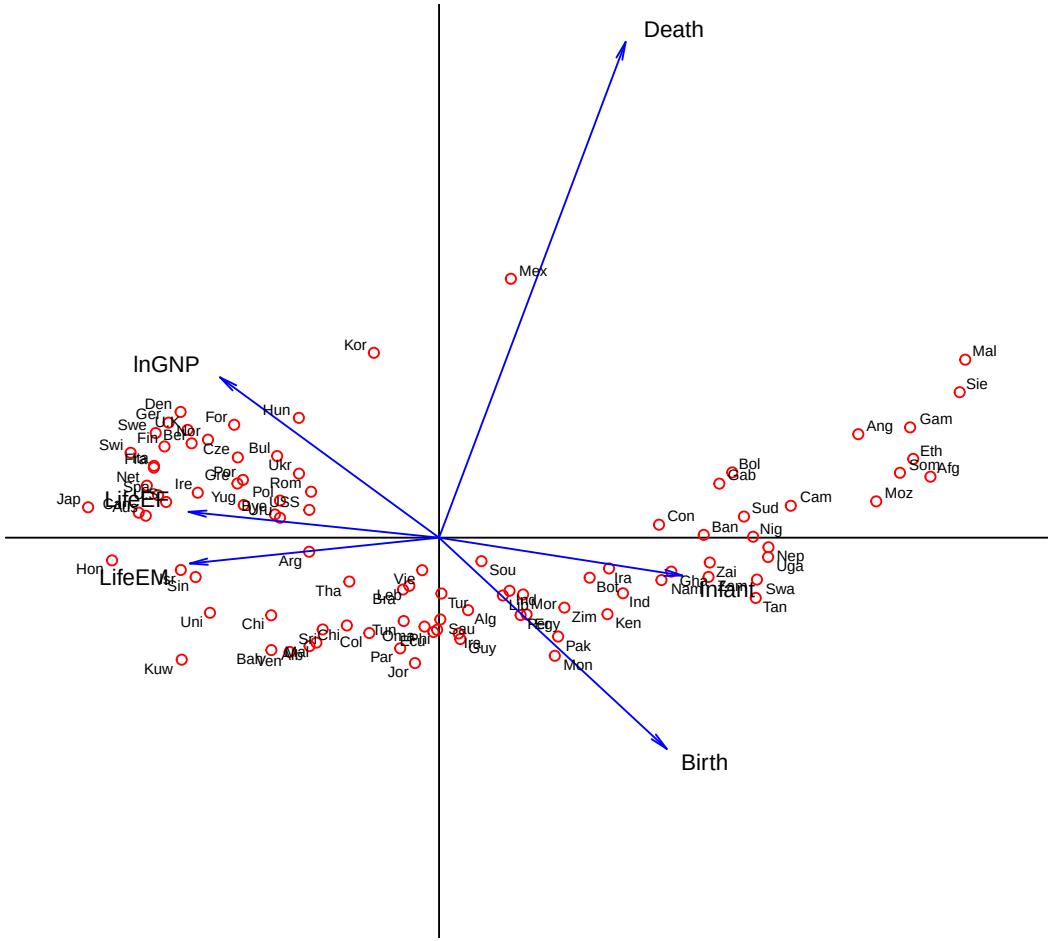
Coefficients

	PC1	PC2	PC3	PC4	PC5	PC6
Birth	0.40	-0.37	0.46	-0.65	0.23	-0.07
Death	0.33	0.88	-0.06	-0.28	0.20	0.02
Infant	0.43	-0.07	0.17	0.65	0.58	-0.14
LifeEM	-0.44	-0.05	-0.09	-0.14	0.66	0.59
LifeEF	-0.44	0.05	-0.11	-0.17	0.36	-0.79
lnGNP	-0.39	0.28	0.85	0.16	-0.11	0.03

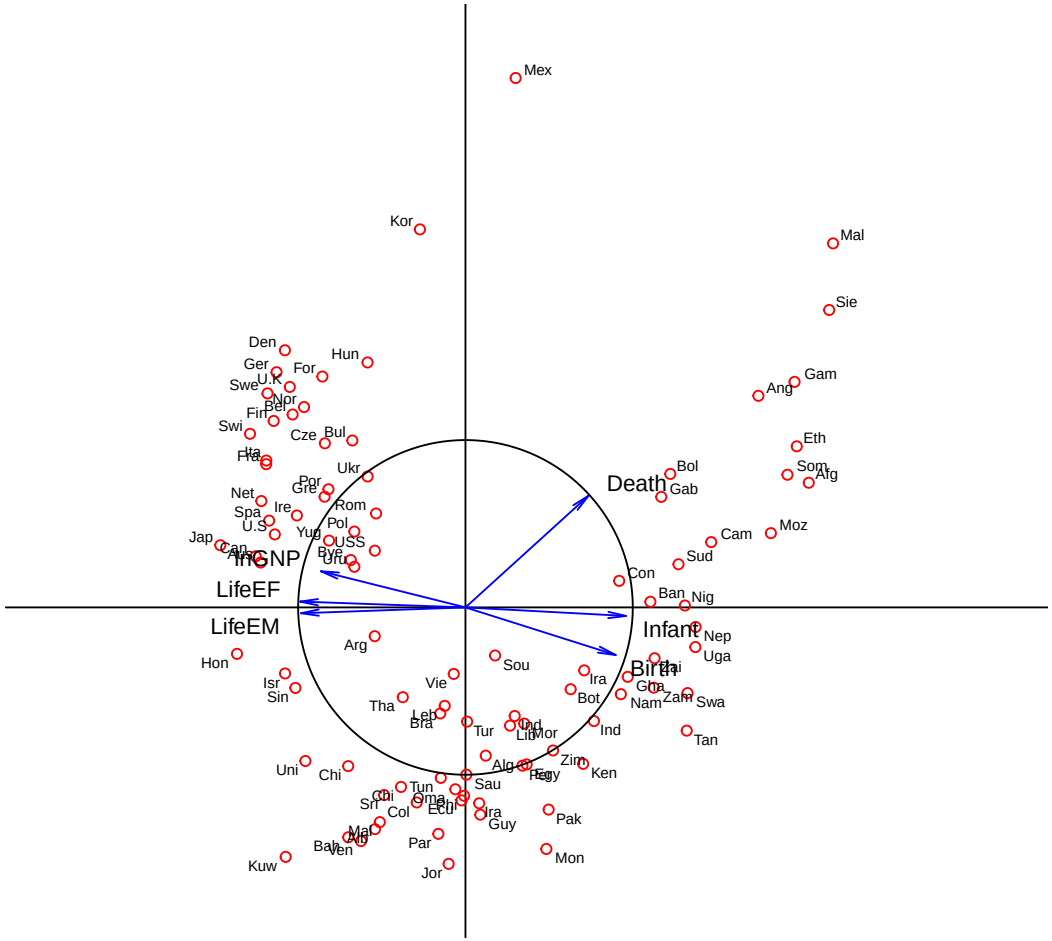
Correlations

	PC1	PC2	PC3	PC4	PC5	PC6
Birth	0.90	-0.28	0.25	-0.22	0.05	-0.01
Death	0.74	0.67	-0.03	-0.09	0.05	0.00
Infant	0.96	-0.05	0.09	0.22	0.14	-0.02
LifeEM	-0.98	-0.03	-0.05	-0.05	0.15	0.07
LifeEF	-0.99	0.03	-0.06	-0.06	0.09	-0.09
lnGNP	-0.86	0.22	0.45	0.06	-0.03	0.00

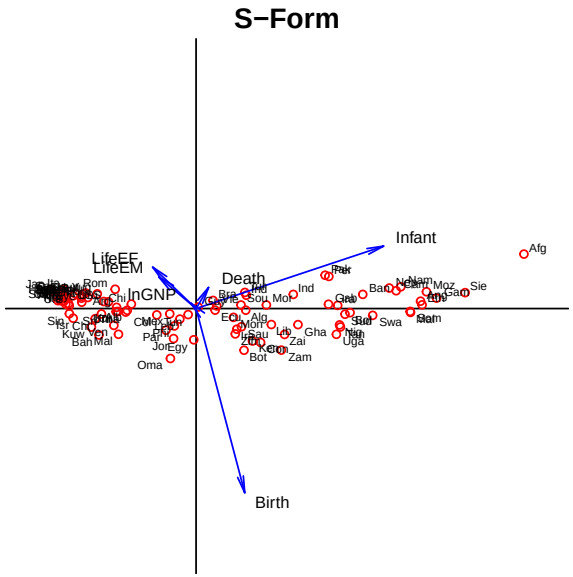
Form biplot (principal components)

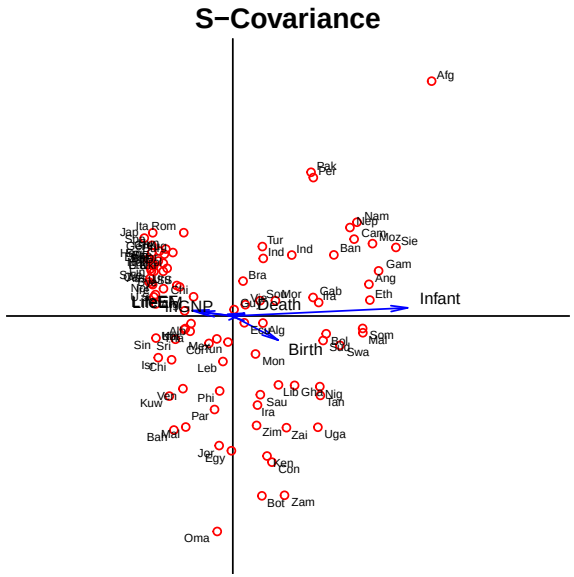


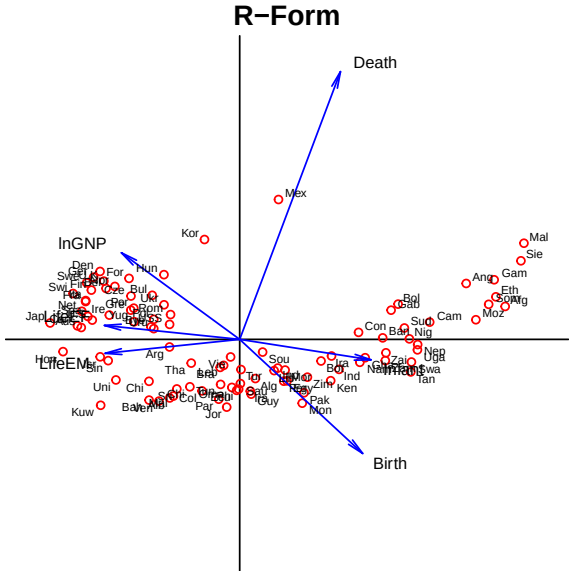
Covariance biplot (standardized principal components)

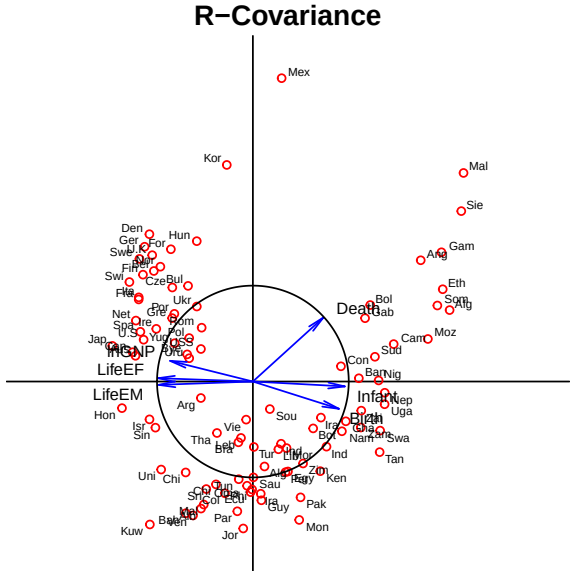


Four PCA biplots









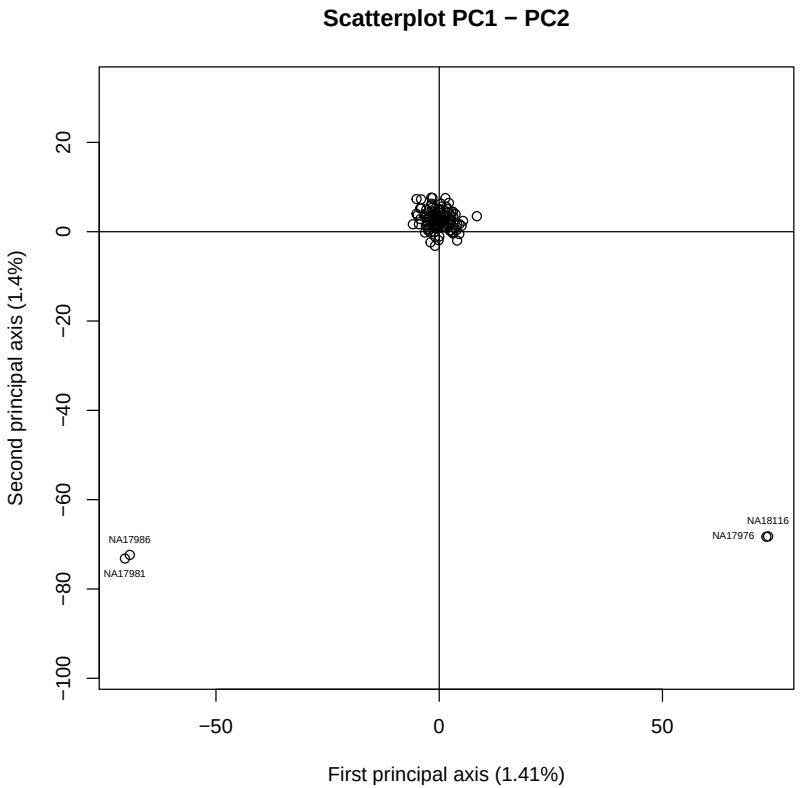
Genetic example: CHD (Chinese in Metropolitan Denver) data; 1,000G project

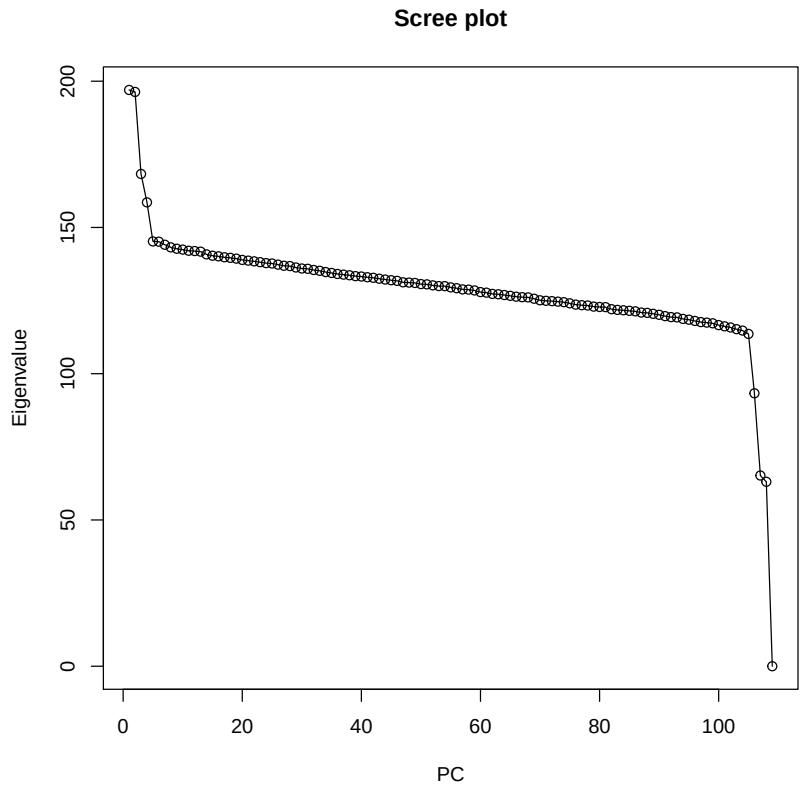
Genetic data typically filtered prior to PCA, using multiple criteria:

- chromosome (X chromosome mostly excluded)
- missingness
- MAF
- test results for HWE
- correlation structure (LD pruning)
- ...

Here we use 24.087 autosomal, complete, highly polymorphic SNPs of 109 CHD individuals.

CHD (Chinese in Metropolitan Denver) data; 1,000G project

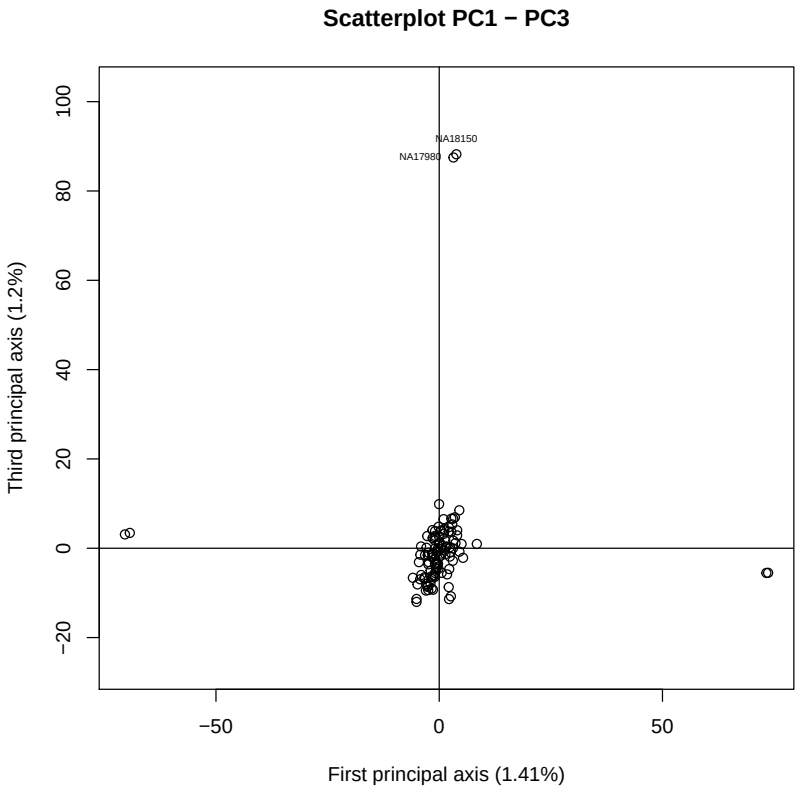


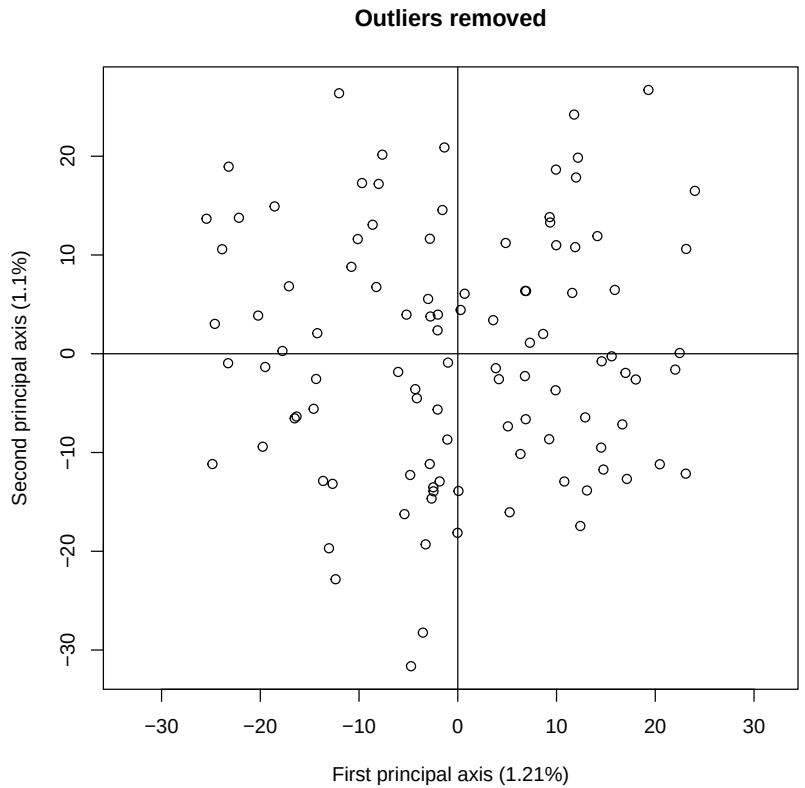


Observations

- The joint plot of individuals and variables is too dense.
- $p > n$ and there are only $n - 1$ PCs with non-zero variance.
- Explained variance is very low (typical for large-scale genetic applications).
- PCA detects outliers.
- In this application, outliers are documented related pairs (one FS; one PO; one 2ND).
- A covariance based PCA is the natural choice.
- The form biplot is the natural choice for this data.
- Computationally not attractive to extract eigenvectors of $\mathbf{S}$.

CHD (Chinese in Metropolitan Denver) data; 1,000G project





References

- Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis*, Second edition, John Wiley, New York. Chapter 11.
- Johnson & Wichern, (2002) *Applied Multivariate Statistical Analysis*, 5th edition, Prentice Hall, Chapter 8.
- Jolliffe, I.T. (1986) *Principal Component Analysis*, Springer-Verlag, New York.
- Manly, B.F.J. (1989) *Multivariate statistical methods: a primer*. 3rd edition. Chapman and Hall, London. Chapter 6.
- Mardia, K.V. et al. (1979) *Multivariate Analysis*. Academic press. Chapter 10.