NA

Alicja Gosiewska

Warsaw University of Technology

14 I 2018

Julie Josse

source: http://juliejosse.com

- a professor of Statistics at Ecole Polytechnique
- XPOP INRIA team.
- My main research fields are: missing values, causal inference, visualization with dimensionality reduction (PCA, correspondence analysis), multi-blocks data, low rank matrix estimation, questionnaire analyses; main application with health data
- Author of more than 100 publications, cited almost 4000 times (Google Scholar).

Research

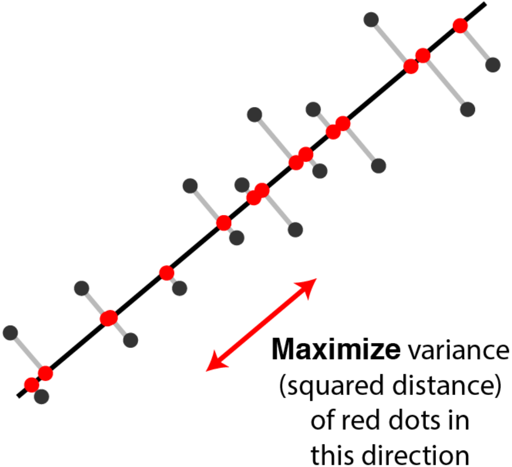
On going projects:

- Handling severe trauma patients
- Causal inference with missing values (with S Hamada, Stefan Wager)
- Low-rank estimation with MNAR data (with Claire Boyer)
- Random Forests with missing values (with Erwan Scornet, Gael Varoquaux)
- Variable selection with adaptive slope (with Gosia Bogdan)
- Students & group's meeting

source: `http://juliejosse.com`

Principal Component Analysis

Geometrical point of view:
providing a subspace that maximizes the variance of the projected points, and therefore represents the diversity of the individuals.



Maximize variance
(squared distance)
of red dots in
this direction

Principal Component Analysis

Equivalently:
providing a subspace that minimizes the Euclidean distance between individuals and their projection onto the subspace. It boils down to finding a matrix of low rank S that gives the best approximation of the matrix $X_{n \times p}$ with n individuals and p variables in the least squares sense.

$$||X_{n \times p} - \hat{X}_{n \times p}||^2$$

Dealing with missing values

A common approach to deal with missing values in PCA involves ignoring the missing values by minimizing the least squares criterion over all non-missing entries.

$$||X_{n\times p} - \hat{X}_{n\times p}||^2$$

This can be achieved by the introduction of a weighted matrix W with

$$w_{ij} = \begin{cases} 0, & \text{if } x_{ij} \text{ is missing.} \\ 1, & \text{otherwise.} \end{cases}$$

in the criterion

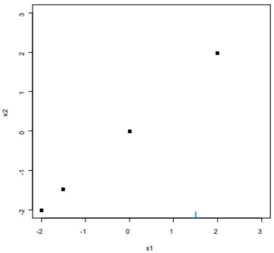
$$||W_{n\times p} * (X_{n\times p} - \hat{X}_{n\times p})||^2$$

where $*$ is the Hadamard product.

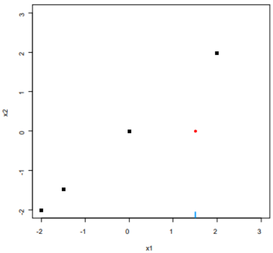
Iterative PCA algorithm

- Kiers H (1997). “Weighted Least Squares Fitting Using Ordinary Least Squares Algorithms.” Psychometrika,
- Josse J, Husson F (2012). “Handling Missing Values in Exploratory Multivariate Data Analysis Methods.”

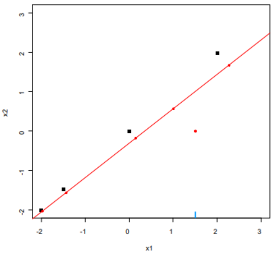
```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98
```



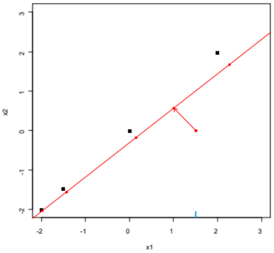
```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.00
 2.0  1.98
```



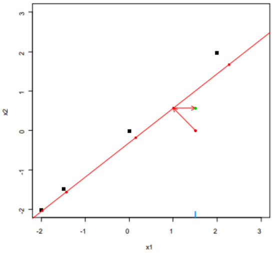
```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.00
 2.0  1.98

   ^     ^
  x1    x2
-1.98 -2.04
-1.44 -1.56
 0.15 -0.18
 1.00  0.57
 2.27  1.67
```



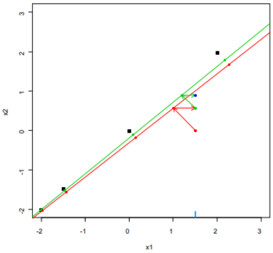
```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.00
 2.0  1.98

   ^     ^
  x1    x2
-1.98 -2.04
-1.44 -1.56
 0.15 -0.18
 1.00  0.57
 2.27  1.67
```



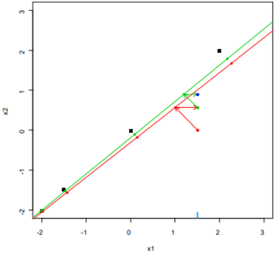
```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.00
 2.0  1.98

   ^     ^
  x1    x2
-1.98 -2.04
-1.44 -1.56
 0.15 -0.18
 1.00  0.57
 2.27  1.67

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.57
 2.0  1.98
```



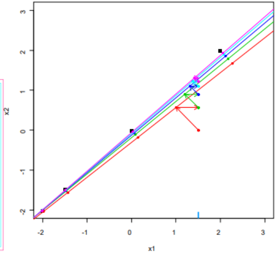
```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.57
 2.0  1.98

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.57
 2.0  1.98
```



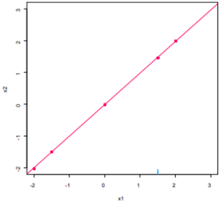
```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.57
 2.0  1.98

   ^     ^
   x1    x2
-2.00 -2.01
-1.47 -1.52
 0.09 -0.11
 1.20  0.90
 2.18  1.78

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.90
 2.0  1.98
```



```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.00
 2.0  1.98

   ^     ^
   x1    x2
-1.98 -2.04
-1.44 -1.56
 0.15 -0.18
 1.00  0.57
 2.27  1.67

  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  0.57
 2.0  1.98
```



```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5    NA
 2.0  1.98
```



```
  x1    x2
-2.0 -2.01
-1.5 -1.48
 0.0 -0.01
 1.5  1.46
 2.0  1.98
```

Regularized Iterative PCA Algorithm

The iterative PCA algorithm provides a good estimation of the PCA parameters when there are very strong correlations between variables and the number of missing values is very small. However, it very rapidly suffers from overfitting problems when data are noisy or there are many missing values.
To tackle this issue, a common strategy is to use regularized methods.

- Josse J, Pagès J, Husson F (2009). "Gestion des Données Manquantes en Analyse en Composantes Principales." Journal de la Société Française de Statistique, 150(2), 28–51.
- Josse J, Husson F (2012). "Handling Missing Values in Exploratory Multivariate Data Analysis Methods." Journal de la Société Française de Statistique, 153(2), 79–99.

Regularization

Verbanck M, Josse J, Husson F (2015). "Regularized PCA to Denoise and Visualise Data." Statistics and Computing, 25(2), 471–486. `doi:10.1007/s11222-013-9444-y`.

$$\hat{X} = U\lambda^{\frac{1}{2}}V^T$$

Without regularization:

$$x_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js}$$

Regularized:

$$x_{ij} = \sum_{s=1}^{S} (\sqrt{\lambda_s} - \frac{\hat{\sigma}^2}{\sqrt{\lambda_s}}) u_{is} v_{js}$$

With: $\hat{\sigma}^2 = \frac{1}{P-S} \sum_{s=S+1}^{P} \lambda_s$

Confidence areas

- Josse J, Husson F (2011). “Multiple Imputation in PCA.” Advances in Data Analysis and Classification, 5(3), 231–246. doi:10.1007/s11634-011-0086-7

A multiple imputation method called MIPCA generates several imputed data sets. The imputed values for missing data differ. Variability across the various imputations reflects variability in the prediction of missing values. The variance of predictions is composed of two parts:

- variability in the estimated values of the PCA parameters
- variability due to noise.

Josse and Husson (2011) used a residuals bootstrap procedure to obtain the variance of parameters.

Model:

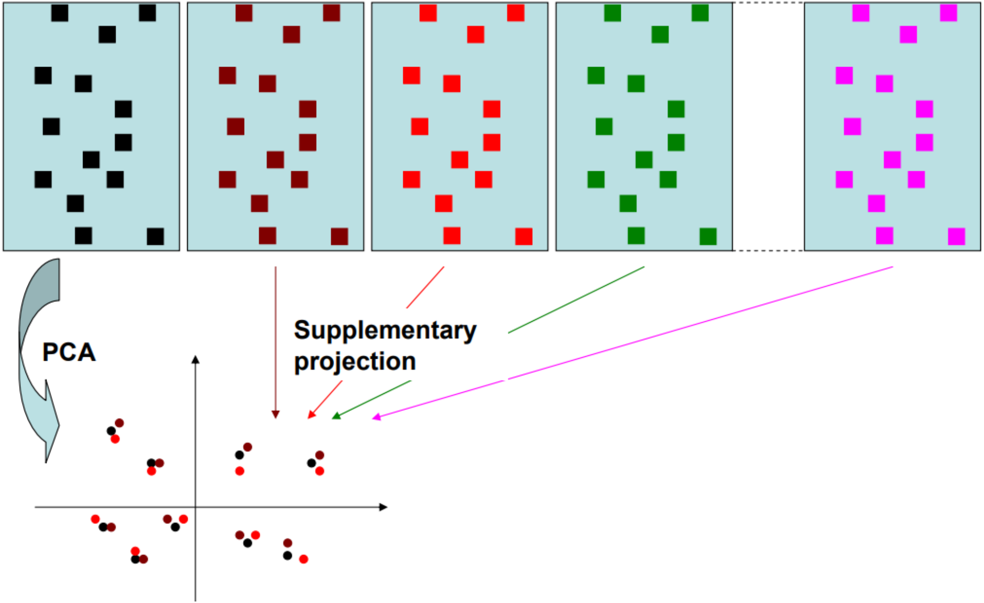
$$x_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} + \epsilon_{ij}$$

Residuals matrix: $\hat{\epsilon} = X - \hat{X}$

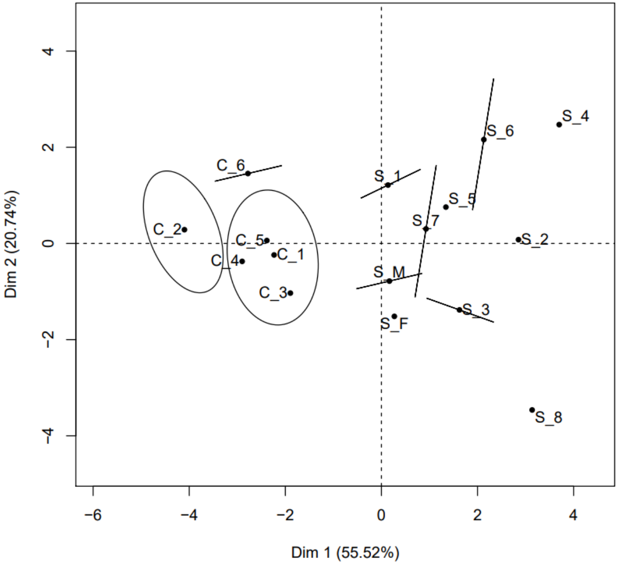
Create B bootstrap replicates of the data X^b, $b = 1, ...B$ by adding to the estimator $\hat{X}$, new matrices of residuals obtained by bootstraping the current residual matrix $\hat{\epsilon}$

Apply (regularized) iterative PCA algorithm to each new matrix X^b.

$\hat{X}^1, ..., \hat{X}^B$ represent the variability of the PCA parameters.


PCA
Supplementary projection

Supplementary projection



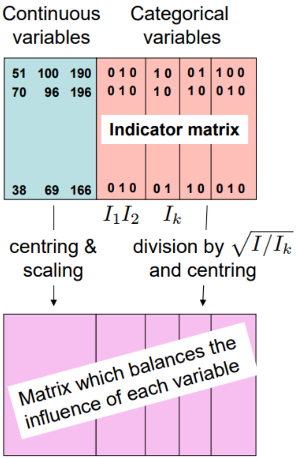
Multiple Correspondence Analysis (MCA)

MCA can be seen as the equivalent of PCA for categorical data. MCA is a PCA on a certain data matrix.

Josse J, Husson F (2012). “Handling Missing Values in Exploratory Multivariate Data Analysis Methods.”
The iterative MCA algorithm:

- in initialization step missing values are imputed by initial values such as the proportion of category.

Factorial Analysis for Mixed Data



Continuous variables

Categorical variables

Indicator matrix

centring & scaling

division by $\sqrt{I/I_k}$ and centring

Matrix which balances the influence of each variable

Verbanck M, Josse J, Husson F (2015). “Regularized PCA to Denoise and Visualise Data.” Statistics and Computing, 25(2), 471–486. doi:10.1007/s11222-013-9444-y.

Multiple Factor Analysis (MFA)

Husson F, Josse J (2013). “Handling Missing Values in Multiple Factor Analysis.” Food Quality and Preferences, 30(2), 77–85. doi:10.1016/j.foodqual.2013.04.013.

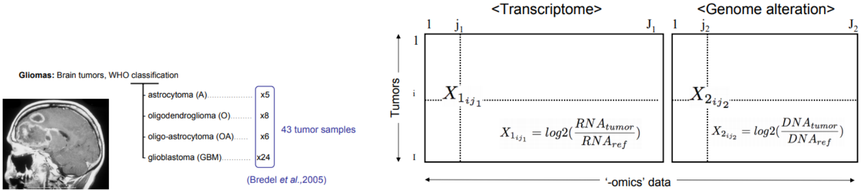
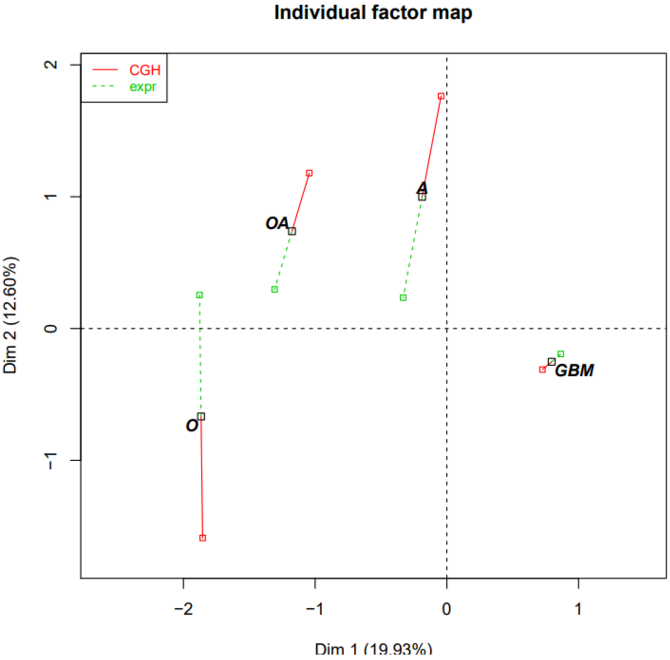


Figure 10: Four brain tumor types characterized by transcriptome and genome data.


Individual factor map
CGH
expr
OA
A
O
GBM
Dim 2 (12.60%)
Dim 1 (19.93%)
2
1
0
−1
−2
−1
0
1

Bibliography

- Josse J, Pagès J, Husson F (2009). "Gestion des Données Manquantes en Analyse en Composantes Principales." Journal de la Société Française de Statistique, 150(2), 28–51.
- Josse J, Husson F (2011a). "Multiple Imputation in PCA." Advances in Data Analysis and Classification, 5(3), 231–246. doi:10.1007/s11634-011-0086-7
- Josse J, Husson F (2012). "Handling Missing Values in Exploratory Multivariate Data Analysis Methods."
- Husson F, Josse J (2013). "Handling Missing Values in Multiple Factor Analysis." Food Quality and Preferences, 30(2), 77–85. doi:10.1016/j.foodqual.2013.04.013.
- Verbanck M, Josse J, Husson F (2015). "Regularized PCA to Denoise and Visualise Data." Statistics and Computing, 25(2), 471–486. doi:10.1007/s11222-013-9444-y.