Contribution to missing values & principal component methods

Julie Josse

Université Paris-Sud

Faculté des sciences d'Orsay
École doctorale de mathématiques Hadamard (ED 574)
Laboratoire de mathématique d'Orsay (UMR 8628 CNRS)

Mémoire présenté pour l'obtention du

Diplôme d'habilitation à diriger les recherches

Discipline : Mathématiques

par

Julie Josse

Contribution to missing values & principal component methods

Rapporteurs : Gilles Celeux
Jerry Friedman
Peter Hoff

Date de soutenance : 30 Août 2016.

Composition du jury :	Gilles Celeux	(Rapporteur)
	Jerry Friedman	(Rapporteur)
	Peter Hoff	(Président)
	Anne-Ruiz Gazen	(Examinateur)
	Pascal Massart	(Examinateur)
	Eric Moulines	(Examinateur)
	Fionn Murtagh	(Examinateur)
	Davy Paindaveine	(Examinateur)

Remerciements

Je souhaite tout d'abord remercier chaleureusement mes trois rapporteurs, Gilles Celeux, Peter Hoff et Jerôme Friedman pour avoir accepté d'évaluer ce travail. C'est un très grand honneur de les avoir dans mon jury.
Gilles, je te remercie sincèrement d'avoir tout mis en oeuvre pour qu'il me soit possible de soutenir mon habilitation dans un délai si court. Par ailleurs, je te remercie pour l'année passée à Orsay même si elle a été courte. C'est très agréable de travailler en ta compagnie, j'ai beaucoup appris à ton contact que ce soit d'un point de vue statistique mais aussi pour tes nombreuses expressions et j'espère que nous allons continuer à travailler ensemble.
Peter, je suis ravie de t'avoir dans mon jury car j'apprécie vraiment ton travail (je lis toujours tes articles avec très grand plaisir) et ta "sensibilité statistique", j'espère sincèrement que nous aurons l'occasion d'interagir dans le futur car je vois beaucoup d'intérêts communs. Je serais également très contente de te représenter lors de conférences.
Jerry, merci infiniment de me faire l'honneur de rapporter mon travail. Bien entendu, c'est vraiment un privilège de t'avoir dans mon jury. Evidemment tous tes travaux sont d'une richesse incroyable et sources d'inspiration pour des générations de statisticiens. J'ai toujours apprécié l'équilibre entre les idées, la théorie et la mise en oeuvre pratique de tes propositions. J'espère que tu apprécieras ce travail et que tu auras envie d'investiguer le sujet des données manquantes. Merci aussi pour l'accueil toujours chaleureux à Stanford.

Je remercie aussi fortement les membres de mon jury Anne-Ruiz Gazen, Pascal Massart, Eric Moulines, Fionn Murtagh et Davy Paindaveine.
Anne nous avons beaucoup d'intérêts communs et je pense que nous avons à gagner à échanger nos points de vue.
Davy, même si en premier lieu nos travaux sont plus éloignés, tu fais des exposés extraordinaires, j'aime beaucoup tes travaux et je vois aussi beaucoup d'échanges potentiels. D'autant que je me mets maintenant à utiliser les concepts de profondeurs...
Pascal, un grand grand merci pour ton écoute, tes conseils et pour m'avoir vraiment facilité la possibilité de soutenir à Orsay.
Eric, merci d'avoir accepté de faire partie du jury, j'espère que cela va être le début de superbes collaborations. En tout cas, je suis plus que ravie de rejoindre ton équipe et je suis convaincue que nous allons passer de bons moments scientifiques et humains. Je suis impatiente de travailler avec toi et j'espère des échanges fructueux de part et d'autre. Je suis impressionnée par ta culture scientifique.
Fionn, merci d'avoir accepté de participer à la soutenance, je suis très contente d'avoir un expert en analyse de données et je suis certaine que vous allez avoir de nombreux commentaires. Je suis toujours impressionnée par le spectre et la puissance de ces méthodes que je ne cesse de redécouvrir sous différents angles.

Je souhaite ensuite témoigner ma reconnaissance et mon amitié à Gérard Biau, qui m'a vraiment accompagnée ces dernières années et m'a toujours donné des conseils avisés. Tu m'as toujours soutenue depuis le début dans tous mes choix et je te remercie vraiment pour tout. Nous formons une bonne équipe je trouve et j'espère que nous allons collaborer scientifiquement dans le futur. En effet, j'ai une grande estime pour ton travail.
Merci aussi à Arnaud et François qui resteront mes piliers et qui m'ont toujours été d'une grande aide. François, j'apprécie toujours autant travailler avec toi et je suis toujours impressionnée par

tes remarques, ton intuition, ta franchise et ton travail en général que je trouve très important pour la communauté. De toute façon je ne pourrai je crois jamais assez te remercier et te témoigner mon amitié. Arnaud, arnaud... toujours là, prêt à rendre service, tu compteras toujours beaucoup. Toujours une petite pensée pour Eric Matzner sans qui je n'aurais jamais fait de recherche.

J'ai eu la chance ces dernières années de rencontrer de nouvelles personnes qui sont devenues des amis et qui m'ont beaucoup appris et inspirée, éveillée, nourris et je voulais les remercier pour cela. Merci à Susan Holmes, qui m'a vraiment permis de me sentir "chez moi" à Stanford en m'intégrant dans ses activités. Je suis impressionnée par ton parcours, ton énergie, ta motivation et tes intérêts scientifiques et je partage avec toi le goût de la bibliographie et des belles données bien entendu...J'ai hate de retravailler avec toi! Merci à Persi Diaconnis qui m'a appris à développer encore plus ma curiosité scientifique et qui prend toujours du temps pour discuter ou me donner des articles avec des explications pédagogiques. Merci à Trevor, Lynda pour tous les agréables moments que nous passons ensemble. Et enfin un grand merci à John, une vrai rencontre heureuse.

Quitte à écrire des remerciements (et à oublier sûrement de nombreuses personnes, je m'en excuse), je remercie mes amis tout simplement pour tous les bons moments passés: Cristof, Serena, Floriane, Suzane, Etienne, Kevin, Marine, Christophe, Laurent, Carole, Caro, Anne, Youen et bien d'autres. Bien entendu, je termine ces remerciements par mes parents chez qui j'ai rédigé ce document dans un cadre idyllique, dans le Finistère Penn-Ar-Bed bien entendu. Evidemment, vous savez que je suis infiniment reconnaissante pour tout et plus que chanceuse de vous avoir.

Je dédie ce manuscrit à ma nièce Lana.

Contents

Chapter 1

Introduction

This manuscript gives a quick overview of a selected part of my research since my PhD work that I realized as an associate professor in the statistics department of Agrocampus Ouest, an agronomy University (Rennes, Brittany). The aim of this document is to obtain the HdR (French's habilitation to supervise research), that would grant me the permission to officially supervise PhD students. My resume is also provided with this manuscript and the full-versions of the papers summarized in this document are available on my webpage:

http://juliejosse.com/

My research mainly focuses on the development of principal components methods to explore and visualize complex data (coming from different sources of information) and on the development of methods to handle missing values and to complete data. These two topics led me to deeply investigate the subjects of low rank matrix estimation and of estimating regularization parameters (such as the number of dimensions, variance of the noise, etc.) with and without missing values. This latter subject can be explained in part because my research has an applied vocation: my aim is to achieve some methodological advances that would be useful for the users. Thus, I pay attention in proposing a complete methodology to users with reasonable choices by default for the input parameters. In addition, I attach great importance to the transfer of my works through the development of statistical software.
My research problematics follow practical issues coming from my environment (such as other laboratories of the University) and investment in research is justified by the potential benefits in the fields of application. My main applications are in bio-sciences and food sciences especially in sensory analysis. Although the aim is often to find operational response to address the specific issues, it is important to conduct a methodological reflection beyond the initial problems, to extend the analysis to a general framework and set the problem in statistical methodology. After the methodological work, appropriate solutions could be proposed to the original questions.

The first part of this manuscript is named **A missing values tour with principal components methods**. First, in Chapter 2, I focus on performing exploratory principal components (PC) methods despite missing values *i.e.* estimating parameters scores and loadings to get biplot representations from an incomplete data set. In Section 2.2, I review a part of my PhD work that I did under the supervision of François Husson and Jérôme Pagès on performing principal component analysis (PCA) with missing values. After my PhD, I naturally continued in this line of research and worked with Stephane Dray (Dray and Josse, 2014) on comparison of different methods to perform PCA with missing values for plant ecology applications. Then, I focused on perfoming other multivariate techniques with missing values. With François Husson, we worked

on multiple factor analysis (MFA) (Husson and Josse, 2013) with sensory analysis applications, as described in Section 2.3 and with Marieke Timmerman and Henk Kiers (Josse, Timmerman, Kiers, and Smilde, 2013) on multi-level simultaneous component analysis. These methods are respectively dedicated to structured data with groups of variables and groups of individuals. To implement these works, we developed with François Husson an R package missMDA (Husson and Josse, 2015), illustrated in Section 2.4 and wrote a companion Journal of Statistical Software (JSS) paper (Josse and Husson, 2015). Software output is an incredibly powerful way to diffuse its work [1].
After studying PC methods with missing values, I deeply investigated the topic of using principal components methods as imputation methods. Indeed, the main algorithm which allows to perform PCA despite missing values outputs an imputed data set and consequently this method can be used to complete data. Even if at first this "imputation" may be seen as an aside to the method, it is in fact very valuable and indeed, the quality of imputation is usually high. This can be explained by the fact that imputation is based on the scores and loadings and thus takes into account similarities between individuals as well as relationships between variables using a rather small number of parameters due to the dimensionality reduction characteristics of the method. This accurate prediction of missing values explains the revival of interest in the topic, especially in the machine learning community with matrix completion problems such as the recommandation system Netflix (Netflix, 2009). In Chapter 3, I summarize the works done with our PhD student Vincent Audigier, (Audigier, Husson, and Josse, 2014a,b, Audigier, Husson, and Josse, 2015), where we built on this imputation property to suggest single imputation for mixed data (continuous, categorical) detailed in Section 3.1 and multiple imputation for continuous and categorical data described in Sections 3.3 and 3.4. One of our advantages was to use our knowledge of the "French exploratory data analysis methods" (Husson, Josse, and Saporta, 2016) such as multiple correspondence analysis (Greenacre and Blasius, 2006) in the framework of inference with missing values. It allowed us to feel a gap in this literature (Little and Rubin, 1987, 2002, van Buuren, 2012) where categorical and mixed data were not well handled as well as small sample size in comparison to large number of variables.

A second area that I investigated and that I review in this manuscript concerns **New practices in visualization with principal components methods**. Some of this research is the direct result of the works on missing values in principal components methods which raise many questions for the complete data analysis. First, from the point of view of the point estimates of the parameters, we suggested with missing values, regularized versions of the principal components methods to avoid overfitting issues. It led me to study in depth the rationale of such regularization in the complete case and the potential impact on the classical graphical outputs. Chapter 4 summarizes four contributions on this topic which are part of the more general framework of low rank matrix estimation methods. In Section 4.2.1, I detail the work with a PhD student Marie Verbanck (Verbanck, Josse, and Husson, 2013) where we studied a regularized PCA optimal in a low-noise asymptotic regime. We also highlighted a Bayesian interpretation of the regularization. Then, Section 4.3 presents the work with Sylvain Sardy (Josse and Sardy, 2015) where we suggested a finite sample estimator inspired by adaptive lasso (Zou, 2006a) which "adapts" to the data at hand thanks to two regularization parameters and consequently which can accurately estimate the signal in many scenarii. We also derived methods (Stein unbiased risk estimates) to appropriately select the regularized parameters. Section 4.4 summarizes the paper Josse and Wager (2015)

[1] It is so powerful that it may explain why some practices, even flawed, could still be in used. The popularity of the algorithm NIPALS (Wold and Lyttkens, 1969) to perform a PCA with missing values is probably an example of such issues. It has been a standard in chemometrics for a while despite its pitfalls (Josse and Husson, 2012).

with Stefan Wager where we defined a new framework for low rank matrix estimation that works by transforming noise models into regularizers via a parametric bootstrap. To implement these estimators, I developed an R package denoiseR with extensions to the missing values case and wrote an associated JSS paper with Sylvain Sardy and Stefan Wager (Josse, Sardy, and Wager, 2016). I illustrate the use of the package to regularized correspondence analysis (CA) in Section 4.6 using a document-words data on perfumes.
Next, after point estimates, Chapter 5 focuses on notions of variability of the parameters. Note that in the framework of missing values, we developed with François Husson in Josse and Husson (2011a) a strategy to visualize the variability due to missing values on the PCA output. Section 5.1 describes the paper Josse, Husson, and Wager (2014a) with Stefan Wager and François Husson aiming at studying variability in the complete case to enhance the interpretability of the PCA outputs. We studied confidence areas for fixed effect PCA using asymptotic regions and non-parametrics ones based on the bootstrap and the jackknife. With Jean-Baptiste Denis and other colleagues, in Josse, van Eeuwijk, Piepho, and Denis (2014b), we focused on these aspects using a Bayesian presentation of PCA detailed in Section 5.2. In this chapter, l insist on the practical implications and show how it may substantially affect the classical graphical outputs and the follow-up interpretations. Furthermore, l discuss the advantages of the approaches to answer practical questions in plant breeding with genotypes-environments (GE) data.

Finally, in a last part, I give a quick overview of my ongoing and future research activities on both principal component methods in Chapter 6 and on missing values in Chapter 7.

Below are the list of my papers described in this manuscript, and the list of papers that are not.

List of papers described in this manuscript.

- Josse, J. , Sardy, S. & Wager, S. (2016). denoiseR a package for low rank-matrix estimation. *On Arxiv: http://arxiv-web3.library.cornell.edu/abs/1310.6602.* Submitted.
- Josse, J. & Wager, S. Stable Autoencoding: A Flexible Framework for Regularized Low-Rank Matrix Estimation. *Journal of Machine Learning Research.*
- Audigier, V., Husson, F. & Josse, J. (2016) MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing.*
- Audigier, V., Josse, J. & Husson, F. (2015). Multiple imputation for continuous variable using Bayesian PCA. *Journal of Statistical Computation and Simulation.*
- Josse, J. & Sardy, S. (2015). Adaptive Shrinkage of singular values. *Statistics and Computing.*
- Josse, J. & Husson, F. (2015). missMDA a package to handle missing values in and with multivariate data analysis methods. *Journal of Statistical Software.*
- Audigier, V., Husson, F. & Josse, J. (2014). A principal components method to impute mixed data. *Advances in Data Analysis and Classification.*
- Josse, J., Wager, S. & Husson, F. (2014). Confidence areas for fixed-effects PCA. *Journal of Computational and Graphical Statistics.* To appear.
- Dray, S & Josse, J. (2014). Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology.* **216** (5), 657-667.

- Josse, J., van Eeuwijk, F., Piepho, H-P, Denis, J.B. (2014). Another look at Bayesian analysis of AMMI models for genotype-environment data. *Journal of Agricultural, Biological, and Environmental Statistics.* **19** (2), 240-257.

- Verbanck, M. & Josse, J. & Husson, F. (2015). Regularized PCA to denoise and visualise data. *Statistics and Computing.* **25** (2), 471-486.

- Josse, J., & Husson, F. (2013). Handling missing values in Multiple Factor Analysis. *Food Quality and Preferences.* **30 (2)**, 77-85.

- Josse, J., & Husson, F. (2013). Handling missing values in multivariate exploratory data analysis. *Journal of the French Statistical Society (SFdS).* **153 (2)**, 79-99. (Paper written for the best Ph.D doctoral thesis prize delivered by the French Statistical Society).

Other papers.

- Fujii H., Josse J., Tanioka M., Miyachi Y., Husson F. and Ono M. (2016). Regulatory T cells in melanoma revisited by a computational clustering of FOXP3+ T cell subpopulations. *Journal of Immunology.*

- Husson, F., Josse, J. & Saporta, G. (2016). Jan de Leeuw and the French school of data analysis. Submitted.

- Groenen, P. & Josse, J. (2016). Multinomial multiple correspondence analysis. *On Arxiv: http://arxiv.org/abs/1603.03174.*

- Josse, J. & Holmes, S (2015). Tests of independence and Beyond. *On Arxiv: http://arxiv-web3.library.cornell.edu/abs/1307.7383* Submitted.

- Josse, J. & Timmerman, M.E. & Kiers, H.A.L. (2013). Missing values in multi-level simultaneous component analysis. *Chemometrics and Intelligent Laboratory Systems.* **129**, 21-32.

- Josse, J., Chavent, M., Liquet, B. & Husson, F. (2012). Handling missing values with Regularized Iterative Multiple Correspondence Analysis. *Journal of Classification.* **29 (1)**, 91-116.

- Josse, J. & Husson, F. (2011). Selecting the number of components in PCA using cross-validation approximations. *Computational Statistics and Data Analysis.* **56 (6)**, 1869-1879.

- Josse, J., Pagès, J. & Husson, F. (2011). Multiple Imputation in PCA. *Advances in Data Analysis and Classification.* **5 (3)**, 231-246.

- Josse, J., Pagès, J. & Husson, F. (2009). Données manquantes en analyse en composantes principales. *Journal de la Société Française de Statitique (SFdS).* **150 (2)**, 28-51.

- Josse, J., Pagès, J. & Husson, F. (2008). Testing the significance of the RV coeffcient. *Computational Statistics and Data Analysis.* **53**, 82-91.

- Lê, S., Josse, J. & Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software.* **25 (1)**, 1-18.

In the pipeline.

- P. Mozharovskyi, Josse, J. & F. Husson. Elliptical distributions under corruption.
- P. Descloux, Josse, J. & S. Sardy. Model selection with Lasso when some data are missing in the design matrix.
- P. Sobczyk, Josse, J. & M. Bogdan. Bayesian dimensionality reduction with PCA using penalized. semi-integrated likelihood.

Part I

A missing values tour with principal components methods

Chapter 2

Principal components methods with missing values

2.1 Introduction

When starting a new project involving statistical analysis, it is first important to describe, explore and visualize the given data. Principal components methods can be useful in such cases, and several methods are available depending on the nature of the data: principal component analysis (PCA) for continuous data, multiple correspondence analysis (MCA) for categorical data (Gifi, 1990a, Greenacre and Blasius, 2006, Husson and Josse, 2014a), factorial analysis for mixed data (FAMD) both continuous and categorical data (Escofier, 1979, Kiers, 1991), and multiple factor analysis (MFA) for data structured in groups of variables (Escofier and Pagès, 2008, Pagès, 2015). These methods involve reducing data dimensionality in order to provide a subspace that best represents the data in the sense of maximizing the variability of the projected points. From a technical point of view, the core of all these methods is the singular values decomposition (SVD) of certain matrices with specific metrics [1]. Unfortunately, data sets often have missing values, and as many statistical methods, principal component methods have been developed for complete data.
The problem of missing values exists since the earliest attempts of exploiting data as a source of knowledge as it lies intrinsically in the process of obtaining, recording, and preparation of the data itself. Clearly, (citing Gertrude Mary Cox) "The best thing to do with missing values is not to have any" , but in the contemporary world of increasingly growing demand in statistical justification and amounts of accessible data this is not always the case, if not to say more. Main references on missing values include Schafer (1997), Little and Rubin (1987, 2002), van Buuren (2012), Carpenter and Kenward (2013) and (Gelman and Hill, 2007)[chp. 25].
Under the classical missing at random mechanism (MAR) assumption (Rubin, 1976), two main strategies are available to deal with missing values. The first one consists of adapting the statistical analysis so that it can be applied on an incomplete data set. For instance, the maximum likelihood (ML) estimator of a parameter of interest ψ can be obtained from incomplete data using an Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) and its standard error can be estimated using a supplemented Expectation-Maximization (SEM) algorithm (Meng and Rubin, 1991). The ML approach is tailored to a specific statistical method but can be difficult to establish and a specific algorithm has to be derived for each statistical method that we would like

[1] "all in all, doing a data analysis, in good mathematics, is simply searching eigenvectors; all the science (or the art) of it is just to find the right matrix to diagonalize", Benzécri, 1973

to apply. That is why the second strategy namely multiple imputation (MI) (Rubin, 1987, Little and Rubin, 1987, 2002) seems to have taken the lead. The principle of MI consists of predicting M different values for each missing value, which leads to M imputed data sets. The variability across the imputations reflects the variance of the prediction of each missing entry. Then, MI consists of performing the statistical analysis on each imputed data set to estimate the parameter ψ and consists of combining the results $(\widehat{\psi}_m)_{1\leq m\leq M}$ to provide a unique estimation for ψ and for its associated variability using Rubin's rules (Rubin, 1987). This ensures that the variance of the estimator is not underestimated and thus good coverage properties.
What is important is that the aim of both approaches is to estimate as well as possible the parameters and their variance despite missing values, *i.e.* taking into account the supplement variability due to missing values [2] and not to impute as accurately as possible the entries. This aim is also pursued when performing principal components methods with missing values. Since, the core of many methods is the PCA of specific matrices, I start by reviewing how to perform PCA with missing values.

2.2 PCA with missing values

PCA in the complete case 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, assumed to be centered by columns, in the least squares sense (with $\| \bullet \|$ the Frobenius norm):

$$\text{argmin}_\mu \left\{ \|X_{n\times p} - \mu_{n\times p}\|_2^2 : \text{rank}\,(\mu) \leq S \right\}$$

The PCA solution (Eckart and Young, 1936) is the truncated singular value decomposition (SVD) of the matrix $X = U\Lambda^{1/2}V'$ at the order S, namely

$$\hat{\mu}_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} \tag{2.1}$$

PCA has been extended for an incomplete data set by replacing the least squares criterion by a weighted least squares

$$\text{argmin}_\mu \left\{ \|W \odot (X - \mu)\|_2^2 : \text{rank}\,(\mu) \leq S \right\} \tag{2.2}$$

where $W_{ij} = 0$ when X_{ij} is missing and 1 otherwise and $\odot$ stands for the elementwise multiplication. The aim is to estimate the PCA parameters despite the missing entries which are skipped (parameters are estimated on the observed values, which makes sense). In contrast to the complete case, there is no explicit solution to minimize criterion (2.2) and it is necessary to resort to iterative algorithms. Many algorithms have been proposed and re-discovered in the literature under different names and in different fields (see Josse and Husson (2012) for references), including the *iterative PCA* algorithm suggested by Kiers (1997) and detailed in Josse and Husson (2012). It performs:

1. initialization $\ell = 0$: substitute missing values with initial values such as the mean of the variables with non-missing entries, the imputed matrix is denoted X^0. Calculate M^0, the matrix of the vector containing the mean of the variables of X^0, repeated in each row of M^0.

2. step $\ell \geq 1$:

[2] Supplement makes sense especially when thinking about missing values as a particular case of a small sample size data. With less data the variance is larger.

(a) perform the SVD of $(X^\ell - M^\ell)$ to estimate quantities Λ^ℓ and U^ℓ, V^ℓ.

(b) compute the fitted matrix $\hat{\mu}^\ell_{ij} = \sum_{s=1}^{S} \sqrt{\lambda^\ell_s} u^\ell_{is} v^\ell_{js}$ and define the new imputed data as $X^\ell = W \odot (X - M^\ell) + (\mathbf{1} - W) \odot \hat{\mu}^\ell$, where $\mathbf{1_{n\times p}}$ is a matrix filled with ones. The observed values are the same and the missing ones are replaced by the fitted values.

(c) $X^\ell = X^\ell + M^\ell$, compute M^ℓ from the new completed matrix X^ℓ [3]

3. steps (2.a), (2.b) and (2.c) are repeated until the change in the imputed matrix falls below a predefined threshold $\sum_{ij}(\hat{\mu}^{\ell-1}_{ij} - \hat{\mu}^\ell_{ij})^2 \leq \varepsilon$, with ε equal to 10^{-6} for example.

The *iterative PCA* algorithm is a "genuine EM" [4] algorithm associated with the Gaussian noise model where the data is generated as a structure of low rank corrupted by noise:

$$\begin{aligned} X_{n\times p} &= \mu_{n\times p} + \varepsilon_{n\times p} \\ x_{ij} &= \sum_{s=1}^{S} \sqrt{d_s} q_{is} r_{js} + \varepsilon_{ij}, \;\; \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{aligned} \qquad (2.3)$$

This EM interpretation reinforces the use of the *iterative PCA algorithm* and implies that the estimated scores (in $U\Lambda^{1/2}$) and loadings (in V) are the maximum likelihood estimates of the parameters. However, in practice the *iterative PCA algorithm* rapidly suffers from overfitting when data are noisy and/or there are many missing values. This means that the fitting error is low whereas the prediction error is large. This issue can be solved by early stopping or by regularization and Josse *et al.* (2009) suggested the following strategy where they replaced step (2.b) of the algorithm with an imputation with $\hat{\mu}^\ell_{ij} = \sum_{s=1}^{S} \left(\sqrt{\lambda^\ell_s} - \frac{(\hat{\sigma}^2)^\ell}{\sqrt{\lambda^\ell_s}} \right) u^\ell_{is} v^\ell_{js}$ with the noise variance estimated as $(\hat{\sigma}^2)^\ell = \frac{\| (X^{\ell-1} - M^{\ell-1}) - U^\ell (\Lambda^\ell)^{1/2} (V^\ell)' \|^2}{np - nS - pS + S^2}$. I will give more details about the shape of the regularization in Section 4.2.1.

Note that the procedure requires the *a priori* knowledge of S, the number of dimensions (the dimensions are no more nested in the incomplete case). We suggested in Josse and Husson (2012) the use of cross-validation (CV) and generalized-cross validation to avoid computational costs of CV. This latter criterion can be simply expressed as $\text{GCV}_S = \frac{np \left\| W \odot (X - \hat{\mu}^S) \right\|_2^2}{(np - |NA| - (nS + pS - S^2))^2}$, with |NA|, the number of missing cells. Thus, it corresponds to the observed residuals sum of squares penalized by the complexity of the model. In Josse and Husson (2012), we interpreted the number of independent parameters as the trace of the projection matrix obtained when writing PCA as a nonlinear smoother $\text{vec}(\hat{\mu}) = P\text{vec}(X)$.

After getting point estimates of the parameters from an incomplete data, a natural step is to study variability. The missing values framework makes this step even more obvious than usual. Indeed, what confidence should be given to the results obtained from an incomplete data set? It is always possible to carry-out an algorithm and to obtain results with graphical outputs in PCA. However, should they be trusted? In Josse and Husson (2011a), we proposed a multiple imputation method based on PCA. After the *iterative PCA algorithm*, it is possible to generate M imputed values for each missing entries, simply by drawing from their predictive distribution $x^b_{ij} \sim \mathcal{N}(\hat{\mu}_{ij}, \hat{\sigma}^2)$, for $b = 1, ..., M$. It leads to M imputed data set. However, this imputation is qualified as "improper"

[3] After each imputation step, means are modified and must be updated. This step is sometimes overlooked in software. Not to forget it, we could write complete PCA as minimizing $\|X - M - \mu\|_2^2$.

[4] I insist on genuine EM, because often EM does not impute data but only sufficient statistics, consequently one can be skeptic with this claim. However, it can be shown that it is an EM algorithm (Josse, Pagès, and Husson, 2009).

by Rubin (1987) since it considers $\hat{\mu}_{ij}$ as fixed. For *proper* MI, we used a residuals bootstrap strategy to generate $x_{ij}^b \sim \mathcal{N}(\hat{\mu}_{ij}^b, \hat{\sigma}^{2^b})$. Then, we suggested to visualize the supplement variability due to missing values either using supplementary projections or using Procrustes rotation (Gower and Dijksterhuis, 2004a) as illustrated in Section 2.4.

2.3 Multi-tables MFA with missing values

Let us now extend the case of one data table to multi-tables. Heterogeneous data coming from different information sources can be organized as illustrated in Figure 2.1 with the rows described by several groups of variables. Multiple factor analysis (MFA) (Pagès, 2015) is a principal components method which aims at studying the similarities between rows from a multidimensional point of view, as well as the correlation between variables (same objectives as for PCA) but also at highlighting similarities and differences between groups of variables and studying what is common to groups and what is specific. Other multi-blocks methods are available (Kroonenberg, 2008) to answer such questions and to visualize results. MFA easily handles groups of continuous and categorical variables as well as contingency tables. One of the aims of MFA is to balance the influence of the groups in such a way that no group (with many correlated variables for instance) dominates the first dimension of variability. To do so, for each group of variables a PCA is performed and then each value in the group is divided by the square root of the first eigen value. Then, MFA consists in performing a global PCA on the concatenated weighted data. One of the rationale is the same as in standardized PCA where variables are weighted to have the same influence in the analysis; here it can be seen as an extension to groups of variables where the first singular value plays the role of the standard deviation. More MFA properties can be found in Pagès (2015).

The risk of being confronted with missing values increases when there are many sources of information. In addition, with multi-table data, we can have specific patterns of missing values involving missing rows per sub-table, as illustrated in Figure 2.1. Since the core of MFA is a weighted PCA, we extended the algorithm to handle missing values in PCA (Section 2.2) and developped in Husson and Josse (2013) a *(regularized) iterative MFA* algorithm (illustrated Section2.4) to perform MFA using an incomplete data set. It also alternates steps of estimation of the parameters and imputation of the missing values but takes into account details specific to MFA, namely the precise weighting. The method was developed to answer practical issues in sensory analysis.

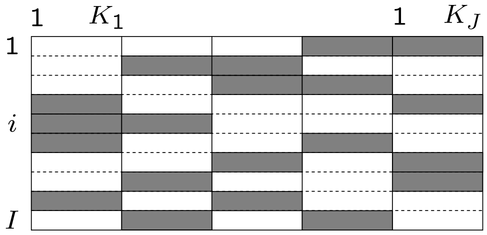


Figure 2.1: Missing rows in tables. $K_1, \ldots, K_J$ are the number of variables in groups 1, …, J.

Sensory analysis

The aim of sensory analysis can be to characterize products, to position them relative to each other, to link hedonic appreciation to organoleptic composition of the products, etc. In sensory analysis, as in many fields, missing values are part of the daily work. One particular example includes missing values planned by design. More precisely, let us consider that a set of judges access a set of products, such as wines. Panelists assign ratings ranging from 0 “I do not like the wine at all” to 10 “I really like the wine". Given the difficulty in assessing a large number of wines (due to a sensory fatigue, a saturation phenomenon after tasting many wines even if the people split), each judge evaluates only a subset of products. This subset is determined by experimental design (Balanced Incomplete Block type). This induces missing (incomplete) information in the data to be analyzed.

In Husson and Josse (2013), we focused on the important issue of the maximum number of products that can be assessed during an evaluation task. We highlighted that it may be a better strategy to assess less products with many judges than all the products with a smaller number of judges.

2.4 Illustration with the R package missMDA

2.4.1 PCA with missing values

Let’s consider the ecological GLOPNET data (Wright, *et al.*, 2004) from Dray and Josse (2014). In ecology, several projects aim to build worldwide repositories by compiling data from preexisting databases. However, due to the wide heterogeneity of measurement methods and research objectives, these huge data sets are often characterized by an extraordinarily high number of missing values. More details about the ecological questions can be found in Dray and Josse (2014). The data describe 2494 species with 6 quantitative variables (LMA (leaf mass per area) LL (leaf lifespan), Amass (photosynthetic assimilation), Nmass (leaf nitrogen), Pmass (leaf phosphorus), Rmass (dark respiration rate), 1 categorical variable (the biome) and have 53.38% of missing entries. Only 72 species have complete information for the 6 traits and the proportion of missing values varied between 4.97 % (LMA) to 89.01 % (Rmass).

Here, I do not dwell on the specific kind of missing values (patterns and reasons for their occurrence) exhibited although it is a first step crucial in a typical analysis. Indeed, methods to deal with missing values and their properties depend on this [5]. Graphical displays, such as the one implemented in the R package *VIM* (Templ, Alfons, Kowarik, and Prantner, 2013), are very useful study the missing values. In Josse and Husson (2015), we suggested using PC methods for such an issue. It works as follows: the data are coded with “o” for observed values and “m” for the missing ones, then a multiple correspondence analysis is performed on the coded data set to study associations between missing and observed entries. Figure 2.4.1 shows that missing values on Nmass are often associated with missing values on LMA. This graph may help in an exploratory analysis to understand the missing values.

Then, PCA despite missing values is obtained using the package missMDA (Husson and Josse, 2015) with the following lines of code:

```
> library(missMDA)
> nbdim <- estim_ncpPCA(don, method.cv = "Kfold", nbsim = 100) # estimate S with cross-validation
> res.comp <- imputePCA(don, ncp = nbdim) # regularized iterative PCA algorithm
> res.pca <- PCA(res.comp, quali.sup=1) # PCA on the completed data
> res.MIPCA <- MIPCA(don, ncp=2) # multiple imputation
> plot(res.MIPCA, choice= "ind.supp"); plot(res.MIPCA, choice= "var")
```

Figures 2.3 highlights that the representation of the species is completely different when imputing with the mean. This first PCA axis obtained by *iterative PCA* corresponds to the "leaf economic spectrum" and separates species with potential for quick returns for investment with high values

[5] The suggested methods have been developed under the MCAR and MAR framework (Little and Rubin, 1987, 2002)

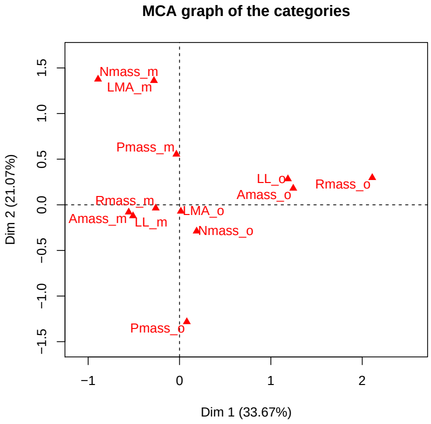


Figure 2.2: Use of MCA to explore the missing values.

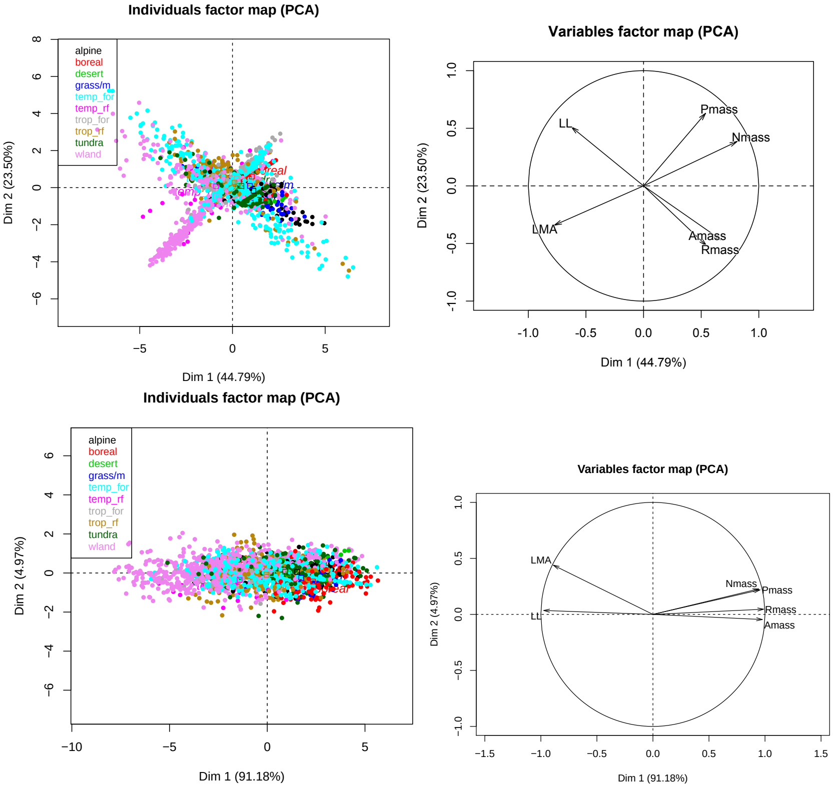


Figure 2.3: Top: species and traits representations obtained by the PCA on the data where missing values were imputed by the mean of the variables. Bottom: results of the *regularized iterative PCA*.

for Nmass, Amass, Rmass and Pmass and low values for LL and LMA (right part) from species with slow returns on the left part.

Figure 2.4 shows the projection of the variables of each imputed data sets obtained by the multiple imputation strategy as supplementary elements onto the reference configuration obtained with the *(regularized) iterative PCA algorithm*. It allows to visualize the position of the variables with different missing values prediction.

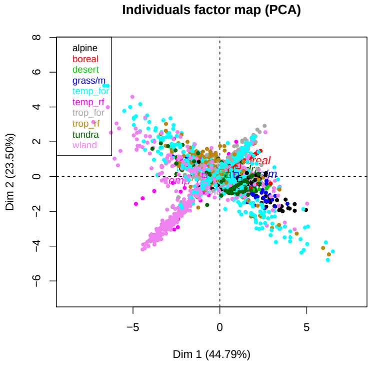


Figure 2.4: Multiple imputation: supplementary projection of the M imputed variables on the PCA configuration of Figure 2.3 (right). Supplementary means that we compute the inner-product between the variables and the axes (loadings) of the reference configuration. An variable without any missing values is projected onto its corresponding point whereas a variable with missing entries is projected around its initial position.

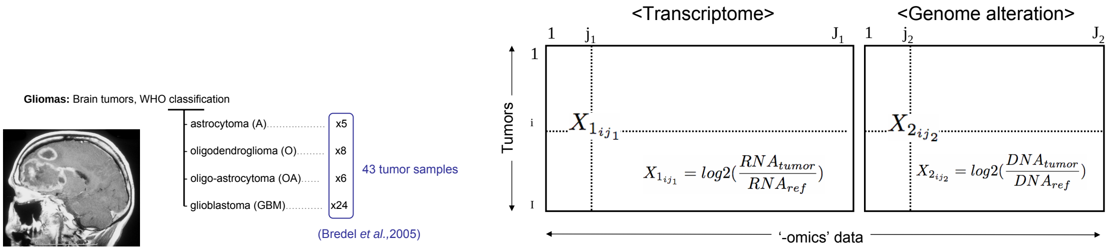


Figure 2.5: Four brain tumor types characterized by transcriptome and genome data.

These areas of variability are valuable to assess the impact of the missing values on the analysis. Here, the plot shows that the variability across different imputations is small and a user can interpret the PCA results with confidence.

2.4.2 MFA with missing values

A biological example studies 43 brain tumors of 4 different types defined by the standard world health organization (WHO) classification (O, oligodendrogliomas; A, astrocytomas; OA, mixed oligo-astrocytomas and GBM, glioblastomas) using data both at the transcriptome level (with 356 continuous variables for expression data) and at the genome level (with 76 continuous variables for CGH data) as illustrated in Figure 2.5.. Ten samples were not available for the expression data. More details about the data and the results of the analysis can be found in de Tayrac, Lê, Aubry, Mosser, and Husson (2009), though note that they deleted samples with missing values to perform their analysis. To compare the information brought by each group, a first step in the analysis is to quantify the relationship between the two sets of variables using coefficients of association and then decide if the association is significant by using a test. In Josse and Holmes (2015) we detailed such measures and tests, such as the RV coefficient, which is an extension of the correlation coefficient for groups. To apply MFA with missing values, we run the following lines of code:

```
> data("gene", package = "missMDA")
> res.impute <- imputeMFA(gene[, -1], group = c(76, 356),  type = rep("s", 2), ncp = 2)
> res.mfa <- MFA(cbind.data.frame(gene[, 1], res.impute$completeObs),  group = c(1, 76, 356),
 type = c("n", rep("s", 2)), name.group = c("WHO", "CGH", "expr"), num.group.sup = 1)
```

Note that the first column of the data corresponds to the nominal variable (coded "n") with the tumor types and is added as supplementary information, only used afterward to help enhance the interpretation; continuous variables are “standardized” (coded "s").
Figure 2.6 on the left shows that that the first dimension is common to both groups whereas the second dimension is mainly due to the group CGH. We are also able to say that this first dimension is close to the first principal component of each group since the values of the L_g coefficients (corresponding to the coordinates) are close to one. Figure 2.6 on the right is the “compromise",

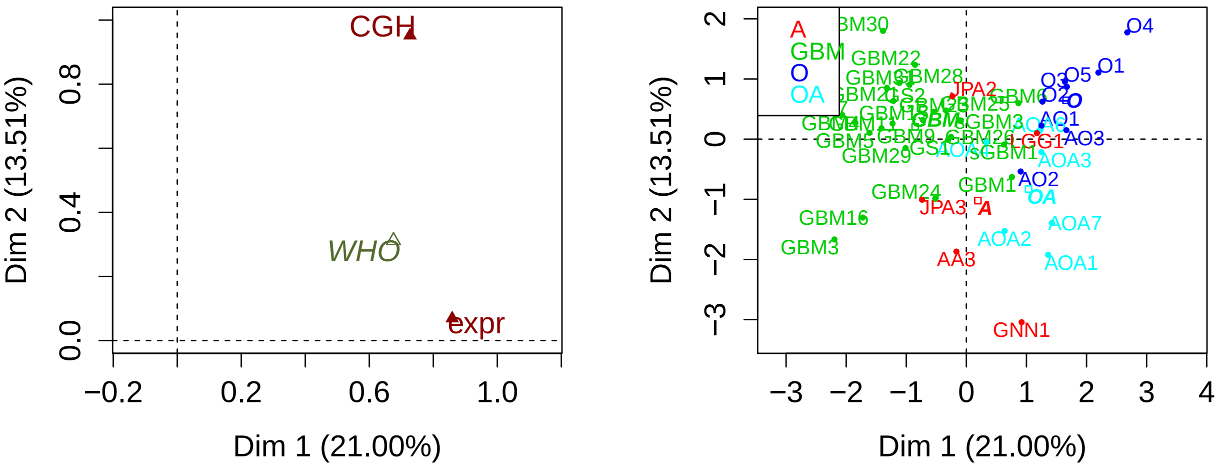


Figure 2.6: MFA groups representation (left) and compromise representation of the tumors (right).

the result of the global PCA and shows that the first dimension of variability opposes the glioblastomas tumors to the lower grade tumors and that the second dimension opposes tumors O to the tumors OA and A. Since as mentioned previously, the first dimension is common to both groups of variables, it means that both the expression data and the CGH data permits to separate the glioblastomas to the other tumors. On the other hand, only the CGH data permits to see differences between the tumor O and the tumors OA and A. Thus, it shows what is common and what is specific to each group. Figure 2.7 on the left is the correlation circle to study the correlation between all the variables and shows that the expression data is much more one-dimensional whereas the CGH data is represented at least on two dimensions (red arrows are hidden by the green arrows). This method also allows to compare the information of both groups at the observation level with the “partial" representation represented Figure 2.7 on the right. The tumor GBM29 is represented using only its expression data (in green) and using only its CGH data (in red). The black dot is at the barycenter of both red and green points and represents the tumor GBM29 taking into account all the data. This tumor is peculiar in the sense that when taking its CGH data, this individual is on the side of the dangerous tumors (small coordinates on the first axis) whereas it is on the side of the other tumors when considering its expression data (positive coordinates on the first axis). There is no consensus for this individual between the two sources of information and it may require more investigation to understand why.

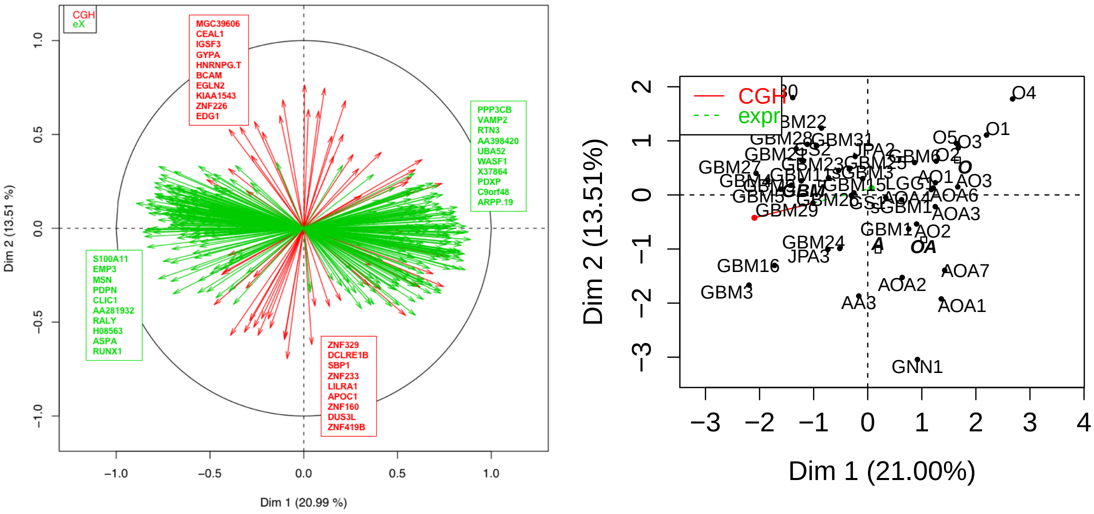


Figure 2.7: MFA variables representation (left) and a “partial" sample (right).

Chapter 3

Single and multiple imputation with principal components methods

After working on performing principal component (PC) methods despite missing entries (getting the graphical outputs and estimating the parameters), I investigated the use of these methods as tools of single and multiple imputation and compared them to the state of the art methods. PC methods sum-up the similarities between the individuals and the relationships between variables using a small number of synthetic variables (principal components) and synthetic observations (loadings). They reduce the dimensionality of the data and require a small number of parameters while keeping as much as possible the information of the data. This trade-off between complexity of the model and preservation of the data structure is particularly relevant to analyse high dimensional data, but also appealing to perform imputation. In the sequel, we will see that these methods are indeed very powerful to impute data.
In this chapter, I describe the works done with François Husson and our PhD student Vincent Audigier on single imputation for mixed data (Audigier *et al.*, 2014a) and multiple imputation for continuous (Audigier *et al.*, 2014b) and categorical data (Audigier *et al.*, 2015). The aim of single imputation is to complete as well as possible the data whereas the aim of multiple imputation is to provide valid inference (point estimates and variances) for quantities of interest.

3.1 Single imputation for mixed data

3.1.1 Introduction

The practice of single imputation, *i.e.* replacing missing values by plausible values can be dangerous [1]. Indeed, if a statistical analysis is performed on an imputed data set, the variance of the estimator is too short since the variability of the missing values prediction is not taken into account. That's why, multiple imputation is recommended. Nevertheless, single imputation, is still paid attention in the statistical literature. This can be appropriate when one just needs to complete a single data set or when no inference is required. In addition, single imputation is a first step to multiple imputation.

[1] "The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases." (Dempster and Rubin, 1983)

There is a huge literature on imputation methods for continuous data or categorical data (Little and Rubin, 1987, 2002, van Buuren, 2012). For mixed data (continuous and categorical), there are fewer options. A solution consists of transforming the categorical variables into dummy variables and then using an imputation based on the assumption of a joint Gaussian distribution for the variables. Alternatively, Schafer (1997) suggested imputing with the general location model which requires computing all the entries of the multi-way cross table and thus is useful only for a small number of categorical variables. Instead of a joint modeling approach, van Buuren, Boshuizen, and Knook (1999) suggested a conditional modeling one which consists of specifying one model for each variable. Kropko, Goodrich, Gelman, and Hill (2014) compared and discussed both approaches in the framework of multiple imputation. Recently, Stekhoven and Bühlmann (2012) proposed a condional imputation method based on random forest which actually serves as a reference and outperforms other imputation methods especially in difficult cases such as high-dimensionality and complex interaction between variables.

3.1.2 Our proposition

We propose an imputation based on the PC method factorial analysis for mixed data (FAMD) (Escofier, 1979, Pagès, 2015), also known as PCAMIX (Kiers, 1991). The method first consists of coding the data as illustrated in Figure 3.1. Categorical variables are transformed into dummy variables and concatenated with the continuous variables. Then, each continuous variable is standardized (centered and divided by its standard deviation) and each dummy variable is divided by the squared root of the proportion of individuals taking the associated category $\sqrt{n_k/n}$ for category k. FAMD consists of performing a principal component analysis (PCA) on this weighted matrix. This specific weighting induces balance in the influence of both variable types. It is close to the rationale of scaling in PCA which gives the same weight for each variable in the analysis; here the specific weights ensure that all continuous and categorical variables play the same role. In addition, the principal components, denoted F_s for $s = 1, ..., S$ maximize the link between the

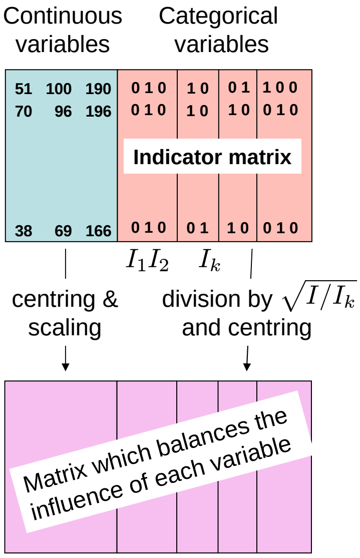


Figure 3.1: Coding the data in factorial analysis for mixed data.

continuous and categorical variables in the following sense:

$$F_s = \underset{F_s \in \mathbb{R}^n}{\arg\max} \sum_{j=1}^{p_{cont}} r^2(F_1, X_j) + \sum_{jc=1}^{p_{cat}} \eta^2(F_1, X_{jc}),$$

with the constraint that F_s is orthogonal to F_s' for all $s' < s$ and with X_j being the variable j, p_{cont} the number of continuous variables, p_{cat} the number of categorical variables, r^2 the square of the correlation coefficient and η^2 the square of the correlation ratio (in an analysis of variance sense). This formulation highlights that FAMD can be seen as the counterpart of PCA for mixed data.

The algorithm which performs FAMD with missing values and which is used to impute data (Audigier *et al.*, 2014a) is inspired by the *iterative PCA algorithm* 2.2 and can be sketched as follows:

1. Initialization: imputation with the means (for continuous data) and proportions (for the dummy variables). Compute weights: standard deviations and column margins (n_k).
2. Iterate until convergence:
   (a) perform PCA on the completed weighted data to estimate the parameters U, Λ, V
   (b) impute the missing values with the fitted values $\hat{\mu} = U\Lambda^{1/2}V'$ using S dimensions
   (c) update the mean, standard deviations and column margins [2]

This method [3] is illustrated in Figure 3.1.2. At the end of the algorithm, the indicator matrix is imputed with real numbers but that verify the property that the sum of the values for one individual and one categorical variable is equal to 1. Thus, the values can be seen as degrees of membership to the categories and we can impute with the most "plausible category" to end-up with a completed mixed data (we impute snore with the value "no" for the second individual). However, negative values can occurs in the imputed data set and consequently, imputed values can not be seen as probabilities.

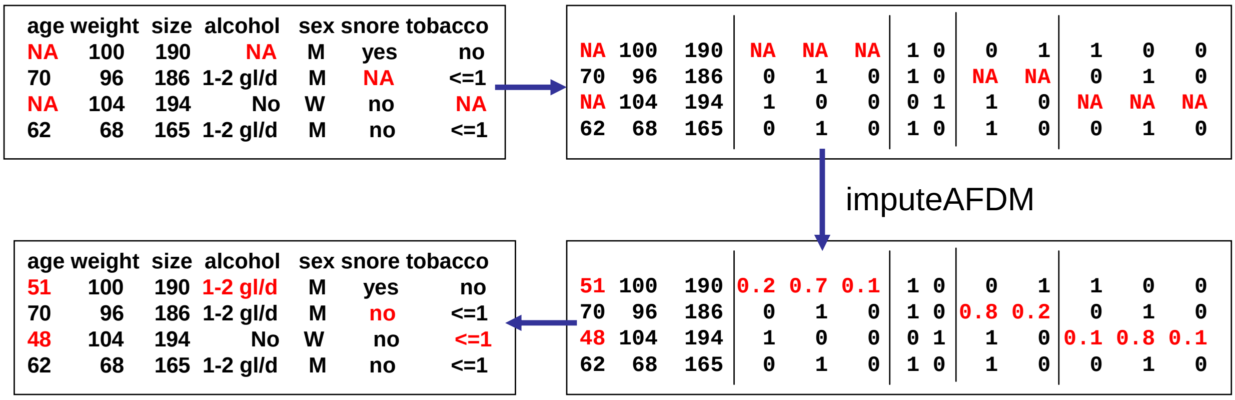
3.1.3 Results

Imputation with FAMD was tested in Audigier *et al.* (2014a) fairly extensively in simulation studies with synthetic and realistic data that rely on complete empirical data sets in which pre-specified percentages of entries have been made missing (10%, 20 %, 30 % of missing completely at random MCAR). Real data aim at covering many situations whereas synthetic data at controlling some aspects. However, note that simulating categorical variables is not (at least for me) trivial [4]. The method was mainly compared with the random forest (RF) based method (Stekhoven and Bühlmann, 2012) and comparison were made in terms of root mean squared errors of prediction for the continuous variables and the proportion of falsely classified entries for categorical variables. The RF method works as follows: it starts with random imputations, then fits a RF on the observed part of the first variable say X_1 using $X_2, ..., X_p$ and updates the prediction for the missing values in X_1. Then, it does the same thing for each variable, one at a time, (a RF of

[2] After each imputations, means, standard deviations are modified hence the need to recenter and rescale the data in such a way that the variables have still the same weight in the analysis. Note that in standardized PCA, this step is not often included in software outputs, leading to erroneous results.

[3] In practice, a regularized version of the algorithm is implemented in the same way as in PCA in 2.2.

[4] Hence the need for a generative model for categorical data, a topic that arouses my curiosity for a while!



age	weight	size	alcohol	sex	snore	tobacco
NA	100	190	NA	M	yes	no
70	96	186	1-2 gl/d	M	NA	<=1
NA	104	194	No	W	no	NA
62	68	165	1-2 gl/d	M	no	<=1

NA	100	190	NA	NA	NA	1	0	0	1	1	0	0
70	96	186	0	1	0	1	0	NA	NA	0	1	0
NA	104	194	1	0	0	0	1	1	0	NA	NA	NA
62	68	165	0	1	0	1	0	1	0	0	1	0

51	100	190	0.2	0.7	0.1	1	0	0	1	1	0	0
70	96	186	0	1	0	1	0	0.8	0.2	0	1	0
48	104	194	1	0	0	0	1	1	0	0.1	0.8	0.1
62	68	165	0	1	0	1	0	1	0	0	1	0

age	weight	size	alcohol	sex	snore	tobacco
51	100	190	1-2 gl/d	M	yes	no
70	96	186	1-2 gl/d	M	no	<=1
48	104	194	No	W	no	<=1
62	68	165	1-2 gl/d	M	no	<=1

Figure 3.2: Illustration of the *regularized iterative FAMD* algorithm on the incomplete snorena data. Top left: the incomplete data; top right: coding of the data; bottom right: the imputed data; bottom left: the original data set completed.

X_2 with $X_1, X_3, .., X_p$ as predictors, etc.) and cycles through all the variables and iterates until stabilization of the predictions.

The properties of the imputation inherit from the properties of the methods. Consequently, imputation with RF has smallest prediction errors when data have nonlinear relationships and complex interactions. Note however, that the *regularized iterative FAMD algorithm* can handle such situations by cutting continuous variables into categories for the former case and introducing an additional variable in the dataset that corresponds to the interaction for the latter. Imputation based on FAMD shows better performances when there are highly linear relationships between continuous variables and also for the imputation of categorical variables and especially for rare category. This can be seen as a great advantage and is the results of the specific FAMD weighting which gives importance to rare categories. The method is also very competitive in term of computational time in comparison to the RF method (as shown in Section 3.4). However, note that the FAMD methods requires to select S the number of components from the incomplete data set. Even in the complete case, there are no methods for selecting this number. In practice, cross-validation is used and the sensitivity of the analysis to this choice was assessed in Audigier *et al.* (2014a) (often it is very stable).

Finally, a major property/drawback of the imputation using RF can be highlighted, which again is related to the inherent characteristics of the method. RF breaks down for small sample size and missing at random (MAR) cases [5]. An extreme case is shown in Figure 3.1.3 where missing values have been inserted in Feat 1, 2 and 3 when Feat 4 is greater that 7 [6]. An imputation based on a PC method can extrapolate whereas imputation based on RF not.

3.2 Multiple imputation

Multiple imputation (MI) consists of replacing each missing value by M plausible values which leads to M imputed data sets. The missing values are predicted using an *imputation model*. In

[5] MAR means that the probability that a value is missing is unrelated to the value itself but can be related to the other variables; however conditionally to the other variables, the probability to have a missing values is the same for all individuals.

[6] Note that in this case, it is impossible to distinguish between MAR and missing non at random (MNAR) values where the probability of missing values depends on the value itself.

```
         Feat1 Feat2 Feat3 Feat4 Feat5
C1           1     1     1     1     1
C2           1     1     1     1     1
C3           2     2     2     2     2
C4           2     2     2     2     2
C5           3     3     3     3     3
C6           3     3     3     3     3
C7           4     4     4     4     4
C8           4     4     4     4     4
C9           5     5     5     5     5
C10          5     5     5     5     5
C11          6     6     6     6     6
C12          6     6     6     6     6
C13          7     7     7     7     7
C14          7     7     7     7     7
Igor         8    NA    NA     8     8
Frank        8    NA    NA     8     8
Bertrand     9    NA    NA     9     9
Alex         9    NA    NA     9     9
Yohann      10    NA    NA    10    10
Jean        10    NA    NA    10    10
```

```
         Feat1 Feat2 Feat3 Feat4 Feat5
C1           1   1.0  1.00     1     1
C2           1   1.0  1.00     1     1
C3           2   2.0  2.00     2     2
C4           2   2.0  2.00     2     2
C5           3   3.0  3.00     3     3
C6           3   3.0  3.00     3     3
C7           4   4.0  4.00     4     4
C8           4   4.0  4.00     4     4
C9           5   5.0  5.00     5     5
C10          5   5.0  5.00     5     5
C11          6   6.0  6.00     6     6
C12          6   6.0  6.00     6     6
C13          7   7.0  7.00     7     7
C14          7   7.0  7.00     7     7
Igor         8   6.87 6.87     8     8
Frank        8   6.87 6.87     8     8
Bertrand     9   6.87 6.87     9     9
Alex         9   6.87 6.87     9     9
Yohann      10   6.87 6.87    10    10
Jean        10   6.87 6.87    10    10
```

⇒ with Random Forests

```
         Feat1 Feat2 Feat3 Feat4 Feat5
C1           1     1     1     1     1
C2           1     1     1     1     1
C3           2     2     2     2     2
C4           2     2     2     2     2
C5           3     3     3     3     3
C6           3     3     3     3     3
C7           4     4     4     4     4
C8           4     4     4     4     4
C9           5     5     5     5     5
C10          5     5     5     5     5
C11          6     6     6     6     6
C12          6     6     6     6     6
C13          7     7     7     7     7
C14          7     7     7     7     7
Igor         8     8     8     8     8
Frank        8     8     8     8     8
Bertrand     9     9     9     9     9
Alex         9     9     9     9     9
Yohann      10    10    10    10    10
Jean        10    10    10    10    10
```

⇒ with PCA

order to perform what is called a *proper* MI (Rubin, 1987), the uncertainty of the *imputation model* parameters must be reflected from one imputation to the next (each imputed data is obtained using a different estimated parameter of the imputation model). Then, MI consists of estimating the parameter ψ of a statistical method (called the *analysis model*) on each imputed data set. Note that several *analysis models* can be applied to a same multiply imputed data set, which is usually used as a strong argument in favor of multiple imputation. Lastly, the $(\widehat{\psi}_m)_{1\leq m\leq M}$ estimates of the parameters are pooled to provide a unique estimation for ψ and for its associated variability (which is composed of the within and between variance) using Rubin's rules (Rubin, 1987). More precisely $\hat{\psi} = \frac{1}{M}\sum_{m=1}^{M}\hat{\psi}_m$ and $\widehat{Var}(\hat{\psi}) = \frac{1}{M}\sum_{m=1}^{M}\widehat{Var}\left(\hat{\psi}_m\right) + \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\psi}_m - \hat{\psi}\right)^2$. This ensures that the variance of the estimator appropriately takes into account the supplement variability due to missing values. To get valid inferences for a large variety of *analysis models*, a desirable property is to have an *imputation model* at least "as general" as the *analysis models* (taking into account as many associations between variables as possible for instance).
There are two classical ways of performing multiple imputation either with an explicit joint model for the data or using a conditional modeling approach (van Buuren, 2012) where one model is defined for each variable with missing data and variables are successively imputed using these models. The conditional approach is often considered as more flexible than the joint modeling one since it can be tailored to the data at hand and. Indeed, it easily deals with interactions and variables of different nature (binary, ordinal, categorical...) by fitting a logistic regression on a variable, a multinomial regression on another one, etc. In addition many statistical models are conditional ones. The conditional approach also see imputed values as draw from a joint distribution even if the joint distribution does not exist (the separate models are not compatible Besag (1974)). Even if its theoretical properties are more obscur, the conditional modeling strategy seems to take the lead and "work quite well" in practice. However the approach can be tedious with many variables. More discussion about advantages and drawbacks of both strategy can be found in Kropko, Goodrich, Gelman, and Hill (2014).

In the next two sections, I describe new methods for multiple imputation of continuous data based on principal component analysis and of categorical data based on multiple correspondence analysis (MCA). MCA is the counterpart of PCA for categorical data. These methods can be seen as member of the "joint modeling" family. The uncertainty of the parameters of the "imputation model" is reflected using a Bayesian strategy for the former case and a non-parametric bootstrap one for the latter case. Through extensive simulations study, the methods were assessed and compared to the reference methods as well as to the latest works on the topic. Contrary to their competitors, they can be easily used on data sets where the number of individuals is less than the number of variables and when the variables are highly correlated and with a high number of categories per variable. In addition, they provide unbiased point estimates of quantities of interest, such as an expectation, a regression coefficient, with a smaller mean squared error. Furthermore, the widths of the confidence intervals built for the quantities of interest are often smaller while ensuring a valid coverage. Finally, they have the great advantage of beeing less time consuming on large data than the other methods.

3.3 Multiple imputation for continuous data with PCA

3.3.1 Competitors

The classical *joint modeling* multiple imputation method, assumes that the data follow a joint multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$. The procedure to get M imputed data can be carried-out as follows using an expectation-maximization bootstrap algorithm as implemented in the R-package Amelia (Honaker, King, and Blackwell, 2014, Honaker, King, and Blackwell, 2011):

1. M bootstrap samples (rows are resampled) are generated from the incomplete data X: X^1, ... , X^M and an EM algorithm is performed on each incomplete data to estimate the parameters of the joint distribution: $(\hat{\mu}^1, \hat{\Sigma}^1)$, ... , $(\hat{\mu}^M, \hat{\Sigma}^M)$

2. Imputation: missing values are drawn x_{ij}^m from their predictive distribution (the conditional distribution) given the observed values and the estimated parameters $\left(\hat{\mu}^m, \hat{\Sigma}^m\right)$.

The classical *fully conditional modeling* (FCM) multiple imputation methods impute using a regression model per variable. The procedure can be carried-out as follows using the algorithm implemented in the R-package mice (Van Buuren, 2014, van Buuren and Groothuis-Oudshoorn, 2011):

1. Initial imputation: missing values are completed with the mean of each variable

2. For a variable j

   2.1 $(\boldsymbol{\beta}^{-j}, \sigma^{-j})$ drawn from their posterior distribution

   2.2 Imputation: stochastic regression x_{ij}^m drawn from $\mathcal{N}\left(X_{-j}\boldsymbol{\beta}^{-j}, \sigma^{-j}\right)$

3. Cycling through variables

4. Repeat M times steps 2 and 3

Note that the uncertainty of the parameters of the imputation models is reflected via a Bayesian approach (and not a bootstrap one as for the previous algorithm although it is also possible). In this situation with one regression per variable, the conditional approach also remains to drawing from $\mathcal{N}(\mu, \Sigma)$. Both the joint and the conditional strategies encounter difficulties with highly correlated variables or when $n < p$. It would be possible to shrink the covariance matrix for joint modeling ($\Sigma + k\mathbb{I}$, but how to select k?) whereas ridge regressions or predictors selection can be applied for conditional modeling but the task is huge with many variables.

3.3.2 Our proposition

In Audigier *et al.* (2014b), we proposed a multiple imputation method based on PCA. Single imputation is achieved using the *regularized iterative PCA* algorithm described in Section 2.2. To perform a *proper* multiple imputation (Rubin, 1987), we reflect the uncertainty of the parameters from one imputation to the next using a Bayesian approach.
More precisely, for the complete case, we use the following modeling:

$$\text{Model (4.2): } x_{ij} = \sum_{s=1}^{S} \mu_{ij}^{(s)} + \varepsilon_{ij} = \sum_{s=1}^{S} \sqrt{d_s} q_{is} r_{js} + \varepsilon_{ij}\ ,\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

$$\text{Prior: } \mu_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2)$$

$$\text{Posterior: } \left(\mu_{ij}^{(s)} | x_{ij}^{(s)}\right) = \mathcal{N}(\Phi_s x_{ij}^{(s)}, \sigma^2 \Phi_s) \text{ with } \Phi_s = \frac{\tau_s^2}{\tau_s^2 + \sigma^2}$$

A prior distribution on each cell of the data matrix per dimension is assumed $\mu_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2)$ and the posterior distribution is obtained by combining the likelihood and the priors. Its expectation depends on unknown quantities and consequently, we use an empirical Bayesian approach by considering the distribution of $x_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2 + \frac{1}{\min(n-1;p)}\sigma^2)$ (obtained using model (4.2)) and estimating τ_s^2 as the maximum likelihood to obtain:

$$\hat{\tau_s}^2 \quad = \quad \left(\frac{1}{np}\lambda_s - \frac{1}{\min(n-1;p)}\sigma^2\right)$$

Consequently we estiamte the shrinkage term as $\hat{\Phi}_s = \frac{\left(\frac{1}{np}\lambda_s - \frac{1}{\min(n-1;p)}\hat{\sigma}^2\right)}{\frac{1}{np}\lambda_s} = \frac{\lambda_s - \frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\lambda_s}$. The estimated expectation of the posterior is denoted $\hat{\mu}_{ij}^{rPCA}$. Note that this estimator can be seen as a truncated version at order S of the one suggested in Efron and Morris (1972). More details about this Bayesian treatment can be found in Section 4.2.2.
Then, with missing values, we can use a data augmentation (DA) algorithm (Tanner and Wong, 1987), which can be seen as the counterpart of EM, to draw from the observed posterior distribution. DA algorithm iterates between the two steps:

(I) impute data from the current parameters and the observed data,

(P) draw new parameters from the posterior distribution computed using the imputed data and a prior distribution on the model's parameters.

In addition of providing a posterior distribution of the parameters from an incomplete data set, the DA algorithm can also be straightforwardly used to get multiple imputed data sets. To do so, after a burn-in step, we keep M independent draws leading to M imputed data sets. Thus, an imputed data set is saved at regular intervals.
Inspired by this strategy, we define our multiple imputation method named BayesMIPCA by iterating the two following steps:

(I) given $\tilde{\mu}_{ij}$ and $\hat{\sigma}^2$, impute the missing values x_{ij}^m by a draw from the predictive distribution $\mathcal{N}\left(\tilde{\mu}_{ij}, \hat{\sigma}^2\right)$

(P) drawing $\tilde{\mu}_{ij}$ from its posterior distribution $\mathcal{N}\left(\hat{\mu}_{ij}^{rPCA}, \frac{\hat{\sigma}^2 \sum_s \hat{\phi}_s}{min(n-1,p)}\right)$ where $\hat{\mu}_{ij}^{rPCA}, \hat{\sigma}^2$ and $(\hat{\phi}_s)_{1 \leq S}$ are calculated from the completed data set obtained at step (I).

Note that the estimates of ϕ and σ, that appear in the posterior distributions of μ, are updated by their maximum likelihood estimates in step (P), and are not fixed. Thus, the BayesMIPCA algorithm can be viewed as a marriage between a DA algorithm and an EM algorithm with unknown convergence properties.

3.3.3 Results

To assess the multiple imputation method based on PCA, we conducted an extensive simulation study. I present here one of the numerous results to show how multiple imputation methods are usually compared. Data are simulated from $\mathcal{N}_p(\mu, \Sigma)$ with a two-block covariance structure (two blocks of independent variables which are correlated within groups with the coefficient correlation ρ set to 0.3 or 0.9). This ensures that the number of underlying components S is 2 (S is fixed for simulations). Then, we vary n (30 or 200) and p (6 or 60). Next, 10% or 30% of missing values using a MCAR mechanism are inserted and $M = 20$ imputed data are generated using the 3 methods presented (called here Joint Gaussian, Conditional Gaussian and BayesMIPCA). On each imputed data set, we estimate parameters such as the mean of the first variable, a regression coefficient and a correlation coefficient between two variables $\psi_1 = \mathbb{E}[X_1], \psi_2 = \beta_1, \psi_3 = \rho$ and their variance using Rubin's rule (Section 3.2). Since the 3 methods give unbiased estimates (which is already a good step), we reporte their confidence intervals width and coverage in In Table 3.3.3.

	parameters				confidence interval width			coverage		
	n	p	ρ	%	*Joint G.*	*Cond. G.*	*BayesMIPCA*	*Joint G.*	*Cond. G.*	*BayesMIPCA*
1	30	6	0.3	0.1	0.803	0.805	0.781	0.955	0.953	0.950
2	30	6	0.3	0.3		1.010	0.898		0.971	0.949
3	30	6	0.9	0.1	0.763	0.759	0.756	0.952	0.95	0.949
4	30	6	0.9	0.3		0.818	0.783		0.965	0.953
5	30	60	0.3	0.1			0.775			0.955
6	30	60	0.3	0.3			0.864			0.952
7	30	60	0.9	0.1			0.742			0.953
8	30	60	0.9	0.3			0.759			0.954
9	200	6	0.3	0.1	0.291	0.294	0.292	0.947	0.947	0.946
10	200	6	0.3	0.3	0.328	0.334	0.325	0.954	0.959	0.952
11	200	6	0.9	0.1	0.281	0.281	0.281	0.953	0.95	0.952
12	200	6	0.9	0.3	0.288	0.289	0.288	0.948	0.951	0.951
13	200	60	0.3	0.1		0.304	0.289		0.957	0.945
14	200	60	0.3	0.3		0.384	0.313		0.981	0.958
15	200	60	0.9	0.1		0.282	0.279		0.951	0.948
16	200	60	0.9	0.3		0.296	0.283		0.958	0.952

Table 3.1: Median of the confidence intervals width and the 95% coverage over the 1000 simulations for different simulations' configurations. In addition, when an algorithm fails on a configuration, no result is given.

Both the Joint Gaussian and Conditional Gaussian approaches collapse when it is expected (when regressions have difficulties). On the contrary, BayesMIPCA can be applied on data sets of various kinds: when the collinearity between variables is weak or strong, when the rate of missing data is large or small, the number of individuals less than or greater than the number of variables. All the

algorithms give valid coverage, close to 95% but BayesMIPCA has often the smallest confidence interval width. This can be explained by the properties of the imputation model. Indeed, PCA is a dimensionality reduction method used to isolate the relevant information of a data set. This makes it very stable and implies that the imputation from a table to another does not change much. The between-variability is lower than for the other methods, which explains that the confidence intervals are shorter. Note also that since the imputation is based on PCA, it is particularly well fitted to situations where the relationships between variables are linear and when the data can be considered as generated from a PCA model (4.2). Simulations were also performed when the data do not have a low rank structure and although MIPCA did not collapse, its performances were indeed weaker and undercoverage issues appeared.
To wrap-up, BayesMIPCA has many advantages and is a flexible alternative to the classical MI procedures. However, it requires tuning a parameter which is the number of dimensions S. In practice, S is selected beforehand using cross-validation (Josse and Husson, 2011b), then the MIPCA is run for a specified S. It means that the uncertainty on S is not reflected in analysis and that all the sources of variability are not taken into account to predict missing values, although in the simulations the impact seemed tiny. Within a Bayesian framework, a solution would be to put a prior distribution on S.

3.4 Multiple imputation for categorical data with MCA

3.4.1 Competitors

Suggesting a MI method for high dimensional data and especially categorical data is a very challenging task. Indeed, the imputation models suffer from estimation issues as the number of parameters quickly grows with large number of categories and variables. Ideally, one would like an imputation model which requires a moderate number of parameters, while preserving as much as possible the relationships between variables. MI with a joint modeling approach is possible via log-linear models (Schafer, 1997). However, this encounters difficulties when there are many variables. An alternative can be the use of a latent class model (Vermunt, van Ginkel, van der Ark, and Sijtsma, 2008) or its nonparametric Bayesian extension suggested in Si and Reiter (2013) using Dirichlet process mixture of products of multinomial distributions. The practice of transforming categorical into dummies and then using Gaussian based methods is also still used, although not respecting the nature of the variables. Imputation can also be obtained by successively drawing from fully conditional (FC) distributions (van Buuren *et al.*, 1999, van Buuren, 2007) using logistic or multinomial logit regressions or either random forests. More details on these methods can be found in Little and Rubin (1987, 2002), van Buuren (2012) and Carpenter and Kenward (2013). Audigier *et al.* (2015) also described in depth their strength and weaknesses and compared them with many simulations.

3.4.2 Our proposition

Multiple correspondence analysis (MCA) has been successfully applied to visualize the relationship between categorical variables in many fields such as social sciences, marketing, health, psychology, educational research, political science, genetics, etc. (Greenacre and Blasius, 2006, Husson and Josse, 2014b). There are different was to define MCA, historically, it was defined as the correspondence analysis (CA) of the indicator matrix (Lebart and Saporta, 2014), but it could also be presented as the CA of the burt matrix (cross-tabulating the categorical variables), or as the minimization of an homogeneity criterion (de Leeuw and Rijckevorsel, 1980, Gifi, 1990b, Michailidis and de Leeuw, 1998), etc. MCA is also known under different names such as HOMALS or the "Dutch version" (Husson *et al.*, 2016). Let X being the indicator matrix of dummy variables, MCA can be obtained by the generalized SVD (Greenacre, 1984a) of the triplet (data, columns weight (rows metric), rows weight (column metric)) $\left(X, \frac{1}{p}\left(D_{\Sigma}\right)^{-1}, R\right)$ with D_{Σ}, the diagonal matrix with the margins of the categories (n_k) and R usually equals to $(1/n\mathbb{I}_n)$. GSVD implies that $X = U\Lambda^{1/2}V'$ with $U'RU = \mathbb{I}$ and $V'\frac{1}{p}D_{\Sigma}^{-1}V = \mathbb{I}$. [7] The choice of weights ensures the properties of the method such as the Chi-square interpretation of the distances between rows as well as the fact that the principal components are “new” variables the most related to the set of variables, with the relationship measured here by the squared correlation ratio of analysis of variance η^2 as defined in Section 3.1 [8].

It is possible to single impute categorical data with MCA using the *iterative MCA algorithm* (Josse, Chavent, Liquet, and Husson, 2012) which aims at performing MCA with missing values and which is very similar to the *iterative PCA algorithm* 2.2. Then, to perform MI with MCA, the uncertainty of the *imputation model* (here MCA) parameters has to be reflected from one imputation to the next and for that we used in Audigier *et al.* (2015) a non-parametric bootstrap approach. In fact, we did not proceed to explicit resampling of the rows thanks to the previous definition with the GSVD but instead defined M weightings $(R_m)_{1\leq m\leq M}$ for the rows. Then, with incomplete X, MCA parameters are estimated with *the iterative MCA algorithm* using the GSVD of $\left(X, \frac{1}{J}\left(D_{\Sigma}\right)^{-1}, R_m\right)$ for $1 \leq m \leq M$. Consequently, M sets $(U_{n\times S}, \Lambda^{1/2}_{S\times S}, V'_{p\times S})$ of parameters are obtained and used to impute the data as illustrated on the top of Figure 3.4.2. Then, categories are drawn from a multinomial distribution according to the imputed values as illustrated on the bottom of Figure 3.4.2 to reflect the distribution of the data. Note that a post-processing step may be required if negative values occur.

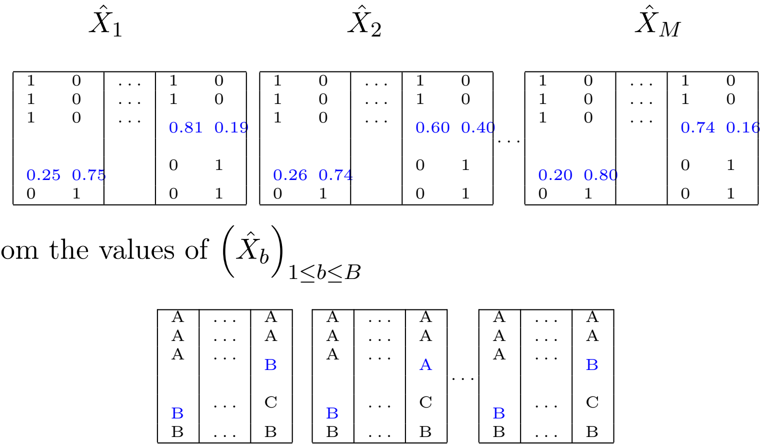


Figure 3.3: Multiple imputation with MCA.

3.4.3 Results

This new MI method was assessed in terms of the quality of inferences as well as computational time with real and simulated data covering many situations [9]. The principle is the same as the one

[7] Benzecri, 1976: "Doing a data analysis, in good mathematics, is simply searching eigenvectors, all the science of it (the art) is just to find the right matrix to diagonalize".

[8] MCA can also be obtained using FAMD with only categorical variables.

[9] The task of designing simulations could be a topic in itself - real data are often used as "population data" to have a true value for the parameters ψ and then samples are drawn and missing values are added (Audigier *et al.*,

detailed in Section 3.3.3. From a data with missing values, M imputed data are generated with a MI method, then a *statistical model* (such as a logistic regression) is performed on each imputed data and the $\hat{\psi}_m$ are combined with Rubin's rules. The MI methods are compared regarding, biais, confidence intervals width and coverage.

The main results can be wrapped-up as follows: MI using the loglinear model performs well on the two data sets where it can be performed (small enough). Its coverages are close to the nominal level, the biases are close to zero, and the confidence interval widths are small. MI using the non-parametric version of the latent class model (DPMPM) (Si and Reiter, 2013) performs well (coverage close to 95%) in some scenario and not so well in others. More on this heratic behavior in Si and Reiter (2013), Vidotto, Kapteijn, and Vermunt (2014). The FC using logistic regressions encounters difficulties on the data with large number of categories whereas FC using random forests performs well except with small sample sizes. Concerning MI using MCA, all the coverages observed are satisfying. The confidence interval widths are of the same order of magnitude than the other MI methods. In addition, the method can be applied whatever the number of categories per variables, the number of variables or the number of individuals. Thus, it appears to be the easiest method to impute categorical data. MI methods can be time consuming and the running time of the algorithms could be considered as an important property of a MI method from a practical point of view. Table 3.4.3 gathers the times required to perform MI with $M = 5$ and 20% of missing values. The tables speaks by itself. Note that once again, S is assumed known and selected using cross-validation beforehand in practice.

	Titanic	Galetas	Income
MIMCA	2.750	8.972	**58.729**
Loglinear	0.740	4.597	NA
DPMPM	10.854	17.414	143.652
FCS logistic	4.781	38.016	881.188
FCS forests	265.771	112.987	6329.514

Table 3.2: Time consumed (in seconds) to impute data sets. Calculation has been performed on an Intel®Core$^{\text{TM}}$2 Duo CPU E7500, running Ubuntu 12.04 LTS equipped with 3 GB ram.

	Titanic	Galetas	Income
Number of individuals	2201	1192	6876
Number of variables	4	4	14
Number of categories	≤ 4	≤ 11	≤ 9

2015).

Part II

New practices in visualization with principal components methods

Chapter 4

Regularized principal components methods: low rank matrix estimation

4.1 Introduction

4.1.1 Genotype-Environment data

I started working on this topic through a collaboration with J.B Denis from the applied mathematics and informatics team (MIA) of INRA [1]. Let's consider the following Genotype-Environment (GE) data (Royo, Rodriguez, and Romagosa, 1993) where $I = 16$ *Triticale* lines (or genotypes) were sown in $J = 10$ experiments (environments) across Spain. Each experiment consisted of a randomized complete blocks experiment with 4 replicates, although only the means across the replicates are available. The first six genotypes corresponds to the so-called "complete" type, while the next eight are of the "substituted" type. Finally, two check genotypes were included. The difference between the complete and substituted lines resid in the fact that the substituted genotypes have in their full set of chromosomes one chromosome stemming from rye replaced by a chromosome stemming from wheat. This substitution is thought to produce day length insensitivity, and reduce drought stress tolerance and acidity tolerance. One of the key objectives of the data analysis is to determine the clustering of the varieties in view of selecting stable genotypes across the environment.

One of the main models used (Gauch, 1988, Gauch and Zobel, 1996, Cornelius, Crossa, and Seyedsadr, 1996) to analyze such data is a biadditive model (Denis and Pazman, 1999, Denis and Gower, 1996) also known as AMMI (additive main effects and multiplicative interaction) model. Such model is defined as follows:

$$y_{ij} = g + \alpha_i + \beta_j + \sum_{s=1}^{S} \sqrt{d_s} r_{is} q_{js} + \varepsilon_{ij} \quad \text{with} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{4.1}$$

where y_{ij} is the response for the category i of the first factor and the category j of the second factor, g is the grand mean, $(\alpha_i)_{i=1,\dots,I}$ and $(\beta_j)_{j=1,\dots,J}$ correspond to the main effect parameters and $\left(\sum_{s=1}^{S} d_s r_{is} q_{js}\right)_{i=1,\dots,I;j=1,\dots,J}$ model the interaction. The least squares estimates for the interaction terms are given by the SVD of the residuals, *i.e.*, of the row and column centered data matrix as illustrated Figure 4.1 [2]. From a computational point of view, this model is similar to the PCA one, the main difference being that the linear part only includes the grand mean and column

[1] French National Institute for Agricultural Research.

[2] The shape of the data influences a lot the way we think about the associated methods.

main effect in PCA. Such models are useful for studying the interaction between two factors when no replication is available.

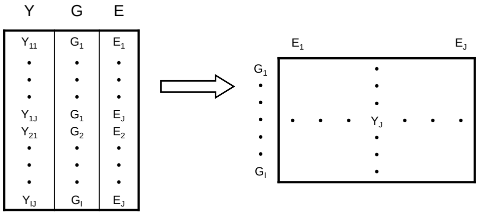


Figure 4.1: From an analysis of variance table to a "PCA" table for GE data.

PCA (Figure 4.2) is often used to visualise the interaction and to see if the genotypes have different performances in different environments. Here, there is a clear opposition between complete (C) and substitute (S) genotypes on the first dimension. Genotypes SM and SF which are the check genotypes are close to the center of gravity of the cloud which is in expected since they are known for their wide adaptivity. More interpretation is given in Josse *et al.* (2014b).

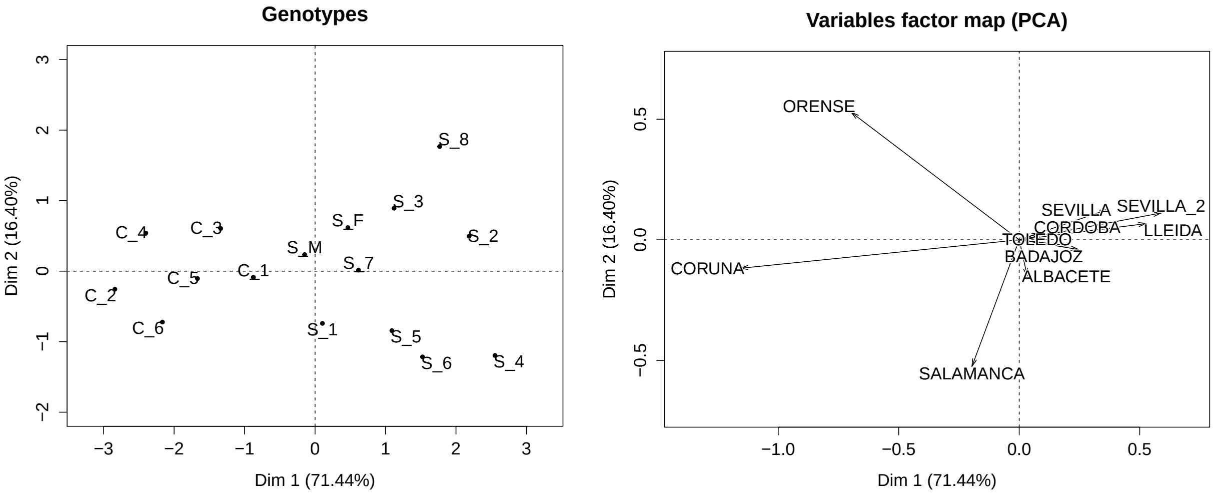


Figure 4.2: PCA on the GE data.

It is common to complement the analysis by performing k-means algorithms or hierarchical trees on the first S principal components (PC) of PCA. The rationale is to first denoise the data to get a more stable clustering. Figure 4.3 represents the results of such analysis.

In the sequel of this chapter, I detail how regularization schemes can improve the faithfulness of these commonly used PC analysis graphical outputs.

4.1.2 Fixed-effects PCA model

The previous analysis can be put into the more general framework of low rank matrix estimation. Low-rank matrix estimation plays a key role in many scientific and engineering tasks, including collaborative filtering (Koren, Bell, and Volinsky, 2009), genome-wide studies (Leek and Storey, 2007, Price, Patterson, Plenge, Weinblatt, Shadick, and Reich, 2006), and magnetic resonance imaging (Candès, Sing-Long, and Trzasko, 2013, Lustig, Donoho, Santos, and Pauly, 2008). Low-

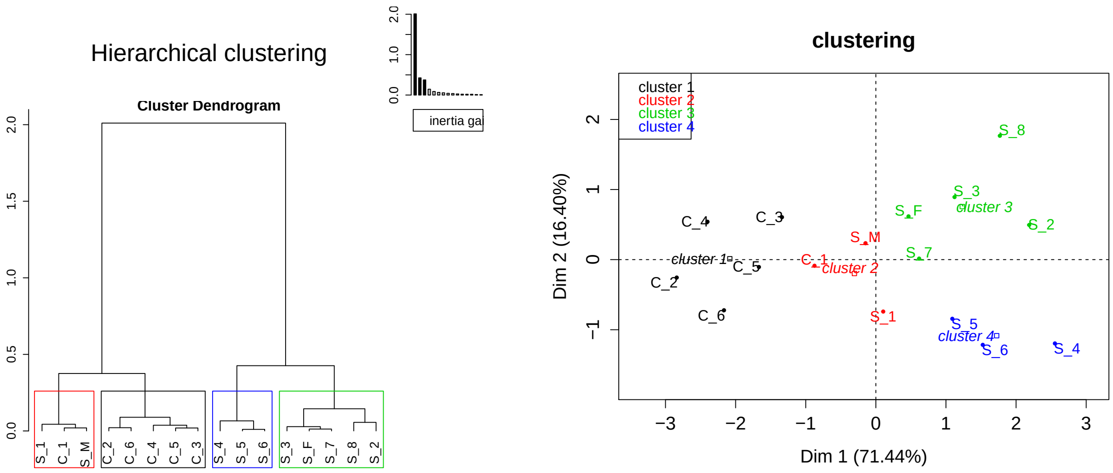


Figure 4.3: Clustering on the PC for the GE data.

rank procedures are often motivated by the simple model (4.2):

$$\begin{aligned} X_{n\times p} &= \mu_{n\times p} + \varepsilon_{n\times p} \text{ with } \mu \text{ of rank } S \\ x_{ij} &= \sum_{s=1}^{S} \sqrt{d_s} q_{is} r_{js} + \varepsilon_{ij}, \;\; \varepsilon_{ij} \sim \mathcal{N}(0,\sigma^2) \end{aligned} \tag{4.2}$$

The statistical aim is to estimate the signal μ from the noisy data X. Such a model is also known as the fixed-effects model (Caussinus, 1986) in PCA since the structure μ is considered as fixed. In this model, both the rows and the columns of X are non-random, and randomness comes only from measurement error. It is in agreement with many examples where the rows of X represent specific subjects whom we want to study, and the columns of X represent different features of the subjects. For instance, in sensory analysis, individuals can be products, such as chocolates, and variables can be sensory descriptors, such as bitterness, sweetness, etc and each cell of the data matrix is the average of the scores given by judges. The aim is to study these specific products and not others (they are not interchangeable, and are not a random sample drawn from a population of individuals). It thus makes sense to estimate the individual parameters (q_s) and to study the graphical representations of both rows and variables.

The natural inferential framework associated with (4.2) is a bit specific since the number of parameters that need to be estimated grows proportionally with the size of X. Thus, the asymptotic regime is only reached by taking the variance of the noise ε down to 0. In our example, the entries x_{ij} are obtained by averaging the scores from multiple panelists. Thus, we will naturally reach $\sigma^2 \to 0$ asymptotic as we have more and more judges. In the example of the previous Section 4.1.1, asymptotic is reached by having more replicates.

The classical PCA solution consists in solving the least squares

$$\operatorname{argmin}_{\mu} \left\{ \|X - \mu\|_2^2 : \operatorname{rank}(\mu) \leq S \right\}, \tag{4.3}$$

for a given S. The solution (Eckart and Young, 1936) is the truncated singular value decomposition (SVD) of the matrix $X = U\Lambda^{1/2}V'$ at the order S, namely

$$\hat{\mu}_{ij} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} \tag{4.4}$$

It also corresponds to the maximum likelihood solution of model (4.2). The so-called *regularization parameter* S plays a key role and a challenge is to select it from the data. As mentioned in the first part of the manuscript (Section 2.2), cross-validation approximation can be used (Josse and Husson, 2012).
Despite the wide use of estimator (4.4), the estimator is found to be noisy and its performance can be improved by regularization. Thus, other regularization techniques have recently been proposed for the estimation of μ to improve the recovery of the signal. In the next sections, I review three contributions on this topic. In addition, the approaches allow graphical representations which are as close as possible to the representations that would be obtained from the signal only.

4.2 Low-noise asymptotic

It is established, for instance in regression, that the maximum likelihood estimators are not necessarily the best for minimizing mean squared error (MSE). However, shrinkage estimators, although biased, have smaller variance which may reduce the MSE. In Verbanck *et al.* (2013), we followed this approach and proposed a regularized version of PCA (rPCA). As shown later, our approach essentially shrinks the first S singular values with a different amount of shrinkage for each singular value. First, we derived the shrinkage terms by minimizing the mean squared error (as reviewed in Section 4.2.1). Then, we showed that rPCA can also be obtained from a Bayesian treatment of the fixed-effects model (4.2) (Section 4.2.2).

4.2.1 MSE point of view

PCA provides an estimator $\hat{\mu}$ which is as close as possible to X in the least squares sense. However, assuming model (4.2), the objective is to get an estimator as close as possible to the unknown signal μ. To achieve such a goal, as in ridge regression, we look for a shrinkage version of the maximum likelihood estimator which is as close as possible to the true structure. More precisely, we look for shrinkage terms $\Phi = (\phi_s)_{s=1,\dots,\min(n-1,p)}$ that minimise:

$$\text{MSE} = \mathbb{E}\left(\sum_{i,j}\left(\sum_{s=1}^{\min(n-1,p)} \phi_s \hat{\mu}_{ij}^{(s)} - \mu_{ij}^{(s)}\right)^2\right)$$
$$\text{with } \hat{\mu}_{ij}^{(s)} = \sqrt{\lambda_s} u_{is} v_{js} \; ; \; \mu_{ij}^{(s)} = \sqrt{d_s} q_{is} r_{js}$$

First, according to equation (4.2), for all $s \geq S+1$, $\mu_{ij}^{(s)} = 0$. Therefore, the MSE is minimised for $\phi_{S+1} = ... = \phi_{\min(n-1,p)} = 0$ and can be written as:

$$\text{MSE} \;=\; \mathbb{E}\left(\sum_{i,j}\left(\sum_{s=1}^{S} \phi_s \hat{\mu}_{ij}^{(s)} - \mu_{ij}^{(s)}\right)^2\right)$$

Using the orthogonality constraints, for all $s \neq s'$, $\sum_i u_{is}u_{is'} = \sum_j v_{js}v_{js'} = 0$, the MSE can be simplified as follows (with $\mu_{ij} = \sum_s \mu_{ij}^s$):

$$\text{MSE} \;=\; \mathbb{E}\left(\sum_{i,j}\left(\sum_{s=1}^{S} \phi_s^2 \lambda_s u_{is}^2 v_{js}^2 \;-\; 2\mu_{ij}\sum_{s=1}^{S}\phi_s\sqrt{\lambda_s} u_{is} v_{js} + (\mu_{ij})^2\right)\right) \qquad (4.5)$$

Finally, equation (4.5) is differentiated with respect to ϕ_s to get:

$$\phi_s \quad = \quad \frac{\sum_{i,j} \mathbb{E}\left(\hat{\mu}_{ij}^{(s)}\right)\mu_{ij}}{\sum_{i,j} \mathbb{E}\left(\hat{\mu}_{ij}^{(s)2}\right)} = \frac{\sum_{i,j} \mathbb{E}\left(\hat{\mu}_{ij}^{(s)}\right)\mu_{ij}}{\sum_{i,j}\left(\mathbb{V}\left(\hat{\mu}_{ij}^{(s)}\right) + \left(\mathbb{E}\left(\hat{\mu}_{ij}^{(s)}\right)\right)^2\right)}$$

Then, to simplify this quantity, we use the results of Denis and Pazman (1999) and Denis and Gower (1996) who studied nonlinear regression models with constraints and focused on bilinear models such as model (4.1) of Section 4.1.1. Using the Jacobians and the Hessians of the response defined by Denis and Gower (1994) and recently rediscovered in Papadopoulo and Lourakis (2000), Denis and Gower (1996) derived the asymptotic bias of the response of model (4.1) and showed that the response estimator is approximately unbiased. Transposed to the PCA framework, it leads to conclude that the PCA estimator is asymptotically (with asymptotic defined as σ^2 tends to 0) unbiased $\mathbb{E}\left(\hat{\mu}_{ij}\right) = \mu_{ij}$ and for each dimension s, $\mathbb{E}\left(\hat{\mu}_{ij}^{(s)}\right) = \mu_{ij}^{(s)}$. In addition, the variance of $\hat{\mu}_{ij}$ can be approximated by the noise variance. Therefore, we estimate $\mathbb{V}\left(\hat{\mu}_{ij}^{(s)}\right)$ by the average variance, that is $\mathbb{V}\left(\hat{\mu}_{ij}^{(s)}\right) = \frac{1}{\min(n-1;p)}\sigma^2$. More details about this approach is also given in Section 5.1.1.

Consequently ϕ_s can be approximated by:

$$\phi_s \approx \frac{\sum_{i,j} \mu_{ij}^{(s)}\mu_{ij}}{\sum_{i,j}\left(\frac{1}{\min(n-1;p)}\sigma^2 + (\mu_{ij}^{(s)})^2\right)} = \frac{\sum_{i,j} \mu_{ij}^{(s)2}}{\sum_{i,j}\left(\frac{1}{\min(n-1;p)}\sigma^2 + (\mu_{ij}^{(s)})^2\right)}$$

since for all $s \neq s'$, the dimensions s and s' of μ are orthogonal. Based on equation (4.2), the quantity $\sum_{i,j}(\mu_{ij}^{(s)})^2$ is equal to d_s the variance of the s^{th} dimension of the signal. ϕ_s is then equal to:

$$\phi_s = \begin{cases} \dfrac{d_s}{\frac{np}{min\{p,n-1\}}\sigma^2 + d_s} \forall s = 1, ..., S \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{4.6}$$

The form of the shrinkage term is appealing since it corresponds to the ratio of the variance of the signal over the total variance (signal plus noise) for the s^{th} dimension. The shrinkage terms (4.6) depend on unknown quantities. We estimate them by plug-in. The total variance of the s^{th} dimension is estimated by the variance of X for the dimension s, *i.e.* by its associated eigenvalue λ_s. The signal variance of the s^{th} dimension is estimated by the estimated total variance of the s^{th} dimension minus an estimate of the noise variance of the s^{th} dimension. Consequently, ϕ_s is estimated by $\hat{\phi}_s = \frac{\lambda_s - \frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\lambda_s}$. Regularized PCA (rPCA) is thus defined by multiplying the maximum likelihood solution by the shrinkage terms which leads to:

$$\hat{\mu}_{ij}^{\text{rPCA}} \quad = \sum_{s=1}^{S}\left(\frac{\lambda_s - \frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\lambda_s}\right)\sqrt{\lambda_s}u_{is}v_{js} = \sum_{s=1}^{S}\left(\sqrt{\lambda_s} - \frac{\frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\sqrt{\lambda_s}}\right)u_{is}v_{js} \tag{4.7}$$

rPCA essentially shrinks the first S singular values. In rPCA, the s^{th} singular value is less shrunk than the $(s+1)^{th}$ one. This can be interpreted as granting a greater weight to the first dimensions. This behavior seems desirable since the first dimensions can be considered as more stable and trustworthy than the last ones. When $\hat{\sigma}^2$ is small, $\hat{\phi}_s$ is close to 1 and rPCA is close to standard

PCA. When $\hat{\sigma}^2$ is high, $\hat{\phi}_s$ is close to 0 and the values of $\hat{\mu}^{\text{rPCA}}$ are close to 0 which corresponds to the average of the variables (in the centered case). From a geometrical point of view, rPCA leads to bring the individuals closer to the centre of gravity. Note that rPCA requires knowledge of S and also to estimate the noise variance. In the R package denoiseR (Josse, Wager, and Sardy, 2015), by default, for this rPCA motivated by low noise asymptotics, we use generalized cross-validation for S and $\hat{\sigma}^2 = \frac{\parallel (X - \hat{\mu}^S \parallel^{\ 2}}{np - nS - pS + S^2}$.

4.2.2 Bayesian point of view

It is possible to define the method without any references to MSE, instead using Bayesian considerations. A parallel can be done with the equivalence between ridge regression and a Bayesian treatment of the regression model.

4.2.3 Probabilistic PCA model

A first Bayesian interpretation of the regularized PCA is given with probabilistic PCA (PPCA). The PPCA model is a particular case of a factor analysis model (Bartholomew, 1987) with an isotropic noise. The idea behind these models is to summarize the relationships between variables using a small number of latent variables. More precisely, denoting x_i a row of the matrix X, the PPCA model is written as follows:

$$x_i \quad = \quad B_{p \times S} z_i + \varepsilon_i \text{ with } z_i \sim \mathcal{N}(0, \mathbb{I}_S),\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_p)$$

with $B_{p \times S}$ being the matrix of unknown coefficients, z_i being the latent variables and $\mathbb{I}_S$ and $\mathbb{I}_p$ being the identity matrices of size S and p. This model induces a Gaussian distribution on the individuals (which are *i.i.d*) with a specific structure of variance-covariance:

$$x_i \sim \mathcal{N}(0, \Sigma) \text{ with } \Sigma = BB' + \sigma^2 \mathbb{I}_p$$

There is an explicit solution for the maximum likelihood estimators:

$$\hat{B} = V(\Lambda - \sigma^2 \mathbb{I}_S)^{\frac{1}{2}} R \tag{4.8}$$

with V and Λ defined as usual with the SVD of X, as the matrix of the first S left singular vectors of X and as the diagonal matrix of the eigenvalues, $R_{S \times S}$ a rotation matrix (usually equal to $\mathbb{I}_S$) and σ^2 estimated as the mean of the last eigenvalues.

In contrast to the fixed effect model (4.2), the PPCA model can be seen as a random effect model since the structure is random because of the Gaussian distribution on the latent variables. Consequently, this model seems more appropriate when PCA is performed on sample data such as survey data. In such studies, at first, it does not make sense to consider "estimates" of the "individual parameters" since no parameter is associated with the individuals, only random variables (z_i). However, estimators of the "individual parameters", known as BLUP estimators (Robinson, 1991), are usually calculated as the expectation of the latent variables given the observed variables $\mathbb{E}(z_i | x_i)$. The calculation results in:

$$\hat{Z} = X\hat{B}(\hat{B}'\hat{B} + \sigma^2 \mathbb{I}_S)^{-1} \tag{4.9}$$

Thus, using the ML estimator of B (equation 4.8) and equation (4.9), it is possible to build a fitted matrix which turns out to be the same one $\hat{\mu}^{\text{rPCA}}$ as defined in Section 4.2.1:

$$\hat{\mu}^{\text{PPCA}} = U(\Lambda - \sigma^2 \mathbb{I}_S)\Lambda^{-\frac{1}{2}} V'$$

Thus, we can consider the PPCA model (random-effect model) as the fixed effect model on which we assume a prior distribution on the left singular vectors, considered as the “individual parameters”. It is a way to define constraints on the individuals.

Empirical Bayesian analysis

Another Bayesian interpretation of the regularized PCA can be given considering directly an empirical Bayesian treatment of the fixed effects model with a prior distribution on each cell of the data matrix per dimension: $\mu_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2)$. The posterior distribution is obtained by combining the likelihood and the priors:

$$\begin{aligned}
&\text{Model: } x_{ij} = \sum_{s=1}^{S} \mu_{ij}^{(s)} + \varepsilon_{ij} \ , \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\
&\text{Prior: } \mu_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2) \\
&\text{Posterior: } \left(\mu_{ij}^{(s)} | x_{ij}^{(s)}\right) = \mathcal{N}(\Phi_s x_{ij}^{(s)}, \sigma^2 \Phi_s) \text{ with } \Phi_s = \frac{\tau_s^2}{\tau_s^2 + \sigma^2}
\end{aligned}$$

This posterior expectation depends on unknown quantities. We use an empirical Bayesian strategy to estimate them. More precisely, τ_s^2 is obtained as the maximum likelihood estimator using distribution of $\left(x_{ij}^{(s)}\right)_{i=1,\dots,n;j=1,\dots,p}$ which is $x_{ij}^{(s)} \sim \mathcal{N}(0, \tau_s^2 + \frac{1}{\min(n-1;p)}\sigma^2)$ using model (4.2). It leads to estimate ${\tau_s}^2$ as follows:

$$\hat{\tau_s}^2 \quad = \quad \left(\frac{1}{np}\lambda_s - \frac{1}{\min(n-1;p)}\hat{\sigma}^2\right)$$

Consequently the shrinkage term is estimated with $\hat{\Phi}_s = \frac{\left(\frac{1}{np}\lambda_s - \frac{1}{\min(n-1;p)}\hat{\sigma}^2\right)}{\frac{1}{np}\lambda_s} = \frac{\lambda_s - \frac{np}{\min(n-1;p)}\hat{\sigma}^2}{\lambda_s}$ and also corresponds to the regularization term (4.7) defined in Section 4.2.1.

4.3 Adaptive shrinkage of singular values

In Josse and Sardy (2015), we also suggested an estimator to estimate μ in (4.2). The rationale of our approach was to transpose the ideas of adaptive lasso in regression (Zou, 2006a) in the framework of low rank matrix estimation. Thus, we defined the adaptive trace norm estimator (ATN) which uses two regularization parameters (λ, γ) to threshold and shrink the singular values:

$$\psi(\lambda_s) = \sqrt{\lambda_s} \max\left(1 - \frac{\lambda^\gamma}{\sqrt{\lambda_s}^\gamma}, 0\right). \tag{4.10}$$

This estimator denoted $\hat{\mu}_{(\lambda,\gamma)} = \sum_{s=1}^{\min(n,p)} \psi(\lambda_s) u_{is} v_{js}$ is the closed form solution to

$$\text{argmin}_\mu \left\{ \|X - \mu\|_2^2 + \alpha \|\mu\|_{*,w} \right\},$$

where $\|\mu\|_{*,w} = \sum_{l=1}^{\min(n,p)} \omega_l \sqrt{\lambda_s}$ is a weighted nuclear norm with $\omega_l = 1/\sqrt{\lambda_s}^{\gamma-1}$ and $\alpha = \lambda^\gamma$. ATN parametrizes a rich family of estimators, that can more closely approach an ideal thresholding and shrinking function to recover well the structure of the underlying matrix μ. As for the regularized PCA (4.7) observe that, for $\gamma > 1$, the smallest singular values in (4.10) are more shrunk in comparison with the largest ones. However, ATN does not rely on asymptotic derivations and consequently may be appropriate when we are far from the asymptotic regime. If the regularization parameters are well estimated from the data, we expect an estimator with very good MSE properties.

4.3.1 Selecting the tuning parameters

The parameters (λ, γ) can be selected with cross-validation. However, to avoid such a computational intensive method, we suggested in Josse and Sardy (2015) three methods. The first method seeks good ℓ_2-risk. More precisely, extending the results of Candes, Sing-Long, and Trzasko (2013), we defined a Stein unbiased estimate of the risk (Stein, 1981) $\text{MSE}(\lambda, \gamma) = \mathbb{E}\|\mu - \hat{\mu}_{(\lambda,\gamma)}\|^2$ as follows:

$$\text{SURE}(\lambda, \gamma) = -np\sigma^2 + \sum_{l=1}^{\min(n,p)} \lambda_s \min\left(\frac{\lambda^{2\gamma}}{\lambda_s^{\gamma}}, 1\right) + 2\sigma^2 \text{div}(\hat{\mu}_{\lambda,\gamma}), \tag{4.11}$$

where the second term corresponds to the residuals sum of squares (RSS) whereas the last one is the divergence. The form of the divergence is not straightforward (see Candes *et al.* (2013)). The expectation of the divergence corresponds to the degrees of freedom. Thus, SURE has a very classical shape of RSS penalized by the complexity of the model. A selection rule for $\lambda \geq 0$ and $\gamma \geq 1$ finds the pair (λ, γ) that minimizes the bivariate function $\text{SURE}(\lambda, \gamma)$ in (4.11).

Note that SURE requires to know the noise variance σ^2. The second method named generalized SURE (GSURE) is inspired by generalized cross-validation (Craven and Wahba, 1979b):

$$\text{GSURE}(\lambda, \gamma) = \frac{\text{RSS}}{(1 - \text{div}(\hat{\mu}_{\tau,\gamma})/(np))^2}, \tag{4.12}$$

and does not require knowledge of σ^2. Using a first order Taylor expansion $1/(1-\epsilon)^2$ of (4.12), we get that $\text{GSURE} \approx \text{RSS}\,(1 + 2\ \text{div}(\hat{\mu}_{\gamma,\gamma})/(np))$; then considering the estimate of variance $\hat{\sigma}^2 = \text{RSS}/(np)$, one sees how GSURE approximates SURE (4.11).

Finally, the last method aims at good rank recovery. The parameter that determines the estimated rank is the threshold λ since any empirical singular value $\sqrt{\lambda_s} \leq \lambda$ is set to zero by (4.10). The suggested approach is based on the quantile universal threshold (QUT) of Giacobino, Sardy, Diaz Rodriguez, and Hengartner (2015). The rationale of QUT is to select the threshold λ^{QUT} at the bulk edge of what a threshold should be to reconstruct the correct model with high probability under the null hypothesis that $\mu = O$ (the $n \times p$ matrix with zeros for entries). For that specific value λ^{QUT}, $\text{SURE}(\lambda^{\text{QUT}}, \gamma)$ is minimized over γ. More precisely, 1000 data are generated under the null hypothesis of no signal, $\mu = 0$ and the $(1-\alpha)$-quantile of the distribution of the largest empirical singular value is used as a threshold. With α tending to zero with the sample size, null rank estimation is guaranteed with probability tending to one under the null hypothesis. The universal threshold for reduced rank mean matrix estimation can be written as follows:

$$\lambda_{\max(n,p)} = \sigma F_{\Lambda_1}^{-1}\left(1 - \frac{1}{\sqrt{\log(\max(n,p))}}\right), \tag{4.13}$$

where F_{Λ_1} is the cumulative distribution function of the largest singular value under Gaussian white noise with unit variance. Note that this approach again requires knowledge of σ.

4.4 A bootstrap approach

The previous estimators in Sections 4.3 and 4.2 started from the point of view of the singular values decomposition. In Josse and Wager (2014), we suggested an alternative approach for low rank matrix estimation based on a parametric bootstrap. Let's use the notations $X \sim \mathcal{L}(\mu)$ with μ the target rank-S matrix and $\mathcal{L}$ the distribution of the noise model, here the Gaussian noise model (4.2). We started by writting problem (4.3) of finding a rank S matrix the closest to X in the least-squares sense as in the neural network literature (Bourlard and Kamp, 1988, Baldi and Hornik, 1989):

$$\hat{\mu}_S = XB_S, \text{ where } B_S = \text{argmin}_B \left\{ \|X - XB\|_2^2 : \text{rank}\,(B) \leq S \right\}. \tag{4.14}$$

The matrix B, is called a linear *autoencoder* of X, since it encodes the features of X using a low-rank representation. This "regression" of X on X leads to the same solution as the classical truncated SVD. Indeed, $\hat{B}_S = P_V$, with $P_V = V(V'V)^{-1}V'$, the projection matrix onto the S first right singular vectors of X and thus $X\hat{B}_S = U\Lambda^{1/2}V'P_V = U_{n\times S}\Lambda^{1/2}_{S\times S}V'_{S\times p}$.
Now, in the context of the noise model $X \sim \mathcal{L}(\mu)$, the aim is not to compress X but instead to recover μ from X. From this perspective, we would much prefer to estimate μ using an oracle encoder matrix that formally provides the best linear approximation of μ given the noise model:

$$\hat{\mu}^{(S)^*} = XB^{(S)^*} \text{ where } B^{(S)*} = \text{argmin}_B \left\{ \mathbb{E}_{X\sim\mathcal{L}(\mu)} \left[\|\mu - XB\|_2^2 \right] : \text{rank}\,(B) \leq S \right\}. \tag{4.15}$$

Since the signal is unknown, we suggested using a parametric bootstrap approach by taking as a proxy for μ the data X (X is simply plugged in (4.15) instead of μ) which results in:

$$\widehat{B}^{(S)} \quad = \quad \text{argmin}_B \left\{ \mathbb{E}_{\tilde{X}\sim\mathcal{L}(X)} \left[\left\| X - \tilde{X}B \right\|_2^2 \right] : \text{rank}\,(B) \leq S \right\}. \tag{4.16}$$

At first glance, it may seem like a surprising idea, since we have a noisy version of the formulation (4.14):

$$\widehat{B}^{(S)} = \text{argmin}_B \left\{ \mathbb{E}_\varepsilon \left[\|X - (X + \varepsilon)B\|_2^2 \right] : \text{rank}\,(B) \leq S \right\}$$

However, it yields an objective which is the sum of a least squared data fitting term plus a quadratic regularizer that depends on the noise model:

$$\widehat{B}^{(S)} = \text{argmin}_B \left\{ \|X - XB\|_2^2 + \lambda \|B\|_2^2 \right\} \quad \text{with} \quad \lambda = n\sigma^2$$

and the solution can be written as a classical singular-value shrinkage estimator:

$$\hat{\mu}^{(S)}_\lambda = X\widehat{B}^{(S)} = \;=\; \sum_{s=1}^{S} u_{is} \frac{\sqrt{\lambda_s}}{1 + \lambda/\lambda_s} v_j s' \quad \text{with} \quad \lambda = n\sigma^2. \tag{4.17}$$

or in a classical ridge way with S equals to a diagonal matrix with elements $n\sigma^2$:

$$\hat{\mu}^{(S)}_\lambda = X(X^\top X + S)^{-1}X'X. \tag{4.18}$$

Thhe equivalence between feature noising and regularization can be obtained by a simple bias-variance decomposition:

$$\begin{aligned}
\mathbb{E}_\varepsilon \left[\|X - (X + \varepsilon)\, B\|_2^2 \right] &= \|X - XB\|_2^2 + \mathbb{E}_\varepsilon \left[\|\varepsilon B\|_2^2 \right] \\
&= \|X - XB\|_2^2 + \sum_{i,j,k} B_{ij}^2 \,\text{Var}\,[\varepsilon_{jk}] \\
&= \|X - XB\|_2^2 + n\sigma^2 \|B\|_2^2
\end{aligned}$$

and the solution is easily derived by replacing X with its SVD.
Note that this duality between regularization and feature noising schemes is known in regression. As shown by Bishop (1995), linear regression with features perturbed with Gaussian noise, i.e.,

$$\hat{\beta} = \text{argmin}_\beta \left\{ \mathbb{E}_{\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)} \left[\|Y - (X + \varepsilon)\,\beta\|_2^2 \right] \right\},$$

is equivalent to ridge regularization with Lagrange parameter $\lambda = n\sigma^2$:

$$\hat{\beta}_\lambda^{(R)} = \text{argmin}_\beta \left\{ \|Y - X\beta\| + \lambda \|\beta\|_2^2 \right\}.$$

Thus, overfitting is controlled by artificially corrupting the training data. The methodology we developped in Josse and Wager (2014) is based on the same rationale but for matrix. We called the estimator (4.17), stable autoencoder (SA), since it is stable around round perturbations to the data. Note that the singular-value shrinkage form (4.17) is reassuring since estimators applying non linear transformations of the singular values have good MSE properties (Gavish and Donoho, 2014a).
Estimator (4.17) requires both knowledge of the rank S and of the noise variance σ^2. Consequently, as a second step, we suggested to remove the rank constraint and to solve for a fixed point of the proposed denoising scheme. At a high level, the goal is to find a solution to

$$\hat{\mu}^{\text{iter}} = X\widehat{B}, \text{ where } \widehat{B} = \text{argmin}_B \left\{ \mathbb{E}_{\widetilde{X} \sim \mathcal{L}(\hat{\mu}^{\text{iter}})} \left[\left\| \hat{\mu}^{\text{iter}} - \widetilde{X}B \right\|_2^2 \right] \right\}$$

by iteratively updating $\widehat{B}$ and $\hat{\mu}$ (Note that $\hat{\mu}^{\text{iter}}$ is used as a plug-in for μ in (4.15)):

1. $\hat{\mu} = X\hat{B}$
2. $\hat{B} = (\hat{\mu}'\hat{\mu} + S)^{-1}\hat{\mu}'\hat{\mu}$ ((4.20) is solved but with X replaced by the low-rank estimate $\hat{\mu} = X\widehat{B}$ obtained from the optimal B)

This estimator is called iterated stable autoencoder (ISA). In Josse and Wager (2014), we showed that ISA converges to a low rank solution. In the isotropic case with $S = n\sigma^2\mathbb{I}$, ISA converges to

$$\hat{\mu}^{\text{iter}} = \sum_{l=1}^{\min\{n,p\}} u_l\, \psi(\lambda_l)\, v_l', \text{ where } \psi(\lambda) = \begin{cases} \frac{1}{2}\left(\lambda + \sqrt{\lambda^2 - 4n\sigma^2}\right) & \text{for } \lambda^2 \geq 4n\sigma^2, \\ 0 & \text{else.} \end{cases} \tag{4.19}$$

In what appears to be a remarkable coincidence, for the square case $n = p$, the shrinkage rule (4.19) corresponds exactly to the Marchenko–Pastur optimal shrinkage rule of Gavish and Donoho (2014a) under operator–norm loss; i.e., that it minimizes the limit of the operator norm of the matrix $(\hat{\mu} - \mu)$ in an asymptotic regime of n and p grow while the rank stay fixed. At the very least, this connection is reassuring as it suggests that the iterative scheme may yield statistically reasonable estimates $\hat{\mu}^{\text{iter}}$ for other noise models too. It remains to be seen whether this connection reflects a deeper theoretical phenomenon. Note also, the induced shrinkage function of SA (4.17) resembles a first-order approximation to the one proposed by in Section 4.2 (4.7). Finally, although ISA is able to automatically estimate a rank it requires knowledge of σ^2. One last remark that can be made is the that both SA and ISA may certainly be interpreted as an empirical Bayes estimators. However, this also deserves more research but should enhace their comprehension [3]).

[3] I also suspect some relations with automatic relevance determination ARD to explain why the rank is automatically learned or with some non-parametric Bayesian interpretation, but for now my attempts are not successful!

4.5 Results and illustration on the Genotype-Environment data

4.5.1 Graphical impacts

The rationale behind regularized versions of PCA can be illustrated on graphical representations using rPCA of Section 4.2.1. Since rPCA ultimately modifies the singular values, it affects both the representation of the individuals and of the variables. Indeed, in our practice (Husson, Le, and Pagès, 2010) [4], we represent the rows coordinates (scores) by $U\Lambda^{\frac{1}{2}}$ and the variable coordinates by $V\Lambda^{\frac{1}{2}}$. Therefore, the global shape of the rows cloud represents the variance. Similarly, in the variable representation, the cosine of the angle between two variables can be interpreted as the covariance. Other choices are available (Greenacre, 2009). We focus here on the individuals representation. Data are generated according to model (4.2) with an underlying signal $\mu_{4\times10}$ in two dimensions. Then, 300 matrices are generated: $X^{sim} = \mu_{4\times10} + \varepsilon^{sim}$ with $sim = 1, ..., 300$. On each data matrix, PCA and rPCA are performed. Results are represented Figure 4.4. Representing several sets of coordinates from different PCAs can suffer from translation, reflection, dilatation or rotation ambiguities. Thus, all configurations are superimposed using Procrustes rotations (Gower and Dijksterhuis, 2004a) by taking as the reference the true individuals configuration from μ. Compared to PCA, rPCA provides a more biased representation because the coordinates of the average points (square) are systematically closer to the center than the coordinates of the true points (large dots). This is expected because the regularisation term shrinks the individual coordinates towards the origin. In addition, as it is clear for the red individual, the representation is less variable. Figure 4.4 thus gives a rough idea of the bias-variance trade-off. Note that even the PCA representation is biased, but this is also expected since $\mathbb{E}(\hat{\mu}) = \mu$ only asymptotically as detailed in section 4.2.1.

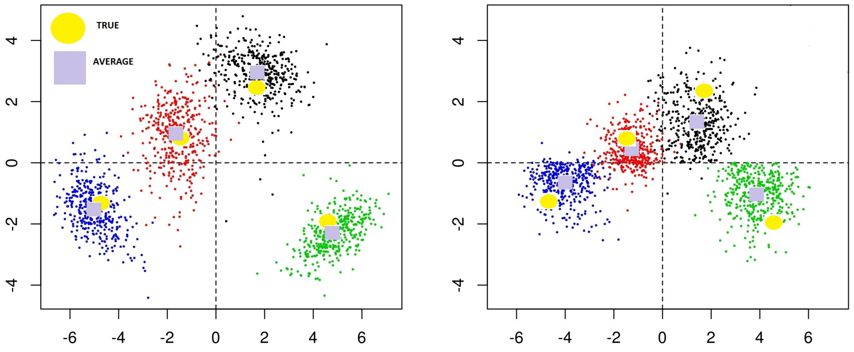


Figure 4.4: Configurations of the PCA (left) and the rPCA (right) of each $X^{sim} = \mu + \varepsilon^{sim}$, with $sim = 1, ..., 300$ represented with small dots.

4.5.2 Simulations

We reproduce here the simulations of Candès *et al.* (2013). Data matrices of size 200 × 500 are generated according to the Gaussian noise model (4.2) with four signal-to-noise ratios SNR$\in \{0.5, 1, 2, 4\}$ calculated as $1/(\sigma\sqrt{np})$, and two values for the underlying rank $S \in \{10, 100\}$. For each combination of SNR and S, the simulation is repeated 50 times and median performance is reported in Table 4.5.2.

[4] We even speak about the French school of data analysis and the French coordinates! Some discussion about this current of research in "Jan de Leeuw and the French School of data analysis" Husson *et al.* (2016)

The three propositions, namely the low-noise (LN) estimator of Section 4.2, the adaptive estimator ATN of Section 4.3 with its parameters selected by SURE (4.11) and the bootstrap estimators stable autoencoder (SA) and the iterated stable autoencoder (ISA) of Section 4.4, are compared to the following estimators:

- Truncated SVD with fixed rank S (TSVD-S). This is the classical approach described in Section 4.1.2.
- Asymptotically optimal singular–value shrinkage (ASYMP) in the Marchenko-Pastur asymptotic regime (both the number of rows ($n = n_p$) and columns (p) tend to infinity while the rank of the matrix stays fixed) given the Frobenius norm loss (Shabalin and Nobel, 2013, Gavish and Donoho, 2014b), with shrinkage function with $n_p/p \to \beta$, $0 < \beta \leq 1$

  $$\begin{cases} \psi(\lambda_s) = \frac{1}{\sqrt{\lambda_s}}\sqrt{(\lambda_s - (\beta-1)n\sigma^2)^2 - 4\beta n\sigma^4} \cdot 1\left(s \geq (1+\sqrt{\beta})n\sigma^2\right), \\ 0 \qquad \text{else.} \end{cases}$$

  If the noise variance is unknown, the authors suggested

  $$\hat{\sigma} = \frac{\lambda_{med}}{\sqrt{n\mu_\beta}},$$

  where λ_{med} is the median of the singular values of X and μ_β is the median of the Marcenko-Pastur distribution.
- Singular value soft thresholding (SVST) (Cai, Candès, and Shen, 2010): $\sqrt{\lambda_s}\max\left(1 - \frac{\lambda}{\sqrt{\lambda_s}}, 0\right)$ where the singular values are soft-thresholded by λ selected by minimizing a Stein unbiased risk estimate (SURE), as suggested by Candès *et al.* (2013)

All the estimators are defined assuming the variance of the noise σ^2 known. In addition, TSVD-S, SA, and LN require the rank S as a tuning parameter. In this case, S is set to the true rank of the underlying signal.
Table 4.5.2 makes clear that the proposed methods have very different strengths and weaknesses. TSVD-S that applies a hard thresholding rule to the singular values provides accurate MSE when the SNR is high but breaks down in low SNR settings. Conversely, the SVST behaves well in low SNR settings, but struggles in other regimes. This is not surprising, as the method over-estimates the rank of μ (given by the number of singular values greater than λ). This behavior is reminiscent of what happens in lasso regression (Tibshirani, 1996a) when too many variables are selected (Zou, 2006b, Zhang and Huang, 2008). Meanwhile, the estimators with non–linear singular–value shrinkage functions, namely SA, ISA, ASYMP, and LN are more flexible and perform well expect in the very difficult scenario where the signal is overwhelmed by the noise. The estimators ASYMPT and LN provide good recovery in their asymptotic regimes. ISA estimates the rank accurately except when the signal is nearly indistinguishable from the noise (SNR=0.5 and S=100). Adaptive estimator is very flexible and is the estimator which performs the best. In **??** we showed that it is also the case when using GSURE. In addition, we underlined that ATN with universal threshold is not the best in terms of MSE but has good rank recovery property which was expected. We could remark that estimating S instead of using its oracle value modify the results. For instance for the case with underlying $S = 100$ and SNR $= 0.5$, cross-validation suggests $S = 0$ which gives MSE errors of 1. Of course, the estimated values of σ and S in practice have an impact on the results. Based on numerous experiments, I would recommend to denoise a low rank matrix, the use of ATN (with GSURE) or ISA (with the noise variance estimated...). Indeed, even if ISA does not provide the best MSE in the Gaussian setting, it has the advantage that it could be used for other noises, as we will see in Section 4.6.

k	SNR	Bootstrap		TSVD	Adaptive (λ, γ)	Asympt	SVST (λ)	LN
		SA	ISA					
		(σ, S)	(σ)	(S)	(σ) - SURE	(σ)	(σ) - SURE	(S)
MSE								
10	4	**0.004**	**0.004**	**0.004**	**0.004**	**0.004**	0.008	**0.004**
100	4	**0.037**	**0.036**	**0.037**	**0.037**	**0.037**	0.045	**0.037**
10	2	**0.017**	**0.017**	**0.017**	**0.017**	**0.017**	0.033	**0.017**
100	2	**0.142**	**0.143**	0.152	**0.142**	0.146	0.156	**0.141**
10	1	**0.067**	**0.067**	0.072	**0.067**	**0.067**	0.116	**0.067**
100	1	0.511	0.775	0.733	**0.448**	0.600	**0.448**	0.491
10	0.5	0.277	**0.251**	0.321	**0.253**	**0.250**	0.353	0.257
100	0.5	1.600	1.000	3.164	**0.852**	0.961	**0.852**	1.474
Rank								
10	4		**10**		11	**10**	65	
100	4		**100**		103	**100**	193	
10	2		**10**		11	**10**	63	
100	2		**100**		114	**100**	181	
10	1		**10**		11	**10**	59	
100	1		29.6		154	**64**	154	
10	0.5		**10**		15	**10**	51	
100	0.5		0		87	15	86	

Table 4.1: Top: MSE of the low rank matrix estimators. Bottom: estimated rank.

4.5.3 Genotype-environment data

Let's see how the classical analysis of Section 4.1.1 is modified. We suggested new point estimates for the PCA parameters (regularized versions) which leads to new graphical representations.

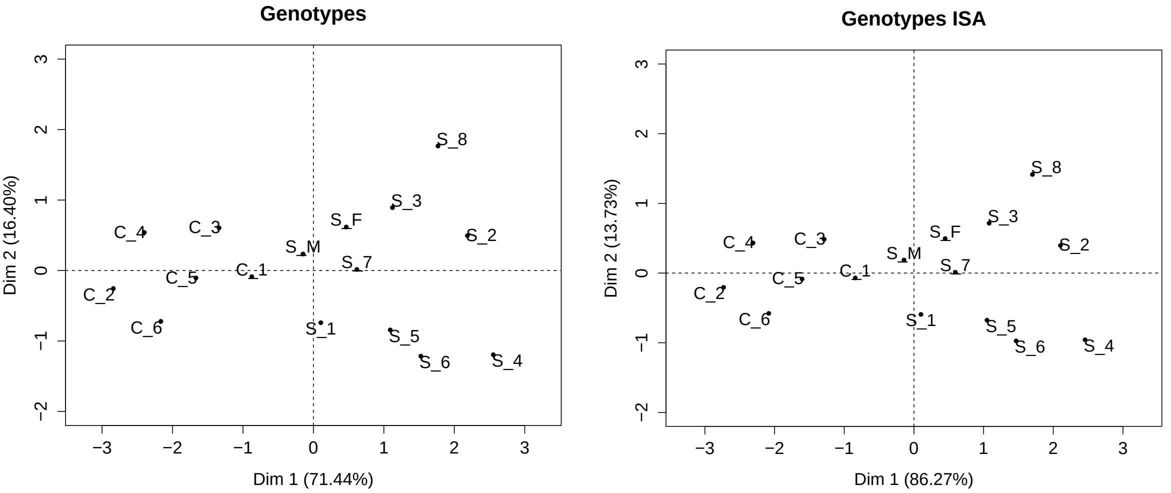


Figure 4.5: PCA (left) and regularized PCA with stable autoencoder (right) on the GE data

Figure 4.5 presents the results obtained after PCA and the stable autoencoder (SA) (4.17) of Section 4.4. The cloud associated with PCA has a higher variability than the cloud associated with SA which is tightened around the origin. The effect of regularisation is stronger on the second axis than on the first one, which is expected because of the regularisation term. If we believe in model (4.2), the representation obtained by SA is more in agreement with the true configuration. Often, the impact of regularisation on the graphical representations is not obvious, but the effect of regularisation is crucial when looking at the clustering results as illustrated in Figures 4.6. Indeed, since SA modifies the distances between genotypes, the SA clustering will differ from the PCA one and results may conduct to a different interpretation whether using one option or the other. The truth is unknown here, but one can expect a better clustering with data denoised by SA since theory and simulations point to better estimation of the underlying signal.

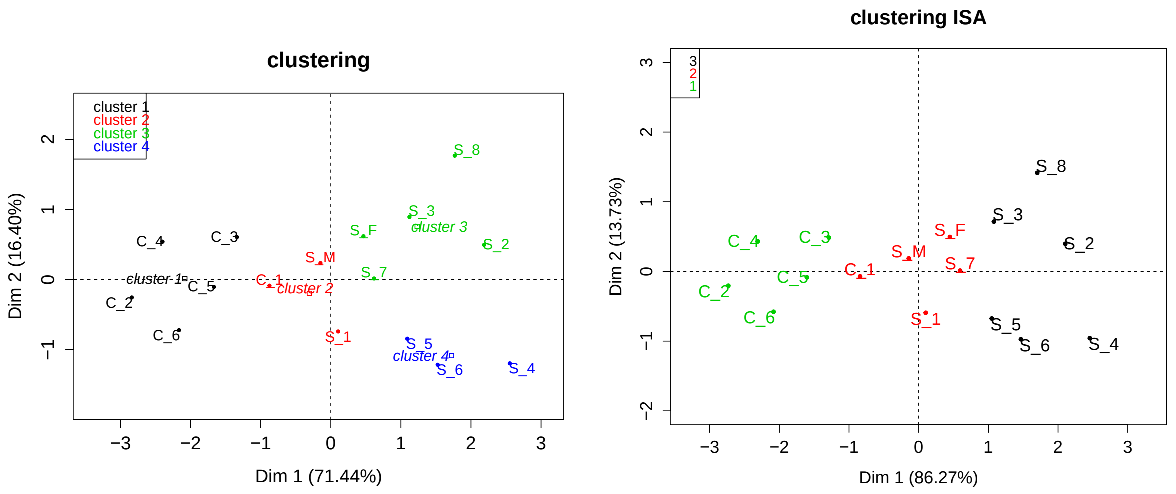


Figure 4.6: Clustering after PCA (left) and regularized PCA (right)

4.6 Outside the Gaussian case: regularized correspondence analysis

When X contains count data, a natural noise model can be $X_{ij} \sim \text{Poisson}(\mu_{ij})$ with the expectation of low rank S. In Josse and Wager (2014), we suggested estimating the signal with the bootstrap method detailled in Section 4.4 using $\widetilde{X}_{ij} \sim \text{Poisson}\left(X_{ij}\right)$. A similar sampling scheme is known to work well for regularizing glm regression with count features, and has desirable theoretical properties (Wager, Fithian, Wang, and Liang, 2014). More precisely, we showed that even in the non-isotropic noise model, low-rank stable autoencoders can still be efficiently solved. Indeed, for a generic noise model $\mathcal{L}(\cdot)$, the matrix $\widehat{B}^{(S)}$ from (4.20),

$$\widehat{B}^{(S)} \quad = \quad \text{argmin}_B \left\{ \mathbb{E}_{\tilde{X} \sim \mathcal{L}(X)} \left[\left\| X - \tilde{X} B \right\|_2^2 \right] : \text{rank}\left(B\right) \leq S \right\} \tag{4.20}$$

can be obtained as follows:

$$\widehat{B}^{(S)} = \text{argmin}_B \left\{ \| X - XB \|_2^2 + \left\| S^{\frac{1}{2}} B \right\|_2^2 : \text{rank}\left(B\right) \leq S \right\}, \tag{4.21}$$

where S is a $p \times p$ diagonal matrix with

$$S_{jj} = \sum_{i=1}^{n} \text{Var}_{\widetilde{X} \sim \mathcal{L}(X)} \left[\widetilde{X}_{ij} \right].$$

From a computational point of view, we can write the solution $\widehat{B}_S$ of (4.21) as

$$\widehat{B}^{(S)} = \text{argmin}_B \left\{ \text{tr} \left(\left(B - \widehat{B} \right)' \left(X'X + S \right) \left(B - \widehat{B} \right) \right) : \text{rank}\left(B\right) \leq k \right\}, \text{ where} \tag{4.22}$$

$$\widehat{B} = \left(X'X + S \right)^{-1} X'X \tag{4.23}$$

is the solution of (4.21) without the rank constraint.
The optimization problem in (4.22) can be easily solved by taking the top S terms from the eigenvalue decomposition of $\widehat{B}'\left(X'X + S\right)\widehat{B}$; the matrix $\widehat{B}_S$ can then be recovered by solving a linear system (e.g., Takane, 2013).
Note that in (4.23), the matrix S is not equal to a constant times the identity matrix due to the non-isotropic noise; that's why the resulting singular vectors of $\hat{\mu}_S^{SA} = X\widehat{B}^{(S)}$ are not the ones of X. This characteristic is unique and implies that the estimator is better than all its competitors to recover the signal (Josse and Wager, 2014).
Note that with count data, X is often analyzed by correspondence analysis (CA) (Greenacre, 1984b, 2007) rather than using a direct singular–value decomposition. CA is a powerful method to visualize contingency tables and consists in applying an SVD on a transformation of the data, denoted M:

$$M = R^{-\frac{1}{2}}\left(X - \frac{1}{N}rc'\right)C^{-\frac{1}{2}}, \tag{4.24}$$

where $R = \operatorname{diag}(r)$, $C = \operatorname{diag}(c)$, N is the the total number of counts, and r and c are vectors containing the row and column sums of X.
In Josse and Wager (2014), we used ISA to regularize correspondence analysis. We showed that if we had chosen to sample $\widetilde{X}$ from an independent contingency table with

$$\mathbb{E}\left[\widetilde{X}\right] = \frac{1}{N}rc^{\top},\ \operatorname{Var}\left[\widetilde{X}_{ij}\right] = \frac{r_i c_j}{N}, \tag{4.25}$$

we would have obtained a regularization matrix $S_M = n\delta/(N(1-\delta))\mathbb{I}_{p\times p}$. Because S_M is diagonal, the resulting estimator $\widehat{M_\lambda}$ could then be obtained from M by singular value shrinkage. Thus, if we want to regularize correspondence analysis applied to a nearly independent table, singular value shrinkage based methods can achieve good performance; however, if the table has strong dependence, our framework provides a more principled way of being robust to sampling noise.

I illustrate the approach with the R package denoiseR (Josse *et al.*, 2016) and FactoMineR (Lê, Josse, and Husson, 2008) on a sensory analysis of perfumes. The data for the analysis were collected by asking consumers to describe 12 luxury perfumes such as *Chanel Number 5* and *J'adore* with words. The answers were then organized in a 12×39 (39 words unique were used) data matrix where each cell represents the number of times a word is associated to a perfume; a total of $N = 1075$ were used overall. The dataset is available at `http://factominer.free.fr/docs/perfume.txt`.
We use correspondence analysis (CA) to visualize the associations between words and perfumes. Here, the technique allows to highlight perfumes that are described using a similar profile of words, and to find words that describe the differences between groups of perfumes.
The results in Figure 4.7 emphasize that, although correspondence analysis is often used as a visualization technique, appropriate regularization is still important, as regularization may substantially affect the graphical output. We know from simulations (Josse and Wager, 2014) that the regularized CA plots are better aligned with the population ones than the unregularized ones are; thus, we may be more inclined to trust insights from the regularized analysis.

4.7 Missing values

Finally, I conlude this chapter with an extension to the missing values setting. In Josse *et al.* (2016), we extended the previous low rank estimators of Sections 4.3 and 4.4 (ISA and ATN) to the missing values case/matrix completion framework. Inspired by *iterative PCA* of Section 2.2,

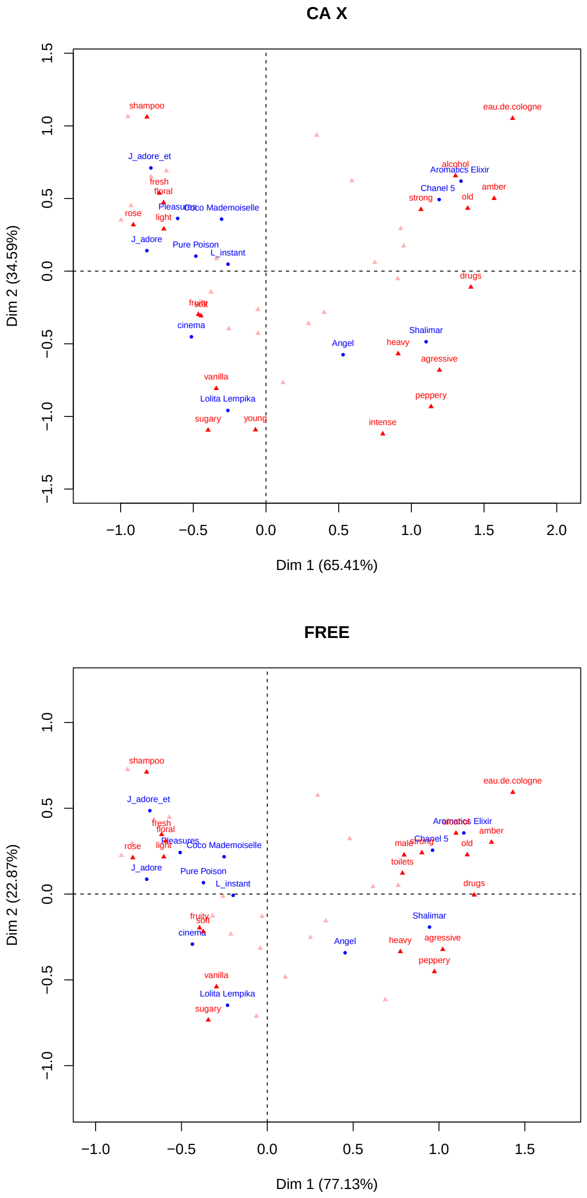


Figure 4.7: Results for CA (top) and Regularized CA (bottom) using ISA. Only, the 20 words that contribute the most to the first two dimensions are represented.

we defined *iterative ATN (ISA) algorithms* where we alternate imputation of the missing values and estimation with ATN (or ISA). Although, theoretical properties of these algorithms have not yet been investigated, we may expect a good recovery of the signal and of the missing values due to their ability to accurately estimate the signal in the complete case. We checked these claims using a small simulation study. To carry-out *iterative ATN* algorithm, we need a method to select both regularized parameters (λ, γ) despite missing values. In Josse *et al.* (2016), we suggested extending the SURE (4.11) criterion for missing values as follows: the first term of the sum is replaced by the number of observed values $(np - |NA|)\sigma^2$ with $|NA|$ the number of missing cells. Then, the second term is replaced by the RSS on the observed values $\sum_{ij \in obs}(X_{ij} - \hat{\mu}^{\text{miss}}_{(\lambda,\gamma)_{ij}})^2$, with $\hat{\mu}^{\text{miss}}_{(\lambda,\gamma)}$, the estimator obtained from the incomplete data with the *iterative ATN algorithm.* Finally, the divergence of the estimator need to be defined from an incomplete data which is a difficult task. Since no explicit form is yet available, we suggested using finite differences (FD) and defined SURE with missing values as:

$$\text{SURE}^{\text{miss}} = -(np - |NA|)\sigma^2 + \sum_{ij \in obs} (X_{ij} - \hat{\mu}^{\text{miss}}_{(\lambda,\gamma)_{ij}})^2 + 2\sigma^2 \text{div}^{\text{miss}}(\hat{\mu}^{\text{miss}}_{(\lambda,\gamma)}) \tag{4.26}$$

with

$$\text{div}^{\text{miss}}(\hat{\mu}^{\text{miss}}_{(\lambda,\gamma)}) = \sum_{ij \in obs} \frac{\hat{\mu}^{\text{miss}}_{(\lambda,\gamma)}(X_{ij} + \delta) - \hat{\mu}^{\text{miss}}_{(\lambda,\gamma)}(X_{ij})}{\delta},$$

where δ is a small variation near machine precision. In the same way, we also suggested a GSURE criterion (4.12) with missing values to deal with unknown noise variance.

We showed with simulation that (4.26) is indeed an unbiased estimate of the risk of the estimator. Although this method provides a neat way to perform ATN with missing values while estimating its regularized parameters by minimizing the risk, it is extremely costly. Indeed, an iterative algorithm is performed for each cell of the matrices.

All the methods discussed in this chapter are implemented in the R package denoiseR (Josse *et al.*, 2015) where I have tried to pay a lot of attention in providing sensible defaults values for all the tuning parameters (S, σ^2, etc.).

Chapter 5

Variability of the principal components parameters

5.1 Confidence areas for fixed-effects PCA

In Chapter 4, I discussed about point estimates of the PCA parameters. In this Section, I summarize the contribution Josse *et al.* (2014a) where we focused on performing statistical inference of the PCA output to enhance the interpretation of the classical graphical outputs. We considered the framework of PCA applied on "population data sets," where both the rows and the columns of the data matrix are fixed, which is in agreement with the fixed-effects model:

$$\begin{aligned} X_{n\times p} &= \mu_{n\times p} + \varepsilon_{n\times p} \\ x_{ij} &= \sum_{s=1}^{S} \sqrt{d_s} q_{is} r_{js} + \varepsilon_{ij}, \;\; \varepsilon_{ij} \sim \mathcal{N}(0,\sigma^2) \end{aligned} \tag{5.1}$$

We developed methods that let us understand the impact of noise on PCA, and to visualize this variability on PCA graphical outputs using confidence ellipses, emphasizing that confidence areas ensure relevant interpretation. We studied the asymptotic variance of the fixed-effects model and proposed several approaches to assess the variability of PCA estimates: a method based on a parametric bootstrap, a new cell-wise jackknife, as well as a computationally cheaper approximation to the jackknife. We suggested visualizing the confidence regions by Procrustes rotation. Then, we compared coverage properties of confidence areas based on the different methods and founded that the resulting variance estimates varied widely. The asymptotic method and the bootstrap do well in low-noise setting, but can fail when the noise level gets high or when the number of variables is much greater than the number of rows. On the other hand, the jackknife has good coverage properties for large noisy examples but requires a minimum number of variables to be stable enough.

5.1.1 Assessing the variance of the PCA parameters

The asymptotic regions

To define asymptotic regions, we rely on the inferential framework associated with model (4.2), *i.e.* we assume that the noise variance tends to zero. Then, we use all the results of Denis and Pazman (1999), Denis and Gower (1994) and Denis and Gower (1996) to find that the PCA estimator $\hat{\mu}_{ij}^{S} = \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js}$ is asymptotically unbiased $\mathbb{E}(\hat{\mu}_{ij}^{(S)}) = \mu_{ij}$ and that the variance of the first order asymptotic approximation is

$$\mathbb{V}\left(\hat{\mu}_{ij}^{(S)}\right) = \sigma^2 P_{ij,ij}, \text{ where} \tag{5.2}$$
$$P = \left(I_p \otimes \frac{1}{n}11'\right) + \left(P_V' \otimes \left(\mathbb{I}_n - \frac{1}{n}11'\right)\right) + (I_p \otimes P_U) - \left(P_V' \otimes P_U\right).$$

Here, 1 is the n-vector of 1's, $\otimes$ is the Kronecker product, and

$$P_U = U(U'U)^{-1}U' \text{ and } P_V = V(V'V)^{-1}V'.$$

This formulation makes sense when looking at the PCA formulation as a smoothing operator (Candès and Tao, 2009, Josse and Husson, 2011b):

$$\text{vec}(\hat{\mu}^{(S)}) \;=\; P\text{vec}(X), \tag{5.3}$$

where vec is the vectorization operator, i.e., $\text{vec}(\hat{\mu}^{(S)})$ is a vector of size np with the columns of $\hat{\mu}^{(S)}$ stacked below each other. Here, the matrix P from (5.2) acts as an orthogonal projection matrix whose elements depend on the data X. Consequently, PCA can be seen as a “non-linear" model [1] where P represents the projection onto the tangent space of the expectation surface. Thus, as in classical non-linear regression models, we can obtain asymptotic forms by studying linear approximation. Remark also, that as in classical linear regression, the estimated number of independent parameters corresponds trace of the projection matrix (here $p+(n-1)\times S+p\times S-S^2$, for the loadings or the p coordinates of variables in S dimensions, the scores or the n coordinates of individuals in S dimensions minus S^2 for the orthonormality constraints. The other terms correspond to the centering of the data). In fact, this estimated number of parameters is the one used to define the estimator of the noise variance and the GCV criterion defined in Section 2.2. These results directly lead to Gaussian asymptotic confidence regions. To compute these regions in practice, we draw matrices

$$\left(\hat{\mu}^{(S)1}, ..., \hat{\mu}^{(S)\star}\right)$$

from a Gaussian distribution with expectation $\hat{\mu}$ and variance given by equation (5.2). We will discuss in Section 5.1.2 how to use these matrices to draw confidence areas around the row and column points in PCA graphical outputs. Confidence areas based on linear approximations are expected to be valid as long as the non-linearity is small. Extending prior work by Bates and Watt (1980) and Pazman (2002), Pazman and Denis (2002) who defined non-linearity measures for biadditive models (4.1), we should expect that the asymptotic confidence areas for PCA perform best when the signal-to-noise ratio (SNR) is high. In addition, the validity of the asymptotics is all the more reliable when n and p are large.

The parametric bootstrap

The classical non-parametric bootstrap where rows are resample is not appropriate for the fixed-effects model (5.1) since the structure of X is non-random. It would have been adapted for survey data for instance. But we can still define a parametric bootstrap (Efron and Tibshirani, 1994) by regenerating residuals. It allows us to study the variability of the parameters (both the scores and the loadings and not only the loadings as with a classical non-parametric bootstrap) due to the noise ε in (4.2). We use the following algorithm:

1. Perform the PCA on X to estimate the parameters $U_{n\times S}$, $\Lambda^{\frac{1}{2}}_{S\times S}$, $V_{p\times S}$.

[1] PCA is more often presented as a bilinear model, due to its two sets of parameters, U and V

2. Calculate the matrix of residuals $\hat{\varepsilon}_{n\times p} = X - \hat{\mu}^{(S)}$, and estimate σ^2.

3. For $b = 1, .., B$:

   (a) Draw ε_{ij}^b from $\mathcal{N}(0, \hat{\sigma}^2)$ to obtain a new matrix ε^b,

   (b) Generate a new data table: $X^b = \hat{\mu}^{(S)} + \varepsilon^b$

   (c) Perform PCA on X^b to obtain new estimates for the parameters $(U^b, (\Lambda^b)^{1/2}, V^b)$ and a new estimator $\hat{\mu}^{(S)^b} = U^b(\Lambda^b)^{\frac{1}{2}}V^{b\prime}$.

At the end, we have B fitted matrices $(\hat{\mu}^{(S)^1} = U^1(\Lambda^1)^{1/2}V^{1\prime}, ..., \hat{\mu}^{(S)^B} = U^B(\Lambda^B)^{1/2}V^{B\prime})$. Note that as in regression, we need an unbiased estimate for σ^2 or if we would like to directly bootstrap the residuals, we should use some kind of studentized residuals. In practice, we use for a centered matrix X :$\hat{\sigma}^2 = \frac{\parallel (X - \hat{\mu}^(S)\parallel \,^2}{np-nS-pS+S^2}$.
Since as shown in Section 5.1.1, the PCA estimator is only asymptotically unbiased, we expect the coverage of the parametric bootstrap confidence areas to reach their nominal level in an asymptotic framework. Following results of Huet, Denis, and Adamczyk (1999) for nonlinear models such as biadditive models (4.1) in an analysis of variance framework, we also expect the parametric bootstrap procedure to give similar results to the asymptotic method.

A cell-wise jackknife

The jackknife is another non-parametric way to study the variability of parameters. In the context of PCA, the "classical jackknife" procedure involves deleting each row of the data matrix X one at a time. It was studied in Daudin, Duby, and Trecourt (1989), Besse and de Falguerolles (1993). Removing one row at a time is analogous to the non-parametric bootstrap where the rows are considered as a random sample. Here, in the setup where the rows are not *i.i.d*, we propose a new form of jackknife for PCA, which consists in removing one cell x_{ij} of the data matrix at a time and estimating the PCA parameters from the incomplete data set. Writing $\hat{\mu}^{(S)^{(-ij)}}$ for the PCA estimator obtained from the matrix without the cell (ij), we get the pseudo-values

$$\hat{\mu}_{jackk}^{(S)^{(ij)}} = \hat{\mu} + \sqrt{np}\left(\hat{\mu}^{(S)^{(-ij)}} - \hat{\mu}\right). \tag{5.4}$$

This procedure is repeated for each cell of the data matrix. These pseudo-values can then be transformed into confidence ellipsoids as shown in Section 5.1.2. Our jackknife procedure requires a method for performing PCA with missing values and thus the *iterative PCA algorithm* of Section 2.2 can be used.
Formally, the jackknife tells us about the variability of the mean of the leave-one-out estimates. In most applications, the mean of the leave-one-out estimates is very close to the actual estimate (in the case of the sample mean, they are the same). Consequently, this specific jackknife is a competitor to both the bootstrap and the asymptotic methods in the sense that they estimate the same variability. Efron and Stein (1981, Theorem 2) showed that the jackknife estimate of variance is biased upwards in expectation, suggesting that the jackknife should lead to conservative confidence areas.

Approximating the jackknife

The jackknife procedure described above is computationally demanding, as it effectively requires us to run the *iterative PCA algorithm* $n \times p$ times. To reduce the computational effort required, we consider an approximation to the jackknife based on the "leave-out-one" lemma of Craven and

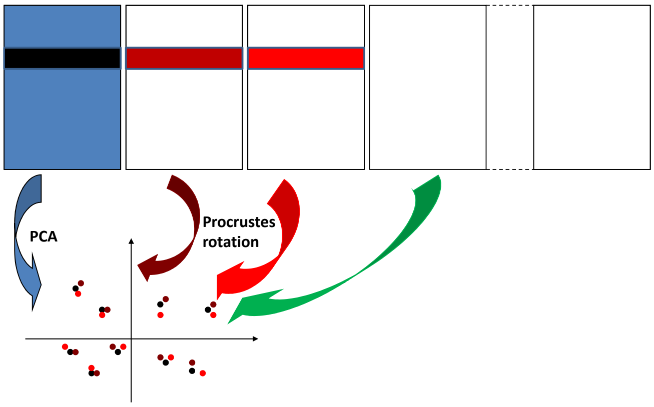


Figure 5.1: Procrustes rotations of the rows of the PCA estimators onto the initial configuration. The first table corresponds to X and the others to $\hat{\mu}^{(S)^1}$, ..., $\hat{\mu}^{(S)^\star}$ obtained either by the asymptotic, the bootstrap or the jackknife strategies. PCA is applied on X to get the representation of the rows of $\hat{\mu}^S$ (black dots), then the rows of each $\mu^{(S)^\star}$ are rotated and represented (red dots, brown dots, etc).

Wahba (1979a). Josse and Husson (2011b) showed that the prediction error $(x_{ij} - \hat{\mu}_{ij}^{(S)^{(-ij)}})$ for a cell (ij) can be approximated by

$$x_{ij} - \hat{\mu}_{ij}^{(S)^{(-ij)}} \simeq \frac{x_{ij} - \hat{\mu}_{ij}^{(S)}}{1 - P_{ij,ij}}. \tag{5.5}$$

Thus, we can define an approximation to the jackknife where instead of deleting the cell (ij), we replace it with the estimate $\hat{\mu}_{ij}^{(S)^{(-ij)}}$ obtained using (5.5) and then perform PCA on this altered matrix. The pseudo-values are computed as in (5.4).

5.1.2 Visualizing confidence areas

All our proposed methods effectively provide a list of pseudo-realizations of the PCA estimator. The spread of these pseudo-realizations can then give us an idea of the stability of PCA. In order to make good use of the pseudo-realizations, we combine them using Procrustes rotation (Gower and Dijksterhuis, 2004b, Krzanowski, 2010, Schoenemann, 1966, Lebart, 2007) as illustrated Figure 5.1 and visualize the results with confidence areas. Procrustes rotations are often used in the context of principal components to compare configuration since they allow to get rid off the potential issues with the different sign for the eigenvectors, or rotations of configuration. In its simplest version, it consists in finding the best translation and rotation that makes two configurations coincides. Then, we define $\breve{\mu}^{(S)^b}$ as

$$\breve{\mu}^{(S)^b} = \hat{\mu}^{(S)^b} R^b \text{ where } R^b = \text{argmin}_R \left\{ \left\| \hat{\mu}^{(S)} - \hat{\mu}^{(S)^b} R \right\|^2 : R'R = \mathbb{I}_p \right\}, \tag{5.6}$$

5.1.3 Results

To assess our proposed methods, we ran simulations with different settings (n/p: 100/20, 50/50 and 20/100; S: 2, 4, the ratio d_1/d_2: 4, 1 and SNR: 4, 1, 0.8) and we counted the number of times that the "true" individual coordinates (of μ) were inside their confidence ellipses as illustrated in Figure 5.2. As our simulation study made clear, the proposed methods have very different strengths and weaknesses. The asymptotic and the parametric bootstrap both perform similarly although the bootstrap is somewhat better. As expected, both methods provide accurate confidence areas close to the nominal level when the SNR is high and when n is large, but can break down in low SNR settings. Conversely, the jackknife and its approximation behave well in low SNR settings, but can struggle when p is small or when there is not much noise. In general, the exact jackknife is more accurate, but can sometimes be prohibitively expensive computationally. Note that we also assessed the method for a number of dimension greater than 2 and good coverage properties were also observed. However, it raises challenges in term of visualisation. Areas on the 2-dimensional map are areas of variability and no more confidence ones.

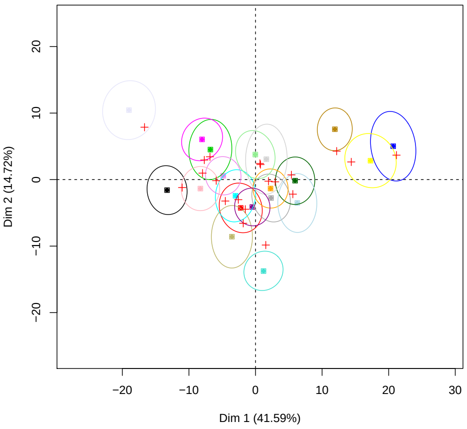
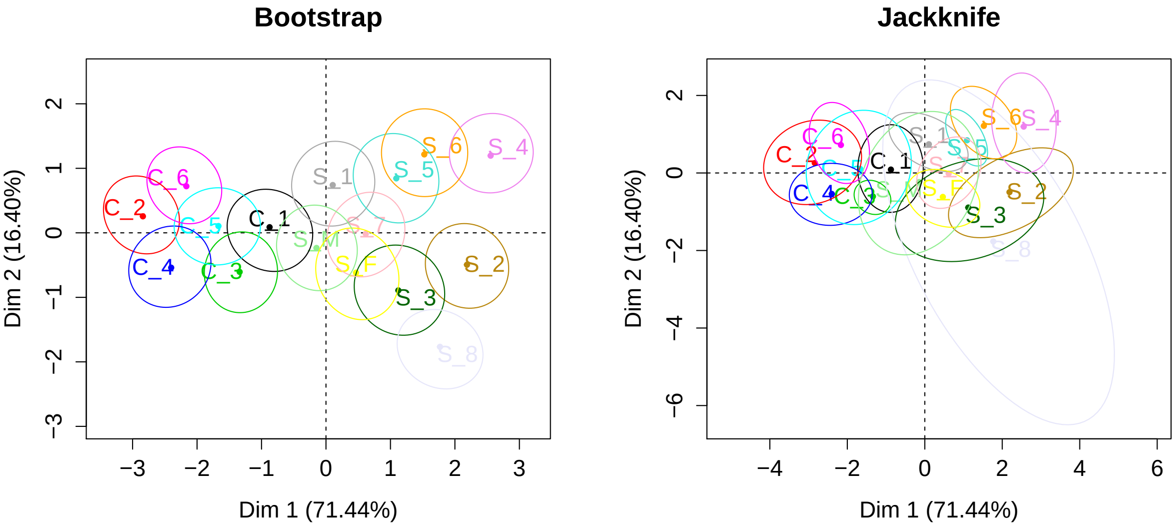


Figure 5.2: True values are represented with a cross. Ellipses are built with the asymptotic method.

The methods are now applied on the GE data as represented in Figure 5.1.3. The confidence ellipses



produced by the asymptotic method and by the bootstrap are all fairly round. In comparison, the ellipses obtained by the jackknife are larger and are much more elongated. In particular, the jackknife appears to suggest that, there is much less signal in the lower-right direction dominated by S8 than in the other directions. Conversely, all methods seem to imply that S4 in the upper-right corner is different from most other genotypes in a significant way.
In order to get a better idea of which method was most trustworthy here, I run a small simulation study built on top of the GE dataset. I use the fitted rank-two PCA matrix as the true signal matrix, and then add isotropic Gaussian noise to it. Results for various noise scales σ are shown in Table 5.1.3. The root-mean residual for the actual PCA fit to the GE data was $\sigma = 0.6$. The small simulation seems to suggest that the bootstrap is slightly anti-conservative here. Thus, the most honest confidence areas for the GE dataset should probably be given by the jackknife ellipses.

	σ	Asympt	Bootstrap	Jack	Approx. Jack.
1	0.05	93.50	93.56	92.56	90.34
2	0.10	94.56	94.47	93.81	91.97
3	0.15	93.47	94.06	94.09	92.25
4	0.30	90.97	93.06	94.66	91.72
5	0.50	84.94	90.81	95.97	91.22
6	0.80	71.72	84.72	97.16	89.44
7	1.00	66.59	81.38	97.16	88.44

Table 5.1: Simulations based on the GE data.

All our methods rely on the validity of the model (5.1), and consequently the choice of the number of dimension S is crucial. In addition, the uncertainty in the number of dimensions S is not taken into account and represents interesting challenges from the point of view of how it can be done and how to visualize confidence areas when S is random.

5.2 A genuine Bayesian strategy

In Chapter 4, we get shrinkage point estimates for the PCA parameters whereas in Section 5.1 of this chapter, we described confidence areas for the biplot representations. Both can be obtained using a Bayesian treatment of the model. We examine such an approach in this section which summarizes the work of Josse *et al.* (2014b) where the focus was on the AMMI models (4.1) in the framework of genotype-by-environment (GE) data. In such a setting, the Bayesian approach is not opportunist since prior information on the phenomenon under study is available: breeders often have a good idea about what the average of the yield should be, what the genetic and environmental variance is and even about the magnitude of the interaction, when some historical information is available.

5.2.1 Introduction

In AMMI models (4.1), we recall that the distribution of the data is:

$$\begin{aligned} \left[y_{ij} \mid \mu_{ij}, \sigma^2\right] &\sim \mathcal{N}(\mu_{ij}, \sigma^2) \\ \text{with } \mu_{ij} &= g + \alpha_i + \beta_j + \sum_{s=1}^{S} \sqrt{d_s} q_{is} r_{js} \end{aligned} \tag{5.7}$$

One difficulty of a Bayesian treatment of such models is to work in an overparametrization framework [2], *i.e.*, to ensure that priors and posteriors on the parameters take into account constraints of models (classical constraints are considered for the main effects such as $\sum \alpha_i = 0$ and orthonomality constraints for the Q and R). Difficulties specifically arise for the interaction terms. Smidl and Quinn (2007) and Hoff (2009), in the framework of PCA, imposed a Uniform prior on the matrices of the left and right singular vectors $Q_{I\times S}$ and $R_{J\times S}$. Such Uniform distributions are special cases of von Mises-Fisher (VMF) distributions (Khatri and Mardia, 1977) which are distributions over the Stiefel manifold representing the set of orthonormal matrices (Chikuse, 2003). Using such priors, which meet the constraints also ensures orthonormality constraints at the posterior level, because the conditional posterior distributions for the matrices Q and R are also VMF distributions (but no more Uniforms). Since, no closed-form expression is available for the joint posterior distribution, the authors implemented a specific Gibbs sampler.

In Josse *et al.* (2014b), we introduced a Bayesian treatment of AMMI models based on another solution to deal with overparametrization in bilinear models. Our proposal has the advantage of being easily implementable in the standard BUGS (Bayesian inference Using Gibbs Sampling) softwares (WinBUGS/OpenBUGS/JAGS) and it is not necessary to implement a specific Gibbs algorithm. The rationale of the approach and the definition of the priors are presented in Section 5.2.2. In Section 5.2.3, I highlight the benefits of using a Bayesian point of view to answer practical questions raised when analyzing GE data.

5.2.2 Rationale of the approach

Prior distributions on the parameters

The rationale of the suggested approach is to get rid of the overparametrization issue simply by disregarding the constraints at the level of the priors. More precisely, priors are defined for the complete set of parameters (on the μ, α_i, β_j, d_s, q_{is} and r_{js}) without considering the constraints. This means that contrary to the previous attempts (Viele and Srinivasan, 2000, Perez-Elizalde, Jarquin, and Crossa, 2011), the orthonormality constraints for the interaction terms are not ensured at the prior level. We thus call our approach the *Unconstrained Bayesian AMMI* approach. The following independent prior distributions are suggested:

$$\begin{aligned}
g &\sim \mathcal{N}\left(m, s_g^2\right) \\
\alpha_i &\sim \mathcal{N}\left(0, s_\alpha^2\right) \\
\beta_j &\sim \mathcal{N}\left(0, s_\beta^2\right) \\
(d_s)_{s=1\ldots S} &\sim \text{ordered sample of } S \text{ independent } \mathcal{N}^+\left(0, s_\lambda^2\right) \qquad (5.8)\\
q_{1s} &\sim \mathcal{N}^+(0,1) \ \text{ for } s = 1, ..., S \\
q_{is} &\sim \mathcal{N}(0,1) \ \text{ for } i > 1 \text{ and for } s = 1, ..., S \\
r_{js} &\sim \mathcal{N}(0,1) \ \text{ for } j \geq 1 \text{ and for } s = 1, ..., S \\
\sigma &\sim U\left(0, S_{ME}\right)
\end{aligned}$$

where $\mathcal{N}^+$ stands for the truncated Normal distribution on the positive values and U is the Uniform distribution. Any supplementary information available from experts or from historical data can be taken into account by assigning specific values to the constants.

[2] Overparametrization can be defined as follows: there is overparametrization when some restrictions on the parameters do not modify the likelihood. A mathematical definition about restrictions on the parameters can be found in Silvey (1975) p79. More discussion can be found in Josse *et al.* (2014b).

Prior distribution on the data

The impact of the chosen priors on the parameters (5.8) can be assessed by inspecting the prior distribution which is implicitly assumed for the data y_{ij}. However, due to the presence of the bilinear terms in model (5.7), no closed-form expression is available. Nevertheless, it is possible to compute explicitly its first two moments (Josse *et al.*, 2014b) and to simulate draws from its prior distribution. It leads to:

$$\begin{array}{rcll}
E\left(y_{ij}\right) & = & m & \\
V\left(y_{ij}\right) & = & s_g^2 + s_\alpha^2 + s_\beta^2 + Qs_\lambda^2 + \frac{1}{3}S_{ME}^2 & \\
Cov\left(y_{ij}, y_{ij'}\right) & = & s_g^2 + s_\alpha^2 & j \neq j' \\
Cov\left(y_{ij}, y_{i'j}\right) & = & s_g^2 + s_\beta^2 & i \neq i' \\
Cov\left(y_{ij}, y_{i'j'}\right) & = & s_g^2 & j \neq j' \text{ and } i \neq i'
\end{array}$$

Inspecting the priors on y_{ij} is an important step in the analysis. Indeed, it is a way for the user to better understand which *a priori* information is included in the analysis. For instance, this allows seeing which range of values for yield are *a priori* considered for the genotypes and the environments. Thus, if the user is not satisfied with these values, he can adjust the constants in the definition of the priors (5.8).

Note that the suggested priors on Q, R and $(d_s)_{s=1,\ldots,S}$ lead to a specific prior for the interaction part of each cell represented by the term $\sum_{s=1}^{S} \sqrt{d_s} q_{is} r_{js}$: all interaction terms have the same expectation and variance while the covariance between terms is zero.

Posterior distributions

Multiplying the likelihood defined by the distribution of the data (5.7) and the priors (5.8) gives the joint posterior distribution of $(y_{ij}, g, \alpha_i, \beta_j, d_s, q_{is}, r_{js}, \sigma)$. As is often the case, no closed-form expression exists. Here, it is not necessary to build and implement a specific Gibbs sampler all that is required is the use of a standard software for Bayesian methods such as JAGS (Martyn, 2003).

As priors are defined for an overparameterized setup, the posteriors relate to the same overparameterized setup: they do not comply with the constraints (the posterior distributions for the interaction terms do not meet the orthonormality constraints). To facilitate the interpretation of the results, the suggested solution consists in considering and working with functions of the parameters that are identifiable (Silvey, 1975). More precisely, we consider the expectation of the data μ_{ij} (defined in (5.7)). Since a posterior distribution is available for all of the parameters (all the g, α_i, β_j, q_{is} and r_{js}), a posterior distribution is also available for all the μ_{ij}. It is possible to get a Markov Chain Monte Carlo (MCMC) sample of size M to estimate the posterior distribution. This means that M matrices of size $I \times J$ are available as draws from the posterior distribution of all the μ_{ij} (respecting the μ_{ij} formulae (5.7)). Thus, it is possible to apply a postprocessing to each matrix ($m = 1, ..., M$) carrying out the classical procedure (in accordance with the chosen constraints): each matrix is centered by row and by column and a SVD is applied to the resulting matrix. Consequently, for each m, new parameters $(g, \boldsymbol{\alpha}, \boldsymbol{\beta}, Q, R, (d_s)_{(s=1,\ldots,S)})$ meeting the constraints are available, which are draws from the posterior distribution for the parameters. Note that the posterior distribution of the expectation μ_{ij} is the same whatever the post-processing applied. Consequently, the postprocessing step is a “neutral" operation and the properties of the posterior are the same when we restrict ourselves to the cell mean parameters μ_{ij}.

5.2.3 Results

Simulations

In the simulation study in Josse *et al.* (2014b), we highlighted that, as expected, when we use as point estimates the mean of the draws from the posterior distribution, we end up with estimators with better MSE than the usual maximum likelihood estimates due to the shrinkage property. In addition, we assessed the influence of the priors with sensitivity analysis (by varying the variance of the priors) and also the impact of selecting a wrong number of dimensions S the method is presently defined for fixed a number of interaction terms). Finally, we briefly discussed the missing values case. Indeed, it is common to hear that "missing values are not a problem in a Bayesian analysis" [3]. Indeed, it is possible to easily carry-out the procedure despite missing values and to get draws from the posteriors. Thus, it is also possible to get "point estimates" and their variability (given by the standard deviation of the posterior distribution) for the parameters despite missing values.

Real data set

We analyse the GE data of Section 4.1.1. In Josse *et al.* (2014b), we showed how the Bayesian approach may help in answering important questions arising in the context of analyses of GE interactions. We considered the five basic following questions: Q1) What is the genotype with the best performance across environments? Q2) What is the genotype with the best performance for a specific environment? Q3) Are some genotypes stable across environments? Q4) Is it possible to rank the genotypes? Q5) Can we estimate the probability that a genotype produces less than a certain threshold? To answer these questions, we also proposed some additional graphs. Here, I present an extract from the analysis.
Assuming that no information is available from experts, the following vague priors are used to illustrate the method: $m = 5$, $s_g = 0.8$, $s_\alpha = 0.5$, $s_\beta = 0.5$, $s_\lambda = 0.5\sqrt{IJ}$ and $S_{ME} = 2$ for definition (5.8). The Bayesian approach is performed with one chain of size 110,000 and a thinning period of 100. A sample of 1,000 draws is thus available to assess the posterior distributions.
Figure 5.3 shows the prior (horizontal axis) and posterior (vertical axis) distributions of the μ_{ij} $(i = 1, ..., 16; j = 1, ..., 10)$. Instead of directly representing the 1000 values drawn using the algorithm, we built credible boxes using the 2.5% and 97.5% quantiles of the distributions. For instance, the box at the top of the figure corresponds to the credible box of μ_{21}. This representation is helpful to see both the impact of the parameters' priors (defined in (5.8)) on the μ_{ij} priors (looking at the coordinates on the horizontal axis) as well as the level of involvement of the dataset in the posteriors (looking at the coordinates on the vertical axis). Here, the priors on the parameters lead to priors on the μ_{ij} outside the permissible range (even negative yields) which suggests to reduce the variability of the priors. Consequently, the following constants are used for the subsequent outputs: $s_g = 0.5$, $s_\alpha = 0.3$ and $s_\beta = 0.3$, $s_\lambda = 0.25\sqrt{IJ}$. More relevant prior information (with historical information for instance) would minimize the probability of negative yield estimates. Figure 5.3 also shows that the data bring substantial information. Indeed, despite the relatively uninformative values for the priors, the posteriors μ_{ij} (which are obtained combining the priors and the data) have values between 0 and 10. In addition, no structure is visible in the prior dimension; this is not the case for the posterior dimension where three distinct clusters appear.
Figure 5.4 focuses on the genotype effect. For each genotype i, each point has as coordinates the pair (main effect term, first interaction effect term): $\left(\alpha_i^m \sqrt{J}, \sqrt{d_1}^m q_{i,1}^m\right)$ with $m = 1, ..., 1000$.

[3] Of course, I am a bit skeptic about this claim.

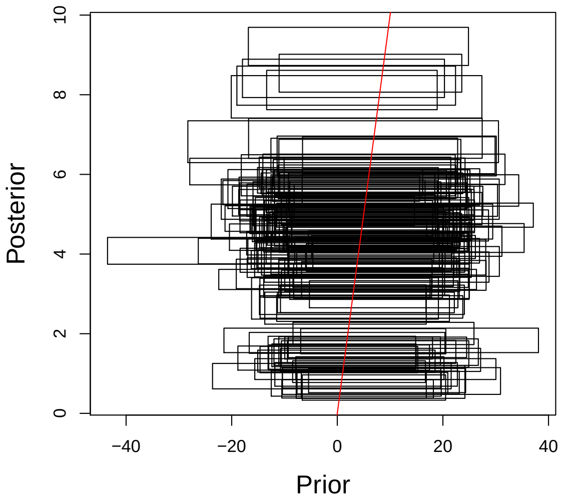


Figure 5.3: Prior-posterior credible boxes for the μ_{ij} $(i = 1, ..., 16\ ;\ j = 1, ..., 10)$. The red line corresponds to the equality between the abscissa and the ordinate.

There are 1000 points representing the posterior distribution which provide a direct view of the variability. The weighting of the parameters by $\sqrt{J}$ and $\sqrt{d_1}$ ensures that their squared norms are equal to the sum of squares (SS) explained by the associated terms in the model ($SS_{Genotype} = J\sum_i \widehat{\alpha}_i^2$ and $SS_{GE(1)} = d_1 = d_1 \sum_i \widehat{q_{i,1}}^2$, respectively) which seems a fair scaling to relate them. Genotype 2, a complete genotype, has a positive main effect (positive values for α_i) whereas

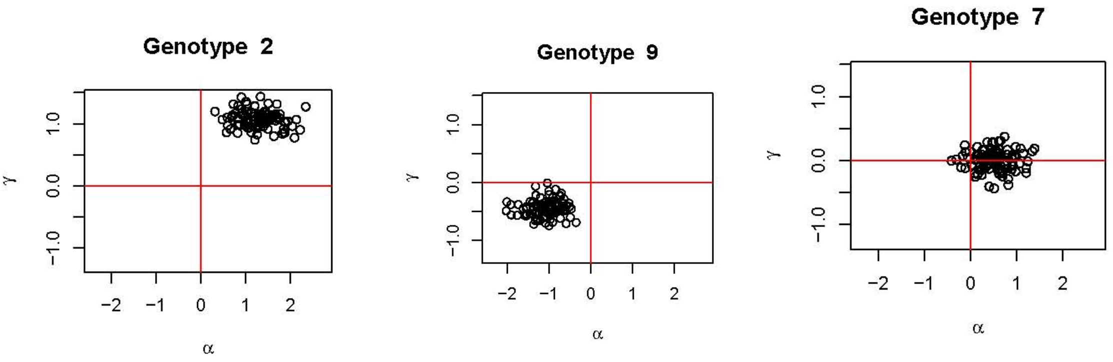


Figure 5.4: Genotype effect: for a subset of genotype $i_{(i=2,7,9)}$ representation of $(\alpha_i^s\sqrt{J}, q_{i,1}^s\sqrt{d_1}^s)$, $s = 1, ...1000$, that is the main effect and interaction of each genotype.

genotype 9, a substituted one, has a negative main effect. Genotype 7 appears to be stable since its values for the interaction terms are the smallest in absolute value (around 0) for the first and second dimension (not shown here). These graphical representations allowed us in Josse *et al.* (2014b) to answer three (Q1, Q3 and Q4) of the questions raised at the beginning of this section. Figure 5.5 focuses on the behavior of the genotypes across environments. For each genotype i, abscissas correspond to mean values for environments and ordinates to the differences between genotype specific performances in particular environments and the mean values for those environments. Thus, for each genotype i, each point has as coordinates $(g^m + \beta_j^m; \alpha_i^m + \sum_{s=1}^S \sqrt{d_s}^m q_{is}^m r_{js}^m)$,

$m = 1, ...1000$. Each panel for a genotype i can be read as follows: the poorest environments (with small β_j) are on the left of the horizontal axis and high values on the vertical axis correspond to high values for the yield for the genotype i in a particular environment. There are strong differences between genotypes: in agreement with what we expected, the complete genotypes 1 to 6 tend to perform better in the poorest environments, whereas the substituted genotypes 8 to 14 show good behavior in the best environments; again this is what we expected on the basis of our biological knowledge of the system, see Royo *et al.* (1993). Indeed, environments with small β_j are environments with acid soil and the authors noted that “the superiority of complete over substituted decreased when the soil pH increased". Genotypes 15 and 16 are very stable across all the environments. These genotypes are official checks with “very low interaction and yield similar or slightly superior to the grand mean". These graphical representations allow us to answer the question Q2 on the local performances *i.e.* to determine which are the best genotypes for a given environment.

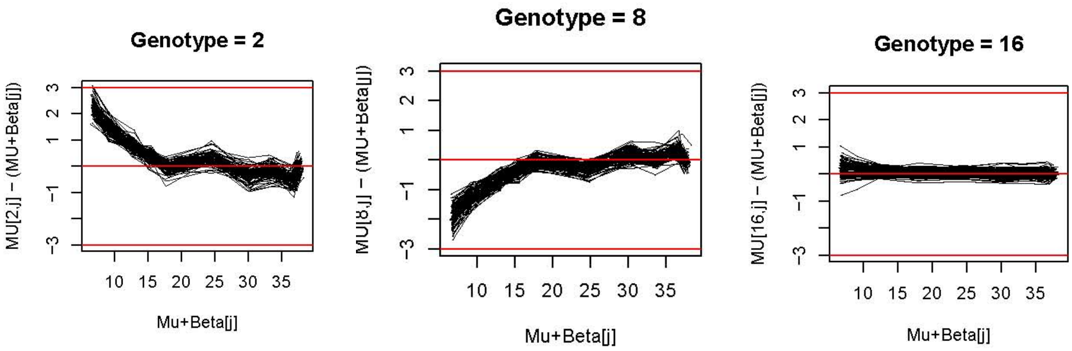


Figure 5.5: Posterior genotype profiles: for each genotype i ($i = 2, 8, 16$), representation of $(\frac{1}{I}\sum_i \mu_{ij}^m; \mu_{ij} - \frac{1}{I}\sum_i \mu_{ij}^m)$, $m = 1, ...1000$, that is the evolution of each genotype according to the potential of the environments as assessed by the environment main effect.

Part III

Works in progress and Perspectives

The ability to easily collect and gather a large amount of data from different sources can be seen as an opportunity to better understand many processes. It has already led to breakthroughs in several application areas. However, due to the wide heterogeneity of measurements and objectives, these large databases often exhibit an extraordinary high number of missing values. Hence, in addition to scientific questions, such data also present some important methodological and technical challenges for data analyst. In, the same way, data reduction, denoising and visualization methods are essential to analyze, to extract relevant information and to have a synthetic view of the underlying phenomena. I wish to continue to invest on these research topics that I consider promising. In addition, my research is still driven by the applications and as statisticians, we are lucky to interact with many scientific fields. In this part, I give a quick overview of my ongoing and future research activities on both principal component methods in Chapter 6 and on missing values in Chapter 7.

Chapter 6

Models for principal components methods

New challenges continuously arise from the evolution of computing and data collection and principal components methods such as correspondence analysis (CA) or multiple correspondence analysis (MCA) have the great advantage of beeing able to visualize large data and of being solved easily by singular value decomposition (SVD) of the data weighted in an appropriate way.
However, there are often motivated and derived using geometric considerations without any references to a probabilistic model in line with Benzécri (1973)'s idea to "let the data speak by themselves." Even so, specific choices of weights and metrics can be viewed as inducing specific models for the data under analysis. Understanding the connections between exploratory multivariate methods and their cognate models can yield insights into the method's properties and may help for inference issues, for selecting tuning parameters, for handling missing values as well, etc. Such relationships have been established for PCA, but there are less attempts for CA and as far a I know no connection has been yet established for MCA.
A part of my current works focuses on this topic. In Section 6.1, I briefly summarize the recently submitted work with Patrick J.F Groenen on "Multinomial MCA" (given in the Appendix A) where we defined a model for categorical variables inspired by MCA and estimated the parameters using a majorization algorithm. Then, I present my ideas to make tighter connections between a multinomial logit model and MCA. In Section 6.2, I focus on low-rank models for count data and their relationship with CA. More precisely, I define a problem of maximization of a penalized likelihood and I am focusing on finding the regularization parameters from the data taken into account the inherent heteroscedasticity.

6.1 Multinomial MCA

This work was motivated by the analysis of a survey data from the French national institute for prevention and health education (INPES [1]) on alcohol usage. Each year, more than 50,000 individuals describe their consumption (kind of beverage, frequency of drinking, frequency of binge drinking, etc) as well as their socio-economic and demographic characteristics. Describing the relationships between these categorical variables is important to monitor alcohol usage in subgroups, to improve the understanding of alcohol usage, to monitor their evolution, and to suggest suitable policies. It is a public health priority since France is paying a heavy price each year to the harmful use of alcohol and it may be preventable. Therefore, scientists need methods

[1] http://www.inpes.sante.fr/default.asp

to explore such data and the task is challenging in part because of the categorical nature of the data and their dimensionality.
On one hand, we have explicit models such as log-linear model (Agresti, 2013, Christensen, 2010) which struggle with high dimensional data since the number of parameters quickly grows with large number of categories and variables. In addition, such models often lack some follow-up graphical representations which may help the user to a great extent to shed more light into the obtained results. On the other hand, MCA is a powerful technique, but clearly lacks from any appropriate modeling of the probability to select a category out of several options.
The *multinomial multiple correspondence analysis* (MMCA model) suggested in Groenen and Josse (2016) can be presented as a fixed-effects latent traits (Lazarsfeld and Henry, 1968) models which aims at bringing the best from both worlds: appropriately modeling the probability of selecting a category out of several options combined with the capability of handling high dimensional data while providing graphical output to explore the relations and gain insight for interpretation. We used a parsimonious low rank representation of the data and derived an efficient majorization algorithm to estimate the parameters. This latter uses an elegant bound for the second derivative derived in unpublished and unfinished notes of De Leeuw (2005). Then, to avoid overfitting issues due to the separability problem inherent of such models, we maximized a regularized maximum likelihood using the nuclear norm. Built on recent results on the selection of the threshold parameters for ℓ_1 type of penalty Giacobino *et al.* (2015), we suggested combining the *universal quantile threshold* to select the rank and cross-validation to determine the amount of shrinkage. Our results enabled us to uncover some interesting insight into the balance between good selection or good prediction properties to select the threshold parameter in lasso regression Tibshirani (1996b).

Although, the previous proposition allows to estimate the parameters of a low rank model for categorical data, it still misses an explicit connection with MCA. Let's define MCA as the correspondence analysis (CA) on the indicator matrix $X_{n\times K}$ by forming the pseudo-residual matrix $Z_X = \frac{1}{\sqrt{nK}}(X - 1(\mathbf{n_k})^T)D_k^{-1/2}$ with D_k the diagonal matrix with the margins $n_{k_{k=1,\dots,K}}$ of X. MCA boils down to low-rank least-squares decomposition of Z_X, $\widehat{\Gamma}_{\text{MCA}} = U_S D_S V_S^{'}$.
In fact it can be shown that the decomposition of MCA can be viewed as a one-step estimate for the cognate model the *multilogit-bilinear* model:

$$\mathbb{P}(x_{ij} = c) = \frac{e^{\theta_{ij}(c)}}{\sum_{c'=1}^{C_j} e^{\theta_{ij}(c')}}, \text{ with } \theta_{ij}(c) = \beta_j(c) + \Gamma_i^j(c) \tag{6.1}$$

and $\Gamma_i^j(c) = \sum_{s=1}^{S} \sqrt{d_s} u_{is} v_{js}(c)$.
The rationale of the approach is the following one. By Taylor expanding the likelihood ℓ of model (6.1) around the *independence model* ($\beta_0 = \log p, \Gamma_0 = 0$) to obtain $\tilde{\ell}(\boldsymbol{\beta}, \boldsymbol{\Gamma})$, a quadratic function of its arguments, then maximizing the latter amounts to a generalized singular value decomposition, which can be performed efficiently. Moreover, the generalized singular value problem is precisely the one we solve to obtain the row and column coordinates MCA ($\beta_0, \widehat{\Gamma}_{\text{MCA}}$). More work has to be done to investigate this result.

6.2 Poisson low rank matrix estimation

Let's consider the case discussed in Section 4.6 of a matrix $X_{n\times p}$ whose entries are independent Poisson counts and with expectation of low-rank:

$$X_{ij} \sim \mathcal{P}(\mu_{ij}), \quad i = 1, \dots, n, \; j = 1, \dots, p,$$

However, here I consider a natural strategy where I would like to estimate the parameters by maximizing the likelihood:

$$\mathcal{L}(\mu) = \sum_{i,j} \left(X_{ij} \log(\mu_{ij}) - \mu_{ij}\right).$$

More precisely, I am interested in estimating a penalized likelihood:

$$\hat{\mu} = \arg\min_{\mu > 0} \{-\mathcal{L}(\mu) + \gamma \|\mu\|_*\}, \tag{6.2}$$

where $\|\mu\|_* = \sum_{s=1}^{S} \sqrt{d_s}$ denotes the nuclear norm of μ. A penalty based on the spectral norm is justified under the low-rank assumption for μ. However, due to my previous works on low rank matrix estimation (Chapter 4), for Gaussian data as well as some numerical experiments, I would like to design a procedure that takes into account the heteroscedastic poisson noise using a specific regularized parameter for each singular value that is learnt from the data:

$$\arg\min_{\lambda \geq 0, \mathbf{u}, \mathbf{v}} \left\{ -\sum_{i,j} \left[X_{ij} \sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} - \exp\left(\sum_{s=1}^{S} \sqrt{\lambda_s} u_{is} v_{js} \right) \right] + \sum_{s=1}^{S} \gamma_s \sqrt{\lambda_s} \right\}$$

To do this, I have different strategies that I would like to investigate. The first one is to extend the works of Ivanoff, Picard, and Rivoirard (2015) that derived such parameters in the framework of regression using concentration inequalities. However, one main problem is that the equivalent of the design matrix in regression for low rank matrix estimation problems depends on the data. Another strategy that could lead to interesting results would be to use a complete Bayesian framework and the works of (Christiansen and Morris, 1997). My idea, would be to put a Gamma *a priori* on each μ_{ij} and then use an empirical Bayesian strategy to estimate the hyper-parameters from the data. I guess, that it would give the appropriate shrinkage terms and thus a good estimator close to the true signal μ. However, there are many technical issues that need to be solved. I am also thinking on defining GCV criterion using the results of S. Wood in the framework of non-linear GAM models coupled with the results of D. Firth and co-authors on non-linear generalized models. On this topic of low-rank models for count data, I am also working on some new insights on the connection (de Falguerolles, 1998, Gower, Lubbe, and Le Roux, 2011) between log-bilinear models (also known as RC models, Goodman, 1985) and CA (4.24).

Chapter 7

Missing values

Modern data analysis has access to wealth of different information which often leads to new objectives with more ambition, the possibility to answer scientific questions with more strength and enable us to hopefully get new discoveries. However, whatever the precaution we take, missing values issues are exacerbated with the amount of available data and they need a particular attention during the statistical analysis. Missing values can be a challenge for all statistical methods in all the fields of application.

I have different ongoing projects on missing values. First, with my post-doc Pavlo Mozharovskyi, who started in September 2015, we are working on single and multiple imputation methods for elliptical distributions. Indeed, many of statistical methods for handling continuous data have been developed based on the assumption of normality, and the machinery for imputation of missing data is not an exception. With Pavlo, we suggest to single imputate the missing entries by maximizing the data depth (Mosler, 2013) (a centrality measure) of the point given its observed values and the data. Let us consider a point $x_0 \in \mathbb{R}^p$ and a sample $X = \{x_1, ..., x_n\}$, a statistical data depth is a function $D(x_0|X) : \mathbb{R}^d \mapsto [0, 1]$ that describes how deep, or central, the observation x_0 is located w.r.t. X. Points closer to the center should have higher depth, and those more outlying smaller one. The most famous depths include the Tukey (or halfspace) depth (Tukey, 1975) and the manahalobis depth. Thus, in the simplest case, we impute data by minimizing the manahalobis distance. The properties of the imputation depend on the depth properties. Our first results highlight interesting behavior: the imputation can be positioned between random forest imputation and regression imputation in the sense that it stays in the data as with the non-parametrics nearest-neighbor imputation while being able to take into account the shape of the global distribution of the data and thus it has "extrapolation properties" close to the ones of regression imputation. Our work also shed lights into the properties of other well-known single imputation methods. For multiple imputation, Pavlo derived a way to draw from conditional distributions, and reflect uncertainty by means of the Markov chain Monte Carlo and a bootstrap. One of the benefits of such methods will also be the ability to deal with data corrupted by both missing and outliers for instance. Note that there are strong connexions between these two topics, that I have not yet investigated.

Then, I am working with Jerry Reiter on a special issue on missing values for the journal Statistical Science. Indeed, following the conference MissData [1] that I organized in June 2015 to gather the main contributors on both on missing values inference and on matrix completion from different fields (social-science, biology, even astrophysics!), I sent an email to the editor of Statistical Sci-

[1] http://missdata2015.agrocampus-ouest.fr/infoglueDeliverLive/

ence to submit my project. My idea is to have a special issue to present the state of the art as well as to spotlight the latests research and challenges on the topic. The following points will be discussed: Dealing with missing values in challenging data (methods for multi-level data, methods for mixed multivariate data, methods for non-iid data); Missing values in big data (matrix completion; impacts on machine learning); Connections to causal inference (potential outcomes; double robustness methods; missing data in clinical trials); Multiple imputation: joint models vs. fully conditional specifications; MNAR data... In fact, there are many questions and challenges, from the theoretical, methodological and practical point of views that need to be addressed. Note also that evolution in the missing values literature are coupled with evolutions in statistics; if a model is available to describe relationship between mixed variables for instance, it can be used for imputation. I mention here some points that I consider would require more attention. First, we can remark, that the theory of D.B. Rubin which is still in used was developed in the 1980's and has not really been modified since. Consequently, the theory was defined under the classical assumption of repeated sampling theory and the asymptotic framework $n \to \infty$ and nothing is available for other settings. In addition, many rules (such as the aggregation rules of multiple imputation) are used just because they "work", even when being far from their scope of applicability, and this need more investigation. Finally, when analyzing data, we have often many kinds of missing data in a same data with different patterns which may be important to take into accounts. I can also mention an ongoing work with Sylvain Sardy and his PhD Student P. Descloux on "Model selection with Lasso when some data are missing in the design matrix" as well as some first attempts on the topic of dealing with missing values in non-standard inferential tasks. Principal components can be considered as such a case, I am now considering clustering methods.

Concerning long time perspectives, I would like to investigate more the role of multiple imputation (MI) as an universal tool to deal with missing values in the framework of large databases. It is of course interesting to have "one" tool that can be used in many different cases. MI does not pretend to be the best tool for each situation but instead to be quite appropriate for many cases. Nevertheless, I have doubts about its future. Recently, I was contacted to be involved in studying a (beautiful) database on the multiple trauma patients in Paris. The data have 7000 patients and 200 variables and the data come from different hospitals. It is full of missing values of all the possible kinds, and considering multiple imputation here seems difficult. To start investigating this line of research, I have planned to supervise a PhD student starting in September 2016 with Arnaud Guyader. The aim is to first examine the actual practices of MI. Then, we could design rules more in agreement with the available data and suggest alternatives to multiple imputation when possible with an associated implementation. Of course, supervising students is a great rewarding experience where both part learn to a great extent. It is also important to make the subject lively and to transfer skills and enthusiasm for this job.

Appendix A

Multinomial MCA

Multinomial Multiple Correspondence Analysis

Patrick J.F. Groenen and Julie Josse

March 10, 2016

Abstract

Relations between categorical variables can be analyzed conveniently by multiple correspondence analysis (MCA). The graphical representation of MCA results in so-called biplots makes it easy to interpret the most important associations. However, a major drawback of MCA is that it does not have an underlying probability model for an individual selecting a category on a variable. In this paper, we propose such probability model called multinomial multiple correspondence analysis (MMCA) that combines the underlying low-rank representation of MCA with maximum likelihood. An efficient majorization algorithm that uses an elegant bound for the second derivative is derived to estimate the parameters. The proposed model can easily lead to overfitting causing some of the parameters to wander of to infinity. We add the nuclear norm penalty to counter this issue and discuss ways of selecting regularization parameters. The proposed approach is well suited to study and vizualise the dependences for high dimensional data.

1 Introduction

Data sets with categorical variables are common in many fields such as social sciences, where surveys with categorical questions are conducted. Although some models are available to describe the dependence between categorical variables, they suffer from estimation issues as the number of parameters quickly grows with large number of categories and variables.

To give a concrete example let us consider a data set from the French national institute for prevention and health education (INPES [1]) on alcohol usage. Each year, more than 50,000 individuals describe their consumption (kind of beverage, frequency of drinking, frequency of binge drinking, etc) as well as their socio-economic and demographic characteristics. Describing the relationships between these categorical variables is important to monitor alcohol usage in subgroups, to improve the understanding of alcohol usage, to monitor their evolution, and to suggest suitable policies. Therefore, scientists need methods to explore such data.

[1] http://www.inpes.sante.fr/default.asp

High dimensional data like this one do not fall within the scope of classical models such as the log-linear models Christensen [2010]. In addition, such models often lack some follow-up graphical representations which may help the user to a great extent to shed more light into the obtained results. On the other hand, principal component methods based such as multiple correspondence analysis (MCA) Greenacre and Blasius [2006] are powerful techniques to explore and visualize large categorical data using biplot representation. MCA has the great advantage of being easily solved by singular value decomposition (SVD). However, MCA is often motivated by geometrical considerations without any reference to probability models.

Our *multinomial multiple correspondence analysis* model aims at bringing the best from both worlds: appropriately modeling the probability of selecting a category out of several options combined with the capability of handling high dimensional data while providing graphical output to explore the relations and gain insight for interpretation. We use a parsimonious low rank representation of the data and derive an efficient majorization algorithm to estimate the parameters. This latter uses an elegant bound for the second derivative derived in unpublished and unfinished notes of De Leeuw [2005]. Then, to avoid overfitting issues due to the separability problem inherent of such models, we maximize a regularized maximum likelihood using the nuclear norm.

Built on recent results on the selection of the threshold parameters for ℓ_1 type of penalty Giacobino et al. [2016], we suggest combining the *universal quantile threshold* to select the rank and cross-validation to determine the amount of shrinkage. Our results enable us to uncover some interesting insight into the balance between good selection or good prediction properties to select the threshold parameter in lasso regression Tibshirani [1996].

The remainder of this paper is organized as follows. After a discussion of related work, we describe in Section 2 the *multinomial multiple correspondence analysis* model. In Section 3, we then present the minimization-majorization algorithm to estimate model's parameters. It is shown that the model and algorithm can easily be extended to allow for missing values. A notorious problem for this type of model is the occurrence of parameters wandering off to infinity when the estimated probabilities get close to one. This form of overfitting is avoided by adding a nuclear norm penalty. We explain in Section 5 our new procedure to select the penalty parameter.

1.1 Related Work

The log-linear model Agresti [2013], Christensen [2010] is the golden standard to study the relationship between categorical variables. However, it encounters difficulties with large number of variables and categories since many cells of the contingency table are equal to zero. Unsaturated log-linear models with main effects and two-way interactions could be used to restrict the number of estimated parameters, but the total number of parameters could still be substantial with many categories. One popular alternative consists of latent variable models that summarize the relationship between the given variables by a small number of latent ones, either categoricals or continuous. The former case, known as latent class models Goodman [1974], boils down to unsupervised clustering for one latent variable and nonparametric Bayesian extensions have recently been proposed Dunson and Xing [2009], Bhattacharya and Dunson [2012] to get rid of the difficult choice of the number of clusters. Our approach, MMCA, can be presented as a fixed effects latent traits models able to handle many latent variables. Other related models were studied by De Leeuw [2006], Li and Tao [2013], Collins et al. [2001], and Buntine [2002] but dedicated to either binary data or random effects.

Finally, another popular approach to examine the relationship between categorical variables is multiple correspondence analysis also known as homogeneity analysis or dual scaling Michailidis and De Leeuw [1998], Nishisato [1980], De Leeuw [2014], le Roux [2010], Greenacre and Blasius [2006]. MCA can be seen as the counterpart of PCA for categorical data and involves reducing data dimensionality to provide a subspace that best represents the data in the sense of maximizing the variability of the projected points. As mentioned, it is often presented without any reference to probabilistic models, in line with Benzécri [1973]'s idea to "let the data speak for itself."

As our model is inspired by the MCA representation, we first briefly discuss how MCA is defined. Consider a dataset with n rows and J categorical variables, with K_j categories each, $j = 1, ..., J$. The data are coded using the $n \times K$ super indicator matrix of dummy variables denoted by $\mathbf{G}$ with $K = \sum_j K_j$ and $g_{ijk} = 1$ if person i chooses category k of variable j and $g_{ijk} = 0$ otherwise. A simple example of such a matrix $\mathbf{G}$ with $n = 10$, $J = 3$ variables with respectively $K_1 = 3, K_2 = 3$, and $K_3 = 2$ categories is given by

$$\mathbf{G} = [\mathbf{G}_1|\mathbf{G}_2|\mathbf{G}_3] \quad = \quad \left[\begin{array}{ccc|ccc|cc}
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1
\end{array}\right].$$

MCA, as all the principal component methods can be derived by performing the SVD of matrices with specific row and column weights. The choice of weights ensures the property of the method such as the Chi-square interpretation of the distances between rows as well as the fact that the first principal component of MCA is the variable the most related to all the categorical variables in the sense of the R^2 of the analysis of variance which strengthen the presentation of MCA as an extension of PCA. More precisely, MCA is obtained by performing the generalized SVD Greenacre [1984] of the triplet *data, column weights, row weights* $\left(\mathbf{JG}, J^{-1}{\mathbf{D}_c}^{-1/2}, n^{-1}\mathbf{I}_n\right)$ with $\mathbf{D}_c$, the diagonal matrix with category frequencies and $\mathbf{J}_{n\times n} = (\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}')$ the row-centering matrix. It boils down to performing the following SVD: $\mathbf{J}\mathbf{G}' = \tilde{\mathbf{U}}\boldsymbol{\Lambda}^{1/2}\tilde{\mathbf{V}}'$ with $\tilde{\mathbf{U}}'(n^{-1}\mathbf{I}_n)\tilde{\mathbf{U}} = \mathbf{I}$ and $\tilde{\mathbf{V}}'(J^{-1}{\mathbf{D}_c}^{-1/2})\tilde{\mathbf{V}} = \mathbf{I}$. MCA can also be defined as finding the best low rank approximation of $\mathbf{J}\mathbf{G}$ with a matrix of rank p according to the Hilbert-Schmidt norm $\| \mathbf{T} \|^2_{\mathbf{D}_c^{-1/2}, \frac{1}{n}\mathbf{I}_n} = \operatorname{tr}\left(\mathbf{T}\mathbf{D}_c^{-1/2}\mathbf{T}'\frac{1}{n}\mathbf{I}_n\right)$:

$$L_{\text{MCA}}(\mathbf{X}, \mathbf{A}) \quad = \quad \|\mathbf{J}\mathbf{G} - \mathbf{X}\mathbf{A}'\|^2_{\mathbf{D}_c^{-1/2}, \frac{1}{n}\mathbf{I}_n}$$

with $\mathbf{A}' = [\mathbf{A}'_1| \dots |\mathbf{A}'_J]$ and $\mathbf{A}_j$ the $K_j \times p$ matrix representing the K_j categories of variable j. The solution is given by $\mathbf{A} = \tilde{\mathbf{V}}\boldsymbol{\Lambda}^{1/4}$ and $\mathbf{X} = \tilde{\mathbf{U}}\boldsymbol{\Lambda}^{1/4}$ truncated at order p.

Thus, the category k chosen by person i on variable j can modeled by

$$\hat{g}_{ijk} \approx \mu_{jk} + \mathbf{x}'_i\mathbf{a}_{jk} \tag{1}$$

with μ_{jk} the main effect for category k of variable j, $\mathbf{x}'_i$ row i of $\mathbf{X}$ and $\mathbf{a}'_{jk}$ row k of $\mathbf{A}_j$. Equation (1) is called the *reconstruction formula* in the MCA literature. Tenenhaus and Young [1985] showed that the row sums of $\hat{\mathbf{G}}$ is equal to 1, which implies that the fitted values can be seen as degree of membership to the associated category. However, negative values may occur. Therefore, $\hat{g}_{ijk}$ cannot be interpreted as the probability of an individual i to select category k of variable j.

2 Multinomial Multiple Correspondence Analysis

To develop a maximum likelihood approach, we consider the probability π_{ijk} of person i choosing category k of variable j. To do so, a natural candidate is the multinomial logit or the so-called softmax function, that is,

$$\pi_{ijk} = \text{softmax}(\boldsymbol{\theta}_i) = \frac{\exp(\theta_{ijk})}{\sum_{\ell=1}^{K_j}\exp(\theta_{ij\ell})}, \tag{2}$$

where the θ_{ijk} denotes a utility that person i attaches to category k of variable j. In MMCA, the θ_{ijk} is modeled by

$$\theta_{ijk} = \mu_{jk} + \mathbf{x}'_i\mathbf{a}_{jk},$$

very much in the same way as MCA (1) in the sense that the rank of the interaction is constrained to be p. By assuming independence between all answers of individuals on all variables and conditional independence between variables given the parameters θ, the joint maximum likelihood of MMCA is

$$\prod_{i=1}^{n}\prod_{j=1}^{J}\prod_{k=1}^{K_j}\pi_{ijk}^{g_{ijk}}.$$

For maximum likelihood, one often minimizes the deviance by taking minus the logarithm of the probabilities that is:

$$
\begin{aligned}
L(\boldsymbol{\mu}, \mathbf{X}, \mathbf{A}) &= -\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K_j} g_{ijk}\log(\pi_{ijk}) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K_j} g_{ijk}\log\left(\frac{\exp(\mu_{jk} + \mathbf{x}_i'\mathbf{a}_{jk})}{\sum_{\ell=1}^{K_j}\exp(\mu_{j\ell} + \mathbf{x}_i'\mathbf{a}_{j\ell})}\right)
\end{aligned}
$$

Without any restrictions, the parameters are not identified. Therefore, we use the following identification constraints:

- $\boldsymbol{\mu}_j'\mathbf{1} = 0$ as adding a constant per column does not change π_{ijk},
- $\mathbf{1}'\mathbf{X} = \mathbf{0}$ to avoid main effects estimated by $\mathbf{x}_i'\mathbf{a}_{jk}$,
- $\mathbf{X}'\mathbf{X} = n\mathbf{I}$ to take care of the rotational indeterminacy between $\mathbf{X}$ and the $\mathbf{A}_j$, and
- $\mathbf{1}'\mathbf{A}_j = \mathbf{0}'$ as adding a constant per column does not change π_{ijk}.

Note that the maximum dimensionality is $p^* = \min(n-1, \sum_{j=1}^{J} K_j - J)$. When $p = p^*$, we have a saturated model with all $\theta_{ijk} \to \infty$ for $g_{ijk} = 1$ and $\theta_{ijk} \to -\infty$ for $g_{ijk} = 0$ because all data points can be perfectly estimated. This problem does not only occur in maximum dimensionality. Even when $p \leq p^*$, we often find that several estimates for π_{ijk} approaching one so that $\theta_{ijk} \to \infty$ when minimizing the deviance $L(\boldsymbol{\mu}, \mathbf{X}, \mathbf{A})$. Consider Figure 1 that shows $-\log(\pi_{ijk})$ for individual j and three categories represented by the vertices of the equilateral triangle. The direction of the left vertex gives an infimum of zero, that is, the further in that direction, the closer $-\log(\pi_{ijk})$ gets to zero. Therefore, there is an attraction to θ_{ijk} becoming ever larger.

This effect can be seen as a form of overfitting and a natural solution to tackle it is to add a regularization term that controls the size of the parameters in a penalized likelihood approach. To do so, we need to reparamatrize $\mathbf{XA}'$ in an SVD-type manner, that is, $\mathbf{XA}' = \mathbf{UDV}'$ with $\mathbf{U}$ the $n \times p$ matrix with $\mathbf{U}'\mathbf{U} = \mathbf{I}$ and $\mathbf{1}'\mathbf{U} = \mathbf{0}$, $\mathbf{D}$ the diagonal $p \times p$ matrix with $d_{ss} \geq 0$, and $\mathbf{V}$ the $(\sum_j K_j) \times p$ matrix with $\mathbf{V}'\mathbf{V} = \mathbf{I}$, $\mathbf{V}_j'\mathbf{1} = \mathbf{0}$, $\mathbf{V}_j' = [\mathbf{v}_{j1}, \ldots, \mathbf{v}_{jK_j}]$ represents all categories of variable j and $\mathbf{v}_{jk}'$ is row jk of $\mathbf{V}$

One extensively used and studied penalty in the framework of singular values decomposition based methods is the use of the nuclear norm penalty Fazel [2002], Srebro [2004], Candes et al. [2013] which is equal to the sum of the singular values $\sum_{s=1}^{p^*} d_{ss}$, leading to the penalized deviance:

$$
L(\boldsymbol{\mu}, \mathbf{U}, \mathbf{D}, \mathbf{V}) = -\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K_j} g_{ijk}\log(\pi_{ijk}) + \lambda\left(\sum_{s=1}^{p^*} d_{ss}\right) \tag{3}
$$

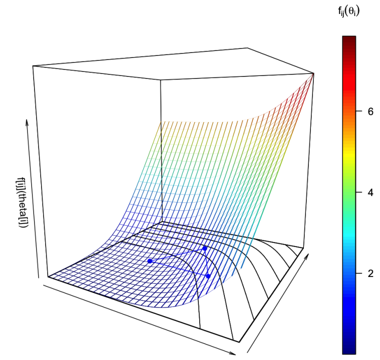


Figure 1: Minus the log likelihood of one observation, $-\log(\pi_{ijk})$, for a variable with three categories, each of them shown by a vertex of the equilateral triangle, spanning a two-dimensional space. The direction in which the left one is pointing, is the direction that asymptotically approaches a probability of one. Without additional constraints, minimization of $-\log(\pi_{ijk})$ will lead to this θ_{ijk} tending to (very) large values.

with

$$\pi_{ijk} = \frac{\exp(\mu_{jk} + \mathbf{u}_i'\mathbf{D}\mathbf{v}_{jk})}{\sum_{\ell=1}^{K_j} \exp(\mu_{j\ell} + \mathbf{u}_i'\mathbf{D}\mathbf{v}_{j\ell})}$$

Whenever $p = p^*$, $L(\boldsymbol{\mu}, \mathbf{U}, \mathbf{D}, \mathbf{V})$ is a convex function minimized over a convex set that has a global minimum for any choice of $\lambda > 0$. For sufficiently large values of λ, the impact of the penalty is to set some of the smallest singular values to zero and thus often results in a lower rank solution. In addition to automatic rank selection, the nuclear-norm also shrinks the non-null singular values.

In Section 5, we will provide more details on a procedure for selecting the threshold parameter λ. In the next section, a majorizing algorithm is derived for minimizing (3).

3 Majorization

Majorization algorithms share as the most important property that they have a guaranteed descent, that is, in each iteration the objective function improves. Under this name, majorization was first proposed by De Leeuw and Heiser [1977]. Since some time, it is probably better known under the name MM, minimization by majorization or maximization by minorization Lange [2004]. The principle is quite simple: in each iteration an auxiliary function (called the majorizing function) $g(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ is set up that satisfies the following requirements

1. $f(\boldsymbol{\theta}_0) = g(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$,

2. $f(\boldsymbol{\theta}) \leq g(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$,

where the current estimate $\boldsymbol{\theta}_0$ is called supporting point and f is the original function to be minimized. In practice, $g(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ only takes simple forms such as linear or quadratic so that its minimum $\boldsymbol{\theta}^+$ is easy to find. At its minimum $\boldsymbol{\theta}^+$ of the majorizing function $g(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$, we necessarily have that $f(\boldsymbol{\theta}^+) \leq g(\boldsymbol{\theta}^+, \boldsymbol{\theta}_0)$. This leads to the so called sandwich inequality

$$f(\boldsymbol{\theta}^+) \leq g(\boldsymbol{\theta}^+, \boldsymbol{\theta}_0) \leq g(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = f(\boldsymbol{\theta}_0)$$

with $\boldsymbol{\theta}^+ = \text{argmin } g(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ proving that an update of the majorizing function also improves f until no improvement is possible.

The advantage of majorization is that in contrast to line search methods such as steepest descent or (quasi-)Newton methods, there is no need for a possibly computationally expensive steplength procedure to guarantee descent.

To derive the majorizing algorithm, the following steps are taken. First, a quadratic majorizing function is derived for the elements of the deviance denoted $f_{ij}(\boldsymbol{\theta}_i)$. Then, a majorizing function is given for the penalized deviance $L(\boldsymbol{\mu}, \mathbf{U}, \mathbf{D}, \mathbf{V})$ in (3). To choose λ, we wish to use cross-validation and thus we need to be able to minimize the penalized deviance in the presence of missing values. This requires an additional majorization function describe in a third step. The last step consists for each of the four sets of parameters in deriving an update for the parameters. Finally, an overview of the entire algorithm is presented.

First step: main majorizing function.
The first step is finding a majorizing function for a single term of the deviance function

$$f_{ij}(\boldsymbol{\theta}_i) = -\sum_{k=1}^{K_j} g_{ijk} \log(\pi_{ijk}) = -\sum_{k=1}^{K_j} g_{ijk} \log\left(\frac{\exp(\theta_{ijk})}{\sum_{\ell=1}^{K_j} \exp(\theta_{ij\ell})}\right). \tag{4}$$

To do so, an explicit expression of its first derivative is needed

$$\nabla f_{ij}(\boldsymbol{\theta}_i) = \mathbf{g}_{ij} - \boldsymbol{\pi}_{ij}, \tag{5}$$

where $\mathbf{g}'_{ij}$ and $\boldsymbol{\pi}'_{ij}$ represent row i of $\mathbf{G}_j$ and $\boldsymbol{\Pi}_j$, respectively with $\boldsymbol{\Pi}$ the $n \times \sum_{j=1}^{J} K_j$ matrix of probabilities $\boldsymbol{\Pi}$ having elements π_{ijk}. Now, the following theorem coming from unfinished notes of De Leeuw [2005] presents a quadratic majorizing function of (4).

Theorem 1.

$$g_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^{(0)}) = f_{ij}(\boldsymbol{\theta}_i^{(0)}) + (\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{(0)})'\nabla f_{ij}(\boldsymbol{\theta}_i^{(0)}) + 1/4\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{(0)}\|^2$$

is a majorizing function of $f_{ij}(\boldsymbol{\theta}_i)$.

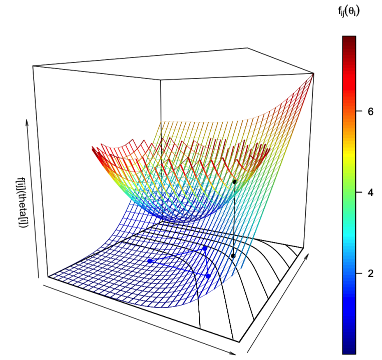


Figure 2: Quadratic majorizing function $g_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^{(0)})$ that touches $f_{ij}(\boldsymbol{\theta}_i)$ at $f_{ij}(\boldsymbol{\theta}_i)$ at $\boldsymbol{\theta}_i^{(0)}$ (black vertical line) and is located above $f_{ij}(\boldsymbol{\theta}_i)$ elsewhere.

In the appendix, this thereom is proved in a slightly different way compared to De Leeuw [2005]. Figure 2 gives an example of the quadratic majorizing function.

Second step: combining majorization.
To obtain a majorizing function for the deviance $L(\boldsymbol{\mu}, \mathbf{U}, \mathbf{D}, \mathbf{V})$, one needs to sum $f_{ij}(\boldsymbol{\theta}_i)$ and $g_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^{(0)})$ over all ijk and to replace the gradient by its value (5), it gives the majorizing function

$$L(\boldsymbol{\mu}, \mathbf{U}, \mathbf{D}, \mathbf{V}) \leq \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{k=1}^{K_j} g_{ijk}(z_{ijk} - \theta_{ijk})^2 + \lambda \left(\sum_{s=1}^{p} d_{ss} \right) + c \tag{6}$$

with

$$z_{ijk} = \mu_{jk}^{(0)} + \mathbf{u}_i^{(0)\prime} \mathbf{D}^{(0)} \mathbf{v}_{jk}^{(0)} + 2(g_{ijk} - \pi_{ijk})$$

and c containing parameters that do not depend on $\boldsymbol{\mu}, \mathbf{U}, \mathbf{D}$, and $\mathbf{V}$.

Third step: including missing values.
Here, we assume that missing values occur due to the most simple process, that is, missing completely at random. If a person i has a missing value on variable j, we can consider that $g_{ijk} = 0$ for $k = 1$ to K_j. In this way, the missing value does not contribute to the penalized deviance (3). When not all $\mathbf{g}_{ij}$ have a one, the majorizing function is a weighted least-squares function. With rank restrictions (as we have when $p \neq p^*$), such a weighted least-squares problem is more difficult to solve. One additional majorizing step is needed. Let $\mathbf{W}$ be an $n \times (\sum_{j=1}^{J} K_j)$ matrix that has $\mathbf{w}_{ij}' = \mathbf{1}'$ if person i does not have a missing value on variable j and $\mathbf{w}_{ij}' = \mathbf{0}'$ otherwise. Using results from Kiers [1997], we

have that

$$\begin{aligned}(w_{ijk}-1)(\theta_{ijk}-\theta_{ijk}^{(0)})^2 &\leq 0\\ w_{ijk}\theta_{ijk}^2 &\leq \theta_{ijk}^2-2\theta_{ijk}(1-w_{ijk})\theta_{ijk}^{(0)}+(1-w_{ijk})\left(\theta_{ijk}^{(0)}\right)^2 (7)\end{aligned}$$

Combining the majorization in (6) and (7) gives

$$\begin{aligned}L(\boldsymbol{\mu},\mathbf{U},\mathbf{D},\mathbf{V}) &\leq \frac{1}{4}\|\mathbf{Z}-(\mathbf{1}\boldsymbol{\mu}'+\mathbf{UDV}')\|^2+\lambda\left(\sum_{s=1}^{p} d_{ss}\right)+c\\ &\leq \frac{1}{4}\|\mathbf{Z}-\mathbf{1}\boldsymbol{\mu}'\|^2+\frac{1}{4}\|\mathbf{JZ}-\mathbf{UDV}'\|^2+\lambda\left(\sum_{s=1}^{p} d_{ss}\right)+c \qquad (8)\end{aligned}$$

with

$$\mathbf{Z}=[(\mathbf{1}\boldsymbol{\mu}^{(0)'}+\mathbf{U}^{(0)}\mathbf{D}^{(0)}\mathbf{V}^{(0)'})+2(\mathbf{G}-\mathbf{W}\odot\boldsymbol{\Pi})]\mathbf{J},$$

with $c=L(\boldsymbol{\mu}^{(0)},\mathbf{U}^{(0)},\mathbf{D}^{(0)},\mathbf{V}^{(0)})-1/4\|\mathbf{Z}\|^2$ and $\odot$ means the elementwise multiplication of two matrices.

Fourth step: update for the four sets of parameters.
The update for $\boldsymbol{\mu}$ simply amounts to minimizing $\|\mathbf{Z}-\mathbf{1}\boldsymbol{\mu}'\|^2$ which is done by

$$\boldsymbol{\mu}=n^{-1}\mathbf{Z}'\mathbf{1}.$$

To update $\mathbf{U}$ and $\mathbf{V}$ it is sufficient to minimize the crosspruduct term $-\text{tr}\mathbf{Z}'\mathbf{JUDV}'$ because the quadratic term in (8) disappears due to their orthonormality restrictions. Let $\mathbf{JZ}=\mathbf{P\Phi Q}'$ be the SVD. So-called Kristof lower bounds are available for linear sums of orthonormal matrices, that is,

$$\begin{aligned}-\text{tr}\mathbf{Z}'\mathbf{JUDV}' &= -\text{tr}\mathbf{Q\Phi P}'\mathbf{UDV}'=-\text{tr}\boldsymbol{\Phi}(\mathbf{P}'\mathbf{U})\mathbf{D}(\mathbf{V}'\mathbf{Q})\\ &\geq -\text{tr}\boldsymbol{\Phi}(\mathbf{I})\mathbf{D}(\mathbf{I})\end{aligned}$$

and the lower bound is attained at

$$\mathbf{U}=\mathbf{P} \text{ and } \mathbf{V}=\mathbf{Q}.$$

To update $\mathbf{D}$, we write the relevant part of the majorizing function (8) as

$$\sum_{s=1}^{p}\left[(\phi_{ss}-d_{ss})^2+\lambda\left(d_{ss}\right)\right] \qquad (9)$$

subject to $d_{ss}\geq 0$. It can be verified that the update

$$d_{ss}=\max(0,\phi_{ss}-\lambda) \qquad (10)$$

is optimal for to minimize (9).

A summary of the updates and the majorization algorithm for MMCA is given by Algorithm 1. After convergence, one may choose to set $\mathbf{X}=n^{1/2}\mathbf{UD}^{1/4}$ and $\mathbf{A}=n^{-1/2}\mathbf{VD}^{1/4}$.

```
Data: $\mathbf{G}, p, \lambda, \alpha, \epsilon$
Result: $\boldsymbol{\mu}, \mathbf{U}, \mathbf{D}, \mathbf{V}$
$t = 0$;
Compute $\mathbf{W}$ from missing values in $\mathbf{G}$;
Initialize $\boldsymbol{\mu}$: $\boldsymbol{\mu}^{(0)} = n^{-1}\mathbf{J}_c\mathbf{G}'\mathbf{1}$ ;
Compute the SVD of $\mathbf{JGJ}_c$: $\mathbf{JGJ}_c = \mathbf{P\Phi Q}'$;
Initialize $\mathbf{U}$: $\mathbf{U}^{(0)} = \mathbf{P}$ ;
Initialize $\mathbf{V}$: $\mathbf{V}^{(0)} = \mathbf{Q}$ ;
Initialize $\mathbf{D}$: $d_{ss}^{(0)} = \max(0, \phi_{ss} - \lambda)$ ;
Compute $\mathbf{\Pi}$ by (4) ;
Compute $L^{(0)} = L(\boldsymbol{\mu}^{(0)}, \mathbf{U}^{(0)}, \mathbf{D}^{(0)}, \mathbf{V}^{(0)})$ ;
while $t = 0$ or $(L^{(t)} - L^{(t-1)})/L^{(t)} \geq \epsilon$ do
    $t = t + 1$;
    $\mathbf{Z} = [(\mathbf{1}\boldsymbol{\mu}^{(t-1)'} + \mathbf{U}^{(t-1)}\mathbf{D}^{(t-1)}\mathbf{V}^{(t-1)'}) + 2(\mathbf{G} - \mathbf{W} \odot \mathbf{\Pi})]\mathbf{J}_c$ ;
    Compute update $\boldsymbol{\mu}$: $\boldsymbol{\mu}^{(t)} = n^{-1}\mathbf{Z}'\mathbf{1}$ ;
    Compute the SVD of $\mathbf{JZ}$: $\mathbf{JZ} = \mathbf{P\Phi Q}'$;
    Update $\mathbf{U}$: $\mathbf{U}^{(t)} = \mathbf{P}$ ;
    Update $\mathbf{V}$: $\mathbf{V}^{(t)} = \mathbf{Q}$ ;
    Update $\mathbf{D}$: $d_{ss} = (1 + \max(0, \phi_{ss} - \lambda)$ ;
    Compute $\mathbf{\Pi}$ by (4) ;
    Compute $L^{(t)} = L(\boldsymbol{\mu}^{(t)}, \mathbf{U}^{(t)}, \mathbf{D}^{(t)}, \mathbf{V}^{(t)})$ ;
end
```

Algorithm 1: The majorizing algorithm for MMCA. ϵ is here a small positive value, for example, $\epsilon = 10^{-8}$.

4 Properties and Interpretation

At a stationary point of the algorithm (in practical cases a local minimum) the solution satisfies several properties.

Property 1. *The main effect μ_{jk} can be interpreted as the log-odds against zero (that is, all categories are equally likely).*

Proof. Assuming that all other parameters are equal zero, then the probability of person i choosing category k of variable j equals $\pi_{ijk} = \exp(\mu_{jk})/\sum_{\ell=1}^{K_j}\exp(\mu_{j\ell})$. The probability for $\mu_{jk} = 0$ equals $\pi_{ijk}^{*} = \exp(0)/\sum_{\ell=1}^{K_j}\exp(\mu_{j\ell})$. The odds are

$$\frac{\pi_{ijk}}{\pi_{ijk}^{*}} = \frac{\left(\frac{\exp(\mu_{jk})}{\sum_{\ell=1}^{K_j}\exp(\mu_{j\ell})}\right)}{\left(\frac{\exp(0)}{\sum_{\ell=1}^{K_j}\exp(\mu_{j\ell})}\right)} = \exp(\mu_{jk})$$

and the log odds equals

$$\log \frac{\pi_{ijk}}{\pi^*_{ijk}} = \mu_{jk}.$$

$\square$

Let $\mathbf{X} = n^{1/2}\mathbf{U}$ and $\mathbf{A} = n^{-1/2}\mathbf{VD}$. In a bilinear biplot, the projection interpretation of $\mathbf{x}_i'\mathbf{a}_{jk}$ implies that a point $\mathbf{x}_i$ representing individual i should be projected onto the vector $\mathbf{a}_{jk}$ representing category k of variable j and multiplied by the length $\|\mathbf{a}_{jk}\|$ of $\mathbf{a}_{jk}$.

Property 2. *The interaction effect $\mathbf{x}_i'\mathbf{a}_{jk}$ can be interpreted as the log-odds against zero that would be obtained by projecting a person onto $\mathbf{a}_{jk}$ at the origin.*

Proof. Assume all parameters equal to zero except $\mathbf{x}_i'\mathbf{a}_{jk}$. Following the same reasoning as in the proof of Property 1 gives the desired result. $\square$

L can be rewritten as squared Euclidean distance between $\mathbf{x}_i$ and $\mathbf{a}_{jk}$.

Property 3. *The fitted category probabilities sum to the observed frequency of that category, that is, $\mathbf{1}'\mathbf{\Pi} = \mathbf{1}'\mathbf{G}$.*

Proof. At convergence we must have that $\boldsymbol{\mu} = \boldsymbol{\mu}^{(0)}$. Therefore,

$$\begin{aligned}\boldsymbol{\mu} &= n^{-1}\mathbf{Z}'\mathbf{1} \\ &= n^{-1}\left(\boldsymbol{\mu}\mathbf{1}' + 2(\mathbf{G} - \mathbf{\Pi})'\right)\mathbf{1} \\ &= \boldsymbol{\mu} + 2n^{-1}(\mathbf{G} - \mathbf{\Pi})'\mathbf{1}\end{aligned}$$

which can only hold if $(\mathbf{G} - \mathbf{\Pi})'\mathbf{1} = \mathbf{0}$, or, equivalently, if $\mathbf{G}'\mathbf{1} = \mathbf{\Pi}'\mathbf{1}$. This completes the proof. $\square$

The weighted centroids of $\mathbf{X}$ with the weights being the probability of choosing the category is given by

$$\text{Diag}(\mathbf{1}'\mathbf{G})^{-1}\mathbf{\Pi}'\mathbf{X}$$

where $\text{Diag}(\mathbf{1}'\mathbf{G})$ is the diagonal matrix of observed counts for each of the categories.

Property 4. $\mathbf{B} = K\mathit{Diag}(\mathbf{1}'\mathbf{G})^{-1}\mathbf{A}$ *is a measure for the bias of the category centroids, that is, the difference between the weighted and unweighted centroids, that is,*

$$\mathbf{B} = K\mathit{Diag}(\mathbf{1}'\mathbf{G})^{-1}\mathbf{A} = \mathit{Diag}(\mathbf{1}'\mathbf{G})^{-1}(\mathbf{G} - \mathbf{\Pi})'\mathbf{X}.$$

Proof. After convergence, it must also be true that the update of a single set of parameters yields the same solution. Consider the update of $\mathbf{A}$ for a given $\mathbf{X}$. The majorizing algorithm says that at convergence we have ... $\square$

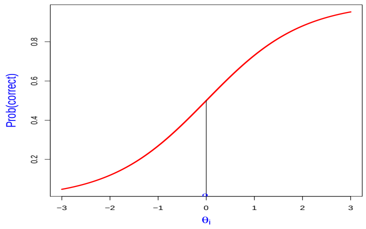


Figure 3: Example of the item characteristic curve used by the three parameter logistic IRT model.

Property 5. *The three parameter logistic model and its multidimensional variant are special cases of MMCA.*

Proof. In the three parameter logistic IRT model, the probability that individual i gives a correct answer to item j is equal to $(1+e^{-\gamma_{ij}})^{-1}$ with $\gamma_{ij} = \alpha_j\theta_i - \beta_j$ and θ_i is the ability of individual i, β_j the item difficulty parameter of item j, and α_j the item discrimination parameter of item j. Figure 3 gives the item characteristic curve by the three parameter logistic model.

The multidimensional IRT model can be written as

$$\begin{aligned}\gamma_{ij} &= -\beta_j + \sum_{s=1}^{p} \theta_{is}\alpha_{js} = -\beta_j + \boldsymbol{\theta}_i'\boldsymbol{\alpha}_j \\ &= \mu_j + \boldsymbol{x}_i'\boldsymbol{a}_j.\end{aligned}$$

If all variables are binary (all $K_j = 2$) and $\lambda = 0$, then the joint maximum likelihood approach to the multidimensional IRT model is equivalent to MMCA. □

The overfitting problem in multidimensional IRT can be easily tackled by the penalty term of the MMCA approach.

5 Selecting the Regularization Parameters

Selecting the threshold parameter λ is crucial in the method. We suggest a two steps procedure. First, we present an appropriate way to estimate the rank p of the interaction. Then, we select λ using cross-validation for a specified rank. To understand the rationale of such an approach, let's start by reviewing some key concepts and recent results to select the threshold parameter in the framework of ℓ_1 based penalty.

Giacobino et al. [2016] highlighted that lasso regression is very often used for its screening properties Bühlmann and van de Geer [2011] and its ability to select variables since it thresholds the coefficients estimate towards 0. However, current methods for selecting λ such as cross-validation or Stein unbiased risk estimation Zou et al. [2007], Tibshirani and Taylor [2012] focus on good predictive performances. They pointed out that the optimal threshold for prediction is typically different from the optimal one for screening and somewhat smaller which generally leads to too complex models. Thus, Giacobino et al. [2016] suggested a *quantile universal threshold* (QUT) that guarantee variable screening with high probability. However, their threshold is not appropriate for prediction since it biases too much the estimates. The rationale of the QUT approach is to selecting the threshold at the bulk edge of what a threshold should be under the null model that all the variables coefficients are equal to zero. Josse and Sardy [2015] used a similar idea of a null model in the context of low rank matrix estimation to estimate the support, *i.e.* the rank.

We can built on both works to suggest a strategy to select λ aiming at good rank recovery. The estimated rank is determined with the threshold λ since any empirical singular value d_{ss} smaller than λ is set to zero by (10). The procedure work as follows: for a given data set $\mathbf{G}$, we estimate the main effect $\boldsymbol{\mu}$ and generate data under the MMCA model (2) and the null hypothesis of no interaction $\mathbf{XA}' = 0$; then we apply the whole procedure (Algorithm 1) and return the first singular value $d_{1,1}$. This procedure is repeated 10000 times. Then, we use the $(1-\alpha)$-quantile of the distribution of the largest empirical singular value under the null hypothesis to determine the selected threshold. Following Donoho and Johnstone [1994] and Giacobino et al. [2016] who implicitly used level of order $\alpha = O(1/\sqrt{\log n})$ when $n = J$, we choose a similar level tending to zero with the maximum of n and $(K-J)$. This leads to the definition of the *quantile universal threshold* for estimating the rank in MMCA:

$$\lambda_{\max(n,K-J)} = \sigma F_{\Lambda_1}^{-1}\left(1 - \frac{1}{\sqrt{\log(\max(n, K-J))}}\right), \tag{11}$$

where F_{Λ_1} is the cumulative distribution function of the largest singular value.

Although this λ enjoy very good property of rank recovery, its value is far too large and it shrinks too much the sequel singular values. That's why, we only use this procedure to select the rank and then for a given rank p, we select λ to minimize the penalized deviance 3 using cross-validation. Note that it is in agreement with the common practice in lasso, where often lasso is used to select variables and then ordinary least squares are applied. Here we claim that in our setting we still need to shrink after the selection step.

Leave-one-out cross-validation, first consists in removing one cell of the categorical data matrix for one individual i on a variable j leading to a row $\mathbf{g}_{ij}$ with missing values. Then, it consists in predicting its value using the estimator obtained from the dataset that excludes these. The value of the predicted elements is denoted π_{ij}^{-ij}. Finally, the deviance is computed with these predicted values. The operation is repeated for all the cells in the categorical data and for a grid for λ. The value of λ that minimizes the deviance is selected. Such method is of course computationally intensive and so we have implemented a K-fold strategy and a parallelized version using the different cores of the machine.

6 Conclusion and Discussion

We introduced the *multinomial multiple correspondence analysis* to model the dependence between categorical variables using a low-rank representation of the data. The challenges in the estimation of the parameters were adressed with a majorization algorithm combined with a nuclear norm penalization. The universal threshold allows to accurately estimate the rank while cross-validation ensures good selection of the thresholding parameter. There appear to be many potential applications for our methods since it is possible to scale majorization algorithm to the difficult cases of large sparse data sets. One drawback is that contrary to MCA, the biplot representation does not enjoy the centroid property. It means that categories are not at the barycenter of the individuals which have selected the categories. Thus, biplot may be less easy to interpret. Note that in terms of graphical outputs MCA enjoys many nice interpretation but as already mentionned, the estimated values for the probabilities sum to one but can take negative values and thus does not represent a proper way to model the probabilities of individuals taken categories.

We finish by discussing some opportunities for further research. We used a two-step approach to select the regularization parameter, after the hard-thresholding step we still shrink with a soft-thresholding approach. This is in the same vein than selecting variables with LASSO and using ordinary least-squares except that we use in the second step an additional regularization. This may indicate that we should consider other penalties and scheme of regularization allowing compromise between hard and soft thresholding. In the framework of low rank matrix approximation with Gaussian noise, recent works showed that the signal was better recovered when non-linear transformation of the singular values were applied. For instance Shabalin and Nobel [2013] and Gavish and Donoho [2014] gave an explicit transformation in a asymptotic framework where both n and J tend to infinity while Josse and Sardy [2015] considered a finite sample situation and suggested an adaptive penalty inspired by adaptive LASSO. Their method provides a large family of thresholding function that goes between hard and soft thresholding. Extending these ideas for categorical data provide some challenges both to get results outside the Gaussian case or to include easily different penalties in the majorization algorithm.

The cross-validation procedure and the capability of the method to handle missing values encourage investigating the use of this method to handle missing values in a broader framework. Indeed, one main strategy avalaible to deal with missing values Little and Rubin [1987, 2002] consists in using imputation methods, it means replacing the missing values by plausible values to get a completed data on which any statistical analysis can be applied. More precisely, the recomemend strategy is to use multiple imputation Rubin [1987] where multiple values are predicted for each missing entrie to take into account the uncertainty of prediction in the sequel analyses. Many multiple imputation techniques are available for continuous data Van Buuren [2012] but the litterature is less abundant for categorical ones. It can be explained by the difficulty to get an imputation model and an estimation strategy which can handle large number of categories per variable, a large number of variables or a small number of individuals. That's why we may expect some interesting results in this direction.

A Proofs

Theorem 1. To prove that $g_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^{(0)}) == f_{ij}(\boldsymbol{\theta}_i^{(0)}) + (\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{(0)})'\nabla f_{ij}(\boldsymbol{\theta}_i^{(0)}) + 1/4\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{(0)}\|^2$ is a majorizing function of $f_{ij}(\boldsymbol{\theta}_i)$ two conditions must hold. The first one is that $f_{ij}(\boldsymbol{\theta}_i) = g_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^{(0)})$ at the supporting point $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^{(0)}$. Working out $g_{ij}(\boldsymbol{\theta}_i^{(0)}, \boldsymbol{\theta}_i^{(0)})$ trivially shows that it is equal to $f_{ij}(\boldsymbol{\theta}_i^{(0)})$ thereby confirming the first assumption.

The second requirement is that $f_{ij}(\boldsymbol{\theta}_i) \leq g_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^{(0)})$ or, equivalently, $h_{ij}(\boldsymbol{\theta}_i) = g_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^{(0)}) - f_{ij}(\boldsymbol{\theta}_i) \geq 0$ for all $\boldsymbol{\theta}_i$ with equality for $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^{(0)}$. Equality was already proven above. The inequality is automatic if (i) the gradient of $h_{ij}(\boldsymbol{\theta}_i)$ is zero at the supporting point $\boldsymbol{\theta}_i^{(0)}$ and (ii) $h_{ij}(\boldsymbol{\theta}_i)$ is convex. The gradient of $h_{ij}(\boldsymbol{\theta}_i)$ is given by

$$\begin{aligned}\nabla h_{ij}(\boldsymbol{\theta}_i) &= \nabla g_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^{(0)}) - \nabla h_{ij}(\boldsymbol{\theta}_i) \\ &= \nabla f_{ij}(\boldsymbol{\theta}_i^{(0)}) + 1/2(\boldsymbol{\theta}_i - \boldsymbol{\theta}_i^{(0)}) - \nabla f_{ij}(\boldsymbol{\theta}_i)\end{aligned}$$

so that

$$\nabla h_{ij}(\boldsymbol{\theta}_i^{(0)}) = \nabla f_{ij}(\boldsymbol{\theta}_i^{(0)}) + 1/2(\boldsymbol{\theta}_i^{(0)} - \boldsymbol{\theta}_i^{(0)}) - \nabla f_{ij}(\boldsymbol{\theta}_i^{(0)}) = \mathbf{0}.$$

and Condition (i) is satisfied. The Hessian of $f_{ij}(\boldsymbol{\theta}_i)$ is given by

$$\nabla^2 f_{ij}(\boldsymbol{\theta}_i) = \mathrm{Diag}(\boldsymbol{\pi}_{ij}) - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{ij}'$$

and that of

$$\nabla^2 g_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^{(0)}) = 1/2\mathbf{I}$$

so that the Hessian of $h_{ij}(\boldsymbol{\theta}_i)$

$$\nabla^2 h_{ij}(\boldsymbol{\theta}_i) = \nabla^2 g_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^{(0)}) - \nabla^2 f_{ij}(\boldsymbol{\theta}_i) = 1/2\mathbf{I} - (\mathrm{Diag}(\boldsymbol{\pi}_{ij}) - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{ij}').$$

For Condition (ii), convexity of $h_{ij}(\boldsymbol{\theta}_i)$, to hold it suffices to prove that $\nabla^2 h_{ij}(\boldsymbol{\theta}_i)$ is positive semidefinite for all $\boldsymbol{\theta}_i$, or, equivalently, that all eigenvalues of $\mathrm{Diag}(\boldsymbol{\pi}_{ij}) - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{ij}'$ are smaller that 1/2. An upper bound of the eigenvalues can be obtained by Gerschgorin disks which say that the eigenvalue ϕ is always smaller than a diagonal element plus the sum of its absolute off-diagonal row (or column) values, i.e.,

$$\begin{aligned}\phi &\leq \pi_{ijk} - \pi_{ijk}^2 + \pi_{ijk}\sum_{\ell \neq k}\pi_{ij\ell} && (12) \\ &= \pi_{ijk} - \pi_{ijk}^2 + \pi_{ijk}\sum_{\ell=1}^{K_j}\pi_{ij\ell} - \pi_{ijk}^2 && (13) \\ &= 2(\pi_{ijk} - \pi_{ijk}^2) = 2\pi_{ijk}(1 - \pi_{ijk}). && (14)\end{aligned}$$

It can be verified that $2\pi_{ijk}(1-\pi_{ijk})$ reaches its maximum of $1/2$ at $\pi_{ijk} = 1/2$ so that the maximum eigenvalue of $\nabla^2 f_{ij}(\boldsymbol{\theta}_i)$ is always smaller than (or equal to) $\phi = 1/2$ and, thus, $\nabla^2 h_{ij}(\boldsymbol{\theta}_i)$ is positive semidefinite and $h_{ij}(\boldsymbol{\theta}_i)$ is convex. □

References

A. Agresti. *Categorical Data Analysis, 3rd Edition.* Wiley, 2013.

J. P. Benzécri. *L'analyse des données. Tome II: L'analyse des correspondances.* Dunod, 1973.

A. Bhattacharya and D. B. Dunson. Simplex factor model for multivariate unordered categorical data. *Journal of the American Statistical Association*, 107 (497):362–377, 2012.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data.* Springer-Verlag, 2011.

Wray Buntine. Variational extensions to EM and multinomial PCA. In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, editors, *Machine Learning: ECML 2002*, volume 2430 of *Lecture Notes in Computer Science*, pages 23–34. Springer Berlin Heidelberg, 2002.

E. J. Candes, C. A. Sing-Long, and J. D. Trzasko. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, 61(19):4643–4657, 2013.

R. Christensen. *Log-Linear Models.* Springer-Verlag, New York, 2010.

Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. A generalization of principal component analysis to the exponential family. In *Advances in Neural Information Processing Systems.* MIT Press, 2001.

J. De Leeuw. Gifi goes logistic. Technical report, Department of Statistics, University of California, Los Angeles, 2005. URL `http://gifi.stat.ucla.edu/janspubs`.

J. De Leeuw. History of non linear principal component analysis. In J. Blasius and M. J. Greenacre, editors, *Visualization and Verbalization of Data.* Chapman & Hall, 2014.

J. De Leeuw and W. J. Heiser. Convergence of correction matrix algorithms for multidimensional scaling. In J. C. Lingoes, E. Roskam, and I. Borg, editors, *Geometric Representations of Relational Data*, pages 735–752. Mathesis Press, 1977.

Jan De Leeuw. Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis*, 50(1):21–39, 2006.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

D. B. Dunson and C. Xing. Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104:1042–1051, 2009.

M. Fazel. *Matrix Rank Minimization with Applications.* PhD thesis, Stanford University, 2002.

M. Gavish and D. L. Donoho. Optimal shrinkage of singular values. *arXiv:1405.7511v2*, 2014.

C. Giacobino, S. Sardy, and N. Hengartner. Quantile universal threshold selection with an application in generalized model with lasso. Technical report, arXiv, Cornell, 2016.

L. A. Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:255–231, 1974.

M. J. Greenacre. *Theory and Applications of Correspondence Analysis.* Acadamic Press, 1984.

M. J. Greenacre and J. Blasius. *Multiple Correspondence Analysis and Related Methods.* Chapman & Hall/CRC, 2006.

J. Josse and S. Sardy. Adaptive shrinkage of singular values. *Statistics and Computing*, pages 1–10, 2015.

H. A. L. Kiers. Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, 62(2):251–266, 1997.

K. Lange. *Optimization.* Springer-Verlag, New York, 2004.

B. le Roux. *Multiple Correspondence Analysis.* SAGE publications, CA: Thousand Oaks, 2010.

J. Li and D. Tao. Simple exponential family PCA. *Neural Networks and Learning Systems, IEEE Transactions on*, 24(3):485–497, 2013.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* Wiley series in probability and statistics, New-York, 1987, 2002.

G. Michailidis and J. De Leeuw. The Gifi system of descriptive multivariate analysis. *Statistical Science*, 13:307–336, 1998.

S Nishisato. *Analysis of Categorical Data: Dual Scaling and its Applications.* University of Toronto Press, Toronto, 1980.

D. B. Rubin. *Multiple Imputation for Non-Response in Survey.* Wiley, 1987.

Andrey A Shabalin and Andrew B Nobel. Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 118: 67–76, 2013.

N Srebro. *Learning with Matrix Factorizations.* PhD thesis, Massachusetts institute of technology, 2004.

M. Tenenhaus and F. W. Young. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50:91–119, 1985.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B: Methodological*, 58:267–288, 1996.

R. J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40 (2):1198–1232, 2012.

S. Van Buuren. *Flexible Imputation of Missing Data (Chapman & Hall/CRC Interdisciplinary Statistics).* Chapman and Hall/CRC, 2012.

H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35 (2):2173–2192, 2007.

Bibliography

Agresti A (2013). *Categorical Data Analysis, 3rd Edition.* Wiley.

Audigier V, Husson F, Josse J (2014a). "A principal components method to impute missing values for mixed data." *Advances in Data Analysis and Classification*, **10**(1), 5–26.

Audigier V, Husson F, Josse J (2014b). "Multiple Imputation for Continuous Variables Using a Bayesian Principal Component Analysis." *Journal of Statistical Computation and Simulation.*

Audigier V, Husson F, Josse J (2015). "MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis." *Statistics and Computing.*

Baldi P, Hornik K (1989). "Neural networks and principal component analysis: Learning from examples without local minima." *Neural Networks*, **2**(1), 53–58.

Bartholomew D (1987). *Latent Variable Models and Factor Analysis.* Charles Griffin and Co Ltd.

Bates D, Watt D (1980). "Relative curvature measure of nonlinearity." *Journal of the Royal Society series B*, **42**, 1–25.

Benzécri JP (1973). *L'analyse des données. Tome II: L'analyse des correspondances.* Dunod.

Besag J (1974). "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**(2).

Besse P, de Falguerolles A (1993). "Application of Resampling Methods to the Choice of Dimension in Principal Component Analysis." In *Computer Intensive Methods in Statistics*, pp. 167–176. Physica-Verlag.

Bishop CM (1995). "Training with noise is equivalent to Tikhonov regularization." *Neural Computation*, **7**(1), 108–116.

Bourlard H, Kamp Y (1988). "Auto-association by multilayer perceptrons and singular value decomposition." *Biological Cybernetics*, **59**(4-5), 291–294.

Cai JF, Candès EJ, Shen Z (2010). "A singular value thresholding algorithm for matrix completion." *SIAM Journal on Optimization*, **20**(4), 1956–1982.

Candès EJ, Sing-Long CA, Trzasko JD (2013). "Unbiased risk estimates for singular value thresholding and spectral estimators." *IEEE Transactions on Signal Processing*, **61**(19), 4643–4657.

Candes EJ, Sing-Long CA, Trzasko JD (2013). "Unbiased risk estimates for singular value thresholding and spectral estimators." *IEEE Transactions on Signal Processing*, **61**(19), 4643–4657.

Candès EJ, Tao T (2009). "The Power of Convex Relaxation: Near-Optimal Matrix Completion." *IEEE Trans. Inform. Theory*, **56(5)**, 2053–2080.

Carpenter J, Kenward M (2013). *Multiple Imputation and its Application.* John Wiley & Sons.

Caussinus H (1986). "Models and Uses of Principal Component Analysis (with Discussion)." In J de Leeuw, W Heiser, J Meulman, F Critchley (eds.), *Multidimensional Data Analysis*, pp. 149–178. DSWO Press.

Chikuse Y (2003). *Statistics on Special Manifolds.* Springer, New York.

Christensen R (2010). *Log-Linear Models.* Springer-Verlag, New York.

Christiansen C, Morris C (1997). "Hierarchical Poisson Regression Model." *Journal of the American Statistical Association*, **92**, 618–632.

Cornelius P, Crossa J, Seyedsadr M (1996). "Statistical tests and estimators of multiplicative models for genotype by environment interaction." In *Genotype by environment interaction*, pp. 199–234. M.S Kang and H.G Gauch (eds), CRC press, Boca Raton, FL.

Craven P, Wahba G (1979a). "Smoothing noisy data with spline functions." *Numer. Math.*, **31**(4), 377–403.

Craven P, Wahba G (1979b). "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation." *Numerische Mathematik*, **31**, 377–403.

Daudin JJ, Duby C, Trecourt P (1989). "PCA stability studied by the bootstrap and the infinitesimal jackknife method." *Statistics: A journal of theoretical and applied statistics*, **20**(2), 255–270.

de Falguerolles A (1998). "Log-bilinear biplot in action." In J Blasius, MJ Greenacre (eds.), *Visualisation Of Categorical Data*, pp. 527–533. Academic Press.

De Leeuw J (2005). "Gifi goes logistic." *Technical report*, Department of Statistics, University of California, Los Angeles. URL `http://gifi.stat.ucla.edu/janspubs`.

de Leeuw J, Rijckevorsel JV (1980). "HOMALS and PRINCALS - Some generalizations of principal components analysis." In *Data analysis and informatics*, pp. 231–242. Springer.

de Tayrac M, Lê S, Aubry M, Mosser J, Husson F (2009). "Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach." *BMC Genomics*, **10**(1), 32.

Dempster AP, Laird NM, Rubin DB (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society B*, **39**(1), 1–38.

Denis JB, Gower JC (1994). "Asymptotic covariances for the parameters of biadditive models." *Utilitas Mathematica*, pp. 193–205.

Denis JB, Gower JC (1996). "Asymptotic confidence regions for biadditive models: interpreting genotype-environment interactions." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **45**, 479–493.

Denis JB, Pazman A (1999). "Bias of least squares estimators in nonlinear regression models with constraints. Part II: Biadditive Models." *Applications of Mathematics*, **44**, 359–374.

Dray S, Josse J (2014). "Principal component analysis with missing values: a comparative survey of methods." *Plant Ecology*, **216**(5), 657–667.

Eckart C, Young G (1936). "The approximation of one matrix by another of lower rank." *Psychometrika*, **1**(3), 211–218.

Efron B, Morris C (1972). "Limiting the risk of Bayes and empirical Bayes estimators. Part II: The empirical Bayes case." *Journal of the American Statistical Association*, **67**(19), 130–139.

Efron B, Stein C (1981). "The Jackknife Estimate of Variance." *Annals of Statistics*, **3**, 586–596.

Efron B, Tibshirani R (1994). *An Introduction to the Bootstrap.* Chapman & Hall/CRC.

Escofier B (1979). "Traitement Simultané de Variables Quantitatives et Qualitatives en Analyse Factorielle." *Les cahiers de l'analyse des données*, **4**(2), 137–146.

Escofier B, Pagès J (2008). *Analyses Factorielles Simples et Multiples.* Dunod.

Gauch H (1988). "Model selection and validation for yield trials with interaction." *Biometrics*, **44**, 705–715.

Gauch H, Zobel R (1996). "AMMI analysis of yield trials." In *Genotype by environment interaction*, pp. 141–150. M.S Kang and H.G Gauch (eds), CRC press, Boca Raton, FL.

Gavish M, Donoho DL (2014a). "Optimal Shrinkage of Singular Values." *arXiv:1405.7511v2.*

Gavish M, Donoho DL (2014b). "Optimal Shrinkage of Singular Values." *arXiv:1405.7511v2.*

Gelman A, Hill J (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

Giacobino C, Sardy S, Diaz Rodriguez J, Hengartner N (2015). "Quantile universal threshold for model selection." *arXiv:1511.05433.*

Gifi A (1990a). *Non-linear Multivariate Analysis.* Wiley, Chichester, England.

Gifi A (1990b). *Nonlinear Multivariate Analysis.* John Wiley & Sons, Chichester.

Gower J, Lubbe S, Le Roux N (2011). *Understanding Biplots.* John Wiley & Sons.

Gower JC, Dijksterhuis GB (2004a). *Procrustes Problems.* Oxford University Press, USA.

Gower JC, Dijksterhuis GB (2004b). *Procrustes Problems.* New York: Oxford University Press.

Greenacre M (1984a). *Theory and Applications of Correspondence Analysis.* Acadamic Press.

Greenacre M (2009). *Biplots in Practice.* Fundacion BBVA (FBBVA).

Greenacre M, Blasius J (2006). *Multiple Correspondence Analysis and Related Methods.* Chapman & Hall/CRC.

Greenacre MJ (1984b). *Theory and Applications of Correspondence Analysis.* Acadamic Press.

Greenacre MJ (2007). *Correspondence Analysis in Practice, Second Edition.* Chapman & Hall.

Groenen PJF, Josse J (2016). "Multinomial MCA." *http://arxiv.org/abs/1603.03174.*

Hoff PD (2009). "Simulation of the Matrix Bingham-von Mises–Fisher Distribution, With Applications to Multivariate and Relational Data." *Journal of Computational and Graphical Statistics*, **18**(2), 438–456.

Honaker J, King G, Blackwell M (2011). "Amelia II: A Program for Missing Data." *Journal of Statistical Software*, **45**(7), 1–47.

Honaker J, King G, Blackwell M (2014). *Amelia II: A Program for Missing Data.* R package version 1.7.2.

Huet S, Denis JB, Adamczyk K (1999). "Bootstrap confidence intervals in nonlinear regression models when the number of observations is fixed and the variance tends to 0. Application to biadditive model." *Statistics*, **32**, 203–227.

Husson F, Josse J (2013). "Handling Missing Values in Multiple Factor Analysis." *Food Quality and Preferences*, **30**(2), 77–85.

Husson F, Josse J (2014a). "Multiple Correspondence Analysis." In M Greenacre, J Blasius (eds.), *Visualization and Verbalization of Data*, pp. 163–181. Chapman & Hall/CRC, London.

Husson F, Josse J (2014b). "Multiple Correspondence Analysis." In *Visualization and Verbalization of Data*, pp. 165–184. CRC Press, Taylor & Francis.

Husson F, Josse J (2015). *missMDA: Handling Missing Values with Multivariate Data Analysis.* R package version 1.9, URL `http://CRAN.R-project.org/package=missMDA`.

Husson F, Josse J, Saporta G (2016). "Jan de Leeuw and the French school of data analysis." *Journal of Statistical Software.*

Husson F, Le S, Pagès J (2010). *Exploratory Multivariate Analysis by Example Using R.* Chapman & Hall/CRC.

Ivanoff S, Picard F, Rivoirard V (2015). "Adaptive Lasso and group-Lasso for functional Poisson regression." *The Journal of Machine Learning Research.*

Josse J, Chavent M, Liquet B, Husson F (2012). "Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis." *Journal of classification*, **29**(1), 91–116.

Josse J, Holmes S (2015). "Test of independence and beyond." *arXiv preprint arXiv:1307.7383.*

Josse J, Husson F (2011a). "Multiple Imputation in PCA." *Advances in data analysis and classification*, **5**(3), 231–246.

Josse J, Husson F (2011b). "Selecting the Number of Components in PCA Using Cross-Validation Approximations." *Computational Statististics and Data Analysis*, **56**(6), 1869–1879.

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.

Josse J, Husson F (2015). "missMDA a package to handle missing values in and with multivariate data analysis methods." *Journal of Statistical Software.*

Josse J, Husson F, Wager S (2014a). "Confidence areas for fixed-effects PCA." *Journal of Computational and Graphical Statistics.*

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, Sardy S (2015). "Adaptive shrinkage of singular values." *Statistics and Computing*, pp. 1–10.

Josse J, Sardy S, Wager S (2016). "denoiseR: a package for low rank matrix estimation." *http://arxiv.org/abs/1602.01206.*

Josse J, Timmerman ME, Kiers HAL, Smilde AK (2013). "Missing values in multi-level simultaneous component analysis." *Chemometrics and Intelligent Laboratory Systems*, **129**(2), 21–32.

Josse J, van Eeuwijk F, Piepho HP, Denis JB (2014b). "Another Look at Bayesian Analysis of AMMI Models for Genotype-Environment Data." *Journal of Agricultural, Biological, and Environmental Statistics*, **19**(2), 240–257.

Josse J, Wager S (2014). "Stable Autoencoding: a flexible framework for regularized low-rank matrix estimation." *arXiv:1410.8275.*

Josse J, Wager S (2015). "Stable Autoencoding: A Flexible Framework for Regularized Low-Rank Matrix Estimation." *arXiv:1410.8275v2.*

Josse J, Wager S, Sardy S (2015). *denoiseR: Regularized low-rank matrix estimation.* R package version 1.0.

Khatri CG, Mardia KV (1977). "The Von Mises-Fisher Matrix Distribution in Orientation Statistics." *Journal of the Royal Statistical Society. Series B*, **39 (1)**, 95–106.

Kiers HAL (1991). "Simple Structure in Component Analysis Techniques for Mixtures of Qualitative and Quantitative Variables." *Psychometrika*, **56**, 197–212.

Kiers HAL (1997). "Weighted Least Squares Fitting Using Ordinary Least Squares Algorithms." *Psychometrika*, **62**(2), 251–266.

Koren Y, Bell R, Volinsky C (2009). "Matrix factorization techniques for recommender systems." *Computer*, **42**(8), 30–37.

Kroonenberg PM (2008). *Applied Multiway Data Analysis (chap.7).* John Wiley & Sons series in probability and statistics.

Kropko J, Goodrich B, Gelman A, Hill J (2014). "Multiple Imputation for Continuous and Categorical Data: Comparing Joint and Conditional Approaches." *Political Analysis.*

Kropko J, Goodrich B, Gelman A, Hill J (2014). "Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches." *Political Analysis.*

Krzanowski WJ (2010). *Principles of multivariate analysis; a user's perspective.* Clarendon Press, Oxford.

Lazarsfeld PF, Henry NW (1968). *Latent Structure Analysis.* Boston: Houghton Mifflin.

Lê S, Josse J, Husson F (2008). "FactoMineR: An R Package for Multivariate Analysis." *Journal of Statistical Software*, **25**(1), 1–18.

Lebart L (2007). "Which Bootstrap for Principal Axes Methods?" In *Selected Contributions in Data Analysis and Classification*, pp. 581–588. Springer Berlin Heidelberg.

Lebart L, Saporta G (2014). "Historical Elements of Correspondence Analysis and Multiple Correspondence Analysis." In *Visualization and Verbalization of Data*, pp. 31–44. CRC Press, Taylor & Francis.

Leek JT, Storey JD (2007). "Capturing heterogeneity in gene expression studies by surrogate variable analysis." *PLoS genetics*, **3**(9), e161.

Little RJA, Rubin DB (1987, 2002). *Statistical Analysis with Missing Data.* John Wiley & Sons series in probability and statistics, New-York.

Lustig M, Donoho DL, Santos JM, Pauly JM (2008). "Compressed sensing MRI." *Signal Processing Magazine, IEEE*, **25**(2), 72–82.

Martyn P (2003). "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, Austria.*

Meng XL, Rubin DB (1991). "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm." *Journal of the American Statistical Association*, **86**(416), 899–909. http://links.jstor.org/sici?sici=0162-1459%28199112%2986%3A416%3C899%3AUETOAV%3E2.0.CO%3B2-H.

Michailidis G, de Leeuw J (1998). "The Gifi system of descriptive multivariate analysis." *Statistical Science*, **13**, 307–336.

Mosler K (2013). "Depth statistics." In *Robustness and Complex Data Struc- tures, Festschrift in Honour of Ursula Gather*, pp. 17–34. Springer, Berlin.

Netflix (2009). "Netflix Challenge." URL http://www.netflixprize.com.

Pagès J (2015). *Multiple Factor Analysis with R.* Chapman & Hall/CRC.

Papadopoulo T, Lourakis MIA (2000). "Estimating the jacobian of the singular value decomposition: Theory and applications." In *In Proceedings of the European Conference on Computer Vision, ECCV00*, pp. 554–570. Springer.

Pazman A (2002). "Results on nonlinear least squares estimators under nonlinear equality constraints." *Journal of statistical planning and inference*, **103**, 401–420.

Pazman A, Denis JB (2002). "Measures of nonlinearity for biadditive ANOVA models." *Metrika*, **55 (3)**, 233–245.

Perez-Elizalde S, Jarquin D, Crossa J (2011). "A general Bayesian estimation method of linear-bilinear models applied to plant breeding trials with genotype x environment interaction." *Journal of Agricultural Biological and Environmental Statistics*, **17(1)**, 15–37.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). "Principal components analysis corrects for stratification in genome-wide association studies." *Nature Genetics*, **38**(8), 904–909.

Robinson GK (1991). “That BLUP is a Good Thing: The Estimation of Random Effects.” *Statistical Science*, **6**(1), 15–32.

Royo C, Rodriguez A, Romagosa I (1993). “Differential Adaptation of Complete and Substitute Triticale.” *Plant Breeding*, **111**, 113–119.

Rubin DB (1976). “Inference and missing data.” *Biometrika*, **63**, 581–592.

Rubin DB (1987). *Multiple Imputation for Non-Response in Survey.* John Wiley & Sons.

Schafer JL (1997). *Analysis of incomplete multivariate data.* Chapman & Hall/CRC, London.

Schoenemann P (1966). “A generalized solution of the orthogonal Procrustes problem.” *Psychometrika*, **31**(1), 1–10.

Shabalin AA, Nobel AB (2013). “Reconstruction of a low-rank matrix in the presence of Gaussian noise.” *Journal of Multivariate Analysis*, **118**, 67–76.

Si Y, Reiter J (2013). “Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys.” *Journal of Educational and Behavioral Statistics*, **38**, 499–521.

Silvey S (1975). *Statistical Inference.* Chapman & Hall.

Smidl V, Quinn A (2007). “On Bayesian principal component analysis.” *Computational Statistics and Data Analysis*, **51**, 4101–4123.

Stein C (1981). “Estimation of the Mean of a Multivariate Normal Distribution.” *The Annals of Statistics*, **9**, 1135–1151.

Stekhoven D, Bühlmann P (2012). “MissForest - Nonparametric missing value imputation for mixed-type data.” *Bioinformatics*, **28**, 113–118.

Takane Y (2013). *Constrained Principal Component Analysis and Related Techniques.* Chapman & Hall.

Tanner MA, Wong WH (1987). “The calculation of posterior distributions by data augmentation.” *Journal of the American Statistical Association*, **82**, 805–811.

Templ M, Alfons A, Kowarik A, Prantner B (2013). *VIM: Visualization and Imputation of Missing Values.* R package version 4.0.0, URL `http://CRAN.R-project.org/package=VIM`.

Tibshirani R (1996a). “Regression shrinkage and selection via the lasso.” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Tibshirani R (1996b). “Regression shrinkage and selection via the Lasso.” *Journal of the Royal Statistical Society, Series B: Methodological*, **58**, 267–288.

van Buuren S (2007). “Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification.” *Statistical Methods in Medical Research*, **16**, 219–242.

van Buuren S (2012). *Flexible Imputation of Missing Data.* Chapman & Hall/CRC, Boca Raton.

Van Buuren S (2014). *mice.* R package version 2.18.

van Buuren S, Boshuizen H, Knook D (1999). "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis." *Statistics in Medicine*, **18**, 681–694.

van Buuren S, Groothuis-Oudshoorn K (2011). "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, **45**(3), 1–67.

Verbanck M, Josse J, Husson F (2013). "Regularized PCA to Denoise and Visualise Data." *Statistics and Computing*, pp. 1–16.

Vermunt JK, van Ginkel JR, van der Ark LA, Sijtsma K (2008). "Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis." *Sociological Methodology*, **33**, 369–397.

Vidotto D, Kapteijn MC, Vermunt J (2014). "Multiple imputation of missing categorical data using latent class models: State of art." *Psychological Test and Assessment Modeling.* . In press.

Viele K, Srinivasan C (2000). "Parsimonious estimation of multiplicative interaction in analysis of variance using Kullback-Leibler information." *Journal of Statistical Planning and Inference*, **84**, 201–219.

Wager S, Fithian W, Wang S, Liang P (2014). "Altitude Training: Strong Bounds for Single-Layer Dropout." In *Advances in Neural Information Processing Systems.*

Wold H, Lyttkens E (1969). "Nonlinear iterative partial least squares (NIPALS) estimation procedures." *Bulletin. Int. Stat. Institut*, **43**, 29–51.

Zhang CH, Huang J (2008). "The sparsity and biais of the lasso selection in high-dimensional linear regression." *The Annals of Statistics*, **36**(4), 1567–1594.

Zou H (2006a). "The Adaptive LASSO and Its Oracle Properties." *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou H (2006b). "The Adaptive LASSO and Its Oracle Properties." *Journal of the American Statistical Association*, **101**, 1418–1429.