Linear Dimensionality Reduction

Alain Franc
Linear Dimensionality Reduction

Alain Franc∗†

Project-Team Pleiade

Research Report n° 9488 — September 2022 — 65 pages

Abstract: These notes are an overview of some classical linear methods in Multivariate Data Analysis. This is an old good domain, well established since the 60’s, and refreshed timely as a key step in statistical learning. It can be presented as part of statistical learning, or as dimensionality reduction with a geometric flavor. Both approaches are tightly linked: it is easier to learn patterns from data in low dimensional spaces than in high-dimensional spaces. It is shown how a diversity of methods and tools boil down to a single core methods, PCA with SVD, such that the efforts to optimize codes for analyzing massive data sets can focus on this shared core method, and benefit to all methods. An extension to the study of several arrays is presented (Canonical Analysis).

Key-words: Dimensionality reduction, Multivariate Data Analysis, Statistical Learning, Principal Components Analysis, Correspondence Analysis, Analysis with Instrumental Variables, Canonical Analysis

∗ Pleiade team and INRAE, Biogeco, University of Bordeaux, 69, route d’Arcachon, 33610, Cestas
† Correspondence: alain.franc@inrae.fr
Méthodes linéaires de Réduction de Dimension

Résumé : Ce document brosse un panorama des méthodes linéaires de l’Analyse de données multivariées. Il s’agit d’un domaine ancien et classique, bien établi depuis les années 60, et redevenu d’actualité en tant qu’étape clé dans l’apprentissage statistique. On peut considérer ces méthodes comme faisant partie d’une approche algébrique de l’apprentissage statistique ou bien comme une réduction de dimension avec une tonalité plus géométrique. Ces deux approches sont étroitement liées : il est plus facile d’apprendre des patterns des données dans des espaces à faible dimension que dans des espaces à grande dimension. Nous montrons comment une apparente diversité de méthodes et outils se réduit en fait pour un tableau à une seule méthode : l’Analyse en Composantes Principales, avec la SVD (Singular Value Decomposition), de telle sorte que les efforts d’optimisation des codes pour l’analyse de jeux de données massives peut se focaliser sur cette méthode centrale partagée, au bénéfice de toutes les méthodes. Une extension à l’étude de plusieurs tableaux est présentée (Analyse canonique).

Mots-clés : Réduction de dimension, Analyse de données multivariées, Apprentissage statistique, Analyse en Composantes Principales, Analyse Factorielle des Correspondances, Analyse avec variables instrumentales, Analyse canonique
Préface

La rédaction de ce document a été motivée par la rencontre de plusieurs observations.

⋄ La diversité des méthodes qualifiées d’Analyse des Données dans les années 70, aujourd’hui rattachées à l’apprentissage statistique, peut s’organiser de façon circulaire où l’une est une déclinaison de l’autre via certains choix. Cette approche globale est essentiellement algébrique, à base de calcul matriciel, et a été très développée par une école française, en parallèle avec des approches plus statistiques dans les pays anglo-saxons.

⋄ Les méthodes principales sont l’ACP (Analyse en Composantes Principales), l’ACP centrée normée, l’Analyse Factorielle des Correspondances de tables de contingence, l’Analyse de deux tableaux par l’AnalyseCanonique, la MDS (Multidimensional Scaling, méthode de construction d’un nuage de points à partir d’un tableau de distances pour laquelle il n’y a pas de terme français consacré par l’usage). Ces méthodes étaient un peu tombées en désuétude comme descriptives, mais ont connu un regain d’intérêt pour l’exploration des structures dans les données massives (ce qui est appelé parfois "pattern discovery" et se rattache à l’apprentissage non supervisé).

⋄ Le point de départ de chacune de ces méthodes est un tableau de données, qui peut être un tableau croisé entre objets et variables, une table de contingence, une matrice de distances ou dissimilarités, etc. ... Une observation clé est que chacune des méthodes s’organise selon le triptyque donné prétraitement SVD post-traitement résultats où chaque méthode peut être vue comme un diabolo, où un prétraitement des données construit une matrice, elle même décomposée via une SVD (Singular Value Decomposition), dont les sorties sont l’objet d’un post-traitement pour produire le résultat recherché. L’étape de la SVD est en général l’étape limitante pour le passage à l’échelle, à savoir le traitement de données massives issues de tableaux de très grandes dimensions : la complexité est cubique avec la taille des tableaux.

Or, des progrès significatifs ont été accomplis récemment pour le passage à l’échelle de la SVD en combinant trois éléments :

- une évolution de l’algorithme du calcul de la SVD en y incluant la Gaussian Random Projection (rSVD),
- une implémentation des calculs en mémoire distribuée pour les opérations élémentaires de calcul matriciel,
- l’utilisation d’un paradigme de programmation à base de tâches pour l’assemblage de ces étapes.

Ces progrès ont notamment été réalisés,

→ pour l’intégration de la rSVD dans l’algorithme de la MDS, par une collaboration entre HiePACS, Pleiade, l’IDRIS et l’INRAE (voir [BCF+18])

RR n° 9488
→ pour l’implémentation numérique (mémoire distribuée, programmation par graphe de tâche),
par une collaboration entre les équipes HiePacs, Tadaam, Storm et Pleiade au cours d’une
ADT appelée ADT Gordon, réalisée en 2019-2020 (voir [ACD⁺22]).

En livrant cette présentation simplifiée de la diversité des méthodes linéaires de réduction de
dimension, en explicitant les étapes successives qui font appel au calcul matriciel, en insistant
sur le gué commun qui est le passage à l’échelle de la SVD pour traiter confortablement des
données massives, en nous appuyant sur les succès de l’ADT Gordon quant au passage à l’échelle
de la MDS, nous espérons faciliter l’extension de cette approche avec le même succès dans le jardin
zoologique des méthodes linéaires de réduction de dimension, pour inférer certains patterns en
révélant une structure de rang faible dans les matrices étudiées.

A.F., à Pierroton, le 22 septembre 2022
## Contents

1 Introduction ................................................. 7

2 Multivariate Data Analysis ................................. 8

3 Principal Component Analysis (PCA) ......................... 11
   3.1 Setting the problem .................................... 12
   3.2 Solving the problem ..................................... 13
   3.3 Link with SVD ......................................... 15
   3.4 Randomized SVD ........................................ 15
   3.5 Core algorithm for PCA ................................ 17
   3.6 Interpretation and plotting ............................. 19
   3.7 Classical analysis ...................................... 22

4 Complements on PCA ........................................... 24
   4.1 Statistical approach ..................................... 24
   4.2 Factor Analysis and PCA ................................. 25
   4.3 Unitarily invariant norms ................................. 27

5 PCA with Instrumental Variables .............................. 28
   5.1 Setting the problem ..................................... 29
   5.2 Solving the problem ..................................... 30
   5.3 Interpretation of PCAiv .................................. 31
   5.4 Non orthonormal basis ................................... 32

6 PCA with metrics on rows and columns ....................... 33
   6.1 Metrics and weights on row and column spaces .......... 33
   6.2 Setting the problem ..................................... 35
   6.3 Solving the problem ..................................... 36
   6.4 Isometry .................................................. 37
   6.5 Interpretation and plotting ................................ 38
   6.6 PCA with metrics and instrumental variables .......... 40

7 Correspondence Analysis ....................................... 41
   7.1 Link with $\chi^2$ distance ............................... 41
   7.2 Description of the method ............................... 43
   7.3 CoA and geometry of point clouds ....................... 44
   7.4 Classical presentation ................................... 44

8 Canonical Correlation Analysis .............................. 46
   8.1 Stating the problem ..................................... 47
   8.2 Solving the problem ..................................... 47
   8.3 Computation of the solution ............................. 49

9 Multiple Correspondence Analysis ........................... 51
   9.1 A tight link between Canonical Analysis and Correspondence Analysis ................................. 51
   9.2 Link between Canonical Analysis and PCA with metric on rows .................................. 52
   9.3 Multiple Correspondence or Canonical Analysis ........................................... 53

RR n° 9488
## Multidimensional Scaling

| Section                                | Page |
|----------------------------------------|------|
| 10.1 The Gram matrix                   | 56   |
| 10.2 Eigendecomposition of the Gram Matrix | 57   |
| 10.3 Dimension reduction               | 58   |
| 10.4 MDS algorithm                     | 58   |

## Summary

| Section                                | Page |
|----------------------------------------|------|
| 11 Summary                             | 60   |

## References in textbooks

| Section                                | Page |
|----------------------------------------|------|
| 12 References in textbooks             | 61   |
1 Introduction

Let us start with an example: supervised learning with Support Vector Machine. Imagine a training set of \( n \) observations, each observation being a pair \((x, y)\), with \( x \in \mathbb{R}^p \) and \( y \in \{-1, 1\} \). One wishes to predict \( y \) as an outcome of a new observation \( x \), not in the training set, and where \( y \) is unknown. Avoiding technicalities, and keeping the presentation short for this introduction, this can be done in adequate situations when there exists a linear discriminating function \( f(x) = \beta + \sum w_i x_i \), and that \( y = 1 \) if \( f(x) > 0 \) and \( y = -1 \) if \( f(x) < 0 \). A separating hyperplane is an hyperplane such that all points \( x \in \mathbb{R}^p \) of pairs \((x, 1)\) (say, blue points) are on one side and all points \( x \) of pairs \((x, -1)\) (say red points) are on the other side of the hyperplane. Such a function \( f \) is not unique, and here support vectors come into the game. The margin is defined as the minimal distance between the points \( x_i \) of the training set and the separating hyperplane \( f(x) = 0 \). SVM is finding a linear function \( f \) with maximum margin. This is equivalent to finding two parallel separating hyperplanes with maximal mutual distance. If the dimension \( p \) of the space where the observations \( x \) are given is large, this can lead to high computation load. However, there is a lemma\(^1\) called Johnson-Lindenstrauss lemma telling that there exists a space \( H \) of smaller dimension \( d \ll p \) (of \( O(\log p) \)) such that all pairwise distances are preserved up to a high accuracy while projecting on \( H \). Therefore, SVM can be fit on the projection of the training set on a space of much lower dimension. It appears that J-L lemma is at the heart of very efficient methods to perform Singular Value Decomposition of very large matrices, and SVD is at the heart of most of, if not all, dimension reduction methods in multivariate data analysis. This establishes a tight link between machine learning and multivariate data analysis.

Multivariate Data Analysis could be read, and is presented here, as an algebraic construction in linear algebra, around Singular Value Decomposition. This is deliberate, as we wish to focus on recent progress in High Performance Computing to handle very large data sets, and this requires a sound basis in linear algebra. Let us keep in mind that data distinguish statistics from probability: statistics are about inference of some models from data. Multivariate Data Analysis is about inferring some patterns from data, like correlation structure between items and features, or individuals and variables. As such, it is now part of Data Mining, Statistical Learning and Machine Learning. This can be summarized with the subtitle of [HTF09]: data mining, inference and prediction. Tibshirani has provided a dictionary between statistics and machine learning\(^2\), referred to in [Mur12], partly given here:

| Machine learning         | Statistics                |
|--------------------------|----------------------------|
| weights                  | parameters                |
| learning                 | fitting                   |
| supervised learning      | regression/classification  |
| unsupervised learning    | density estimation, clustering |

Beyond this dictionary, there is genuine innovation in machine learning, which is about inferring meaningful pattern in data in a very elaborate way, far beyond data analysis (see e.g. [SSBD14]). A grail is to mimic the way a brain learns from experience. This is beyond the scope of those

---

\(^1\)Which deserves to be called a theorem, but is referred to classically as a lemma

\(^2\)see http://www-stat.stanford.edu/~tibs/stat315a/glossary.pdf
notes.

Multivariate Data Analysis is a tool for learning patterns about correlations between features and items, which is one type among many others of data structure. Supervised or unsupervised learning may rely, at one step or another, on techniques inherited from multivariate data analysis, in a same way that multivariate data analysis relies at one step or another on tools inherited from linear algebra. This can be formalized by the following succession of steps:

```
Machine Learning  -->  Statistical learning  -->  Multivariate Data Analysis  -->  Linear Algebra
```

We have this inheritance in mind for these notes, with an objective of implementation of calculations for very large data sets.

There is a second reason for emphasizing the role of Multivariate Data Analysis in Machine Learning. Murphy recognized two approaches in Machine Learning [Mur12, sec. 1.1.2], which fit to the synthetic table of Tibshirani:

- a predictive or supervised approach, where a response variable is predicted from some features, from a set of observations where the response is known, called the training set
- a descriptive or unsupervised approach, where the objective is to find some interesting patterns, and is referred to as "pattern discovery".

Many of these techniques work well in low dimension, i.e. when the number of features to work with is small. Multivariate Data Analysis plays a key role in such a framework to produce a low dimension representation of an array connecting items and features, as close as possible to the original array. This is called Dimensionality Reduction (see e.g. [LV07]), and one of the key (linear) technique therefore is Principal Component Analysis (see section 3). So, one can say that MDA paves the way for elaborate machine learning processes.

Notes and references: There exists many excellent surveys for learning patterns from data. See e.g. [CST00, Mur12, SSBD14]. For a concise introduction to SVM, see [CST00] or [SS04].

2 Multivariate Data Analysis

Multivariate Data Analysis (MDA) is at the crossroad of three domains:

- It is about finding structures in data presented as arrays. This is algebra.
- It is possible to attach to an array a point cloud in a Euclidean space, and study the shape of the cloud. This is geometry.
- Data are modeled as realizations of random variables. This is statistics.

Hence, MDA is at the crossroad between algebra, geometry and statistics.

As arrays of data are matrices, MDA relies heavily on Linear Algebra. Producing a matrix $A$ is one of the most classical mathematical formalization of some information gathered on a set of data.
items. Let us consider a set of \( n \) items each characterized by \( p \) variables. The rows \( i \in [1,n] \) are the items, and the columns \( j \in [1,p] \) are some variables, often referred to as features in machine learning. The value of the feature \( j \) for the item \( i \) is the coefficient \( \alpha_{ij} \) of the matrix. One objective of multivariate data analysis is to describe how items and features are related, which is usually addressed by low rank approximation of matrix \( A \).

A point cloud is a geometric object associated to such a feature matrix. Provided the variables are quantitative, i.e., numbers in \( \mathbb{R} \), a set of \( n \) points in \( \mathbb{R}^p \) is built from \( A \), with one point \( a_i = (\alpha_{i1}, \ldots, \alpha_{ip}) \in \mathbb{R}^p \) for item \( i \). This point cloud as a geometric object is denoted \( \mathcal{A} \). In many real cases, the points are located in a low dimensional manifold. Finding such a manifold is called \textit{dimensionality reduction}. Efficient and well understood techniques exist when the manifold is linear (or affine): the best approximation of \( \mathcal{A} \) by a projection in a space of dimension \( r \) can be solved by finding the best approximation of \( A \) by a matrix of rank \( r \).

Row \( i \) of matrix \( A \) or point \( a_i \in \mathcal{A} \) can modeled as the realization of a set of random variables \( (X_1, \ldots, X_p) \). Questions of interest are the study of the dependence structure of the \( X_j \), given by the variance-covariance matrix of observed data, or exhibiting low dimension latent variables \( z \), such that observations \( x \) can be modeled as \( \mathbb{P}(x | z) \).

Many of these techniques have been progressively selected as core techniques along several decades since the beginning of 20th century. Many of these methods are presented and studied within the algebraic framework of linear algebra. For example, Principal Component Analysis can be built as a consequence of the Singular Value Decomposition of \( A \): \( A = U\Sigma V^T \). These methods have experienced\(^3\) two revolutions

\textbf{computing revolution:} the development of computing infrastructures especially in the 70’s has led to the mushrooming of scientific libraries first for mainframes, and later for a range of machines from laptops to computing clusters

\textbf{massive data revolution:} many sensors produce now myriads of bytes of data, like telescopes, satellites, sequencers, etc., raising the challenge to overcome the walls of time and memory while implementing those methods on massive data sets, leading to matrices of very large dimensions (like \( 10^5 \) to \( 10^6 \) rows or columns).

The progresses in these domains are due simultaneously to the derivation of new algorithms (like the random projection method for computing the Singular Value Decomposition of a large dense matrix, see \cite{HMT11}), development of new paradigms for implementing some algorithms (like Message Passing Interface for distributed parallelization), and progress in technology of computing infrastructures (like Graphic Processing Units).

The high diversity of existing method can be organized into a small set of iconic methods, knowing that such a classification is far from being either unique or universally adopted. However, most of textbooks presenting these methods progressively reach an agreement on the poles around which to organize their variety and similarities/dissimilarities. We will not discuss this here, but select among many possibilities a small set of methods, organized along a variety of questions they answer to. These methods are presented in the following table.

---

\(^3\)at least ...

RR n° 9488
### Acronym Method Section

| Acronym | Method                        | Section |
|---------|-------------------------------|---------|
| PCA     | Principal Component Analysis  | 3       |
| PCAsc   | Scaled-Centered PCA           | 3.7     |
| PCAda   | PCA with double averaging     | 3.7     |
| PCAiv   | PCA with instrumental variables | 5    |
| PCAmet  | PCA with metrics              | 6       |
| CoA     | Correspondence Analysis       | 7       |
| CCA     | Canonical Correlation Analysis| 8       |
| MCoA    | Multiple Correspondence Analysis| 9   |
| MDS     | Multidimensional Scaling      | 10      |

The organization followed here is to analyze each of these methods as a pipeline.

**Notes and references:** Multivariate Data Analysis is a classical domain in data analysis, still relevant, and underestimated (although as reported in [Jol02] and a current research with "Principal Component Analysis" on Google Scholar yielded several millions of hits). It has been developed along several lines, all over 20th century. Most of seminal papers have been published before 1935. Two trends coexist: a statistics oriented trend, developed by an Anglo-Saxon school, in UK, US, India and Scandinavia, and a French school, more algebraic and geometrical, developed in the 60's under leadership of J.-P. Benzecri. A comprehensive and very informative paper for comparing both approaches with historical insight is [TY85]. The seminal paper on PCA by Pearson in 1901 however is geometrical, and Hotelling presentation 30 years later is algebraic. Several recent and excellent textbooks exist for a global presentation of multivariate data analysis or dimensionality reduction, like [And58, MKB79, CC80, LV07, Ize08, Wan12].

Each has a special flavour: [And58] is the seminal book on multivariate data analysis (Anderson was in Stanford) and has been a bedside book of statisticians for decades; [MKB79] is the most comprehensive, with all demonstrations of results presented as theorems, [CC80] establishes a link with statistics, and is easier to read, [LV07] goes beyond linear methods, focusing on nonlinear methods, [Ize08] is comprehensive too, focusing on a diversity of examples, [Wan12] addressed explicitly the new challenge raised by massive data and work in high dimensional spaces. The seminal books for French school, more geometrical, are [Ben73b, Ben73a] and [CP79] who introduced the duality diagram as a unifying tool. See as well [LMT77, LMF82, EP90] among others. The very nice paper [PCY79] gives an historical sketch of the French school as well as a comparison with Anglo-Saxon school.

### Notations

The following notations are adopted throughout these notes:
3 Principal Component Analysis (PCA)

Principal Component Analysis is a Dimensionality Reduction technique which can be read in three different ways:

**geometric:** A point cloud $A$ of $n$ points in $\mathbb{R}^p$ being given, as well as an integer $r < p$, find an affine subspace $E \subset \mathbb{R}^p$ of dimension $r$ such that the projection $A_e$ of $A$ on $E$ is as close as possible to $A$.

**algebraic:** A $n \times p$ matrix $A$ being given, as well as an integer $r < p$, find a matrix $A_r$ of rank $r$ such that $\|A - A_r\|$ is minimum with Frobenius ($\ell^2$) norm.

**statistical:** A set of $p$ random variables being observed on $n$ items independently, find $r$ independent linear combinations of these variables with maximum variance (the first one is with maximum variance, the second one is uncorrelated with the first one and with maximum variance, and so on).

Here, we adopt the algebraic viewpoint, but start with giving some links with the geometric viewpoint, which is important for visualization of point clouds. Geometric approach is about dimension reduction, and algebraic approach is about best low rank approximation. Low rank approximation has many applications in numerical linear algebra. There are many links between algebraic and statistical approach too, which deserve to be further studied.

**Notes and references:** The historical development of Principal Component Analysis is well known and well documented. At the beginning, it appeared in Pearson (1901) under the guise of a geometric derivation (Pearson, K. - 1901 - On lines and planes of closest fit to systems of points in space. *Philos. Mag.* 6(2):559-572). The idea behind is clearly geometric. The term Factor Analysis has been introduced by Thurstone in 1931 (Thurstone, L. - 1931 - *Multiple Factor Analysis, Psychological Review*, 38:406-427). His purpose was to find a general method for finding factors which could explain correlations, following an idea published by Spearman in 1904. The presence of an underlying model in factor analysis has been the cause of numerous and fierce discussions (see [Jol02]). Hotelling gave an algebraic framework in 1933 (in Hotelling, H. - 1933 - *Analysis of a complex of statistical variables into principal components. J. Educ. Psychol.*, 24:498-520), where the term PCA first appeared. Following the work of Pearson, he showed that the principal axis are the eigenvectors of the covariance matrix of the sample. PCA is based on Singular Value Decomposition of a matrix. The link between PCA and SVD is classically attributed to a theorem published by Eckart & Young in 1936 [EY36]. Classical textbooks in anglo-saxon litterature dedicated to PCA are [Jac91], and [Jol02]. PCA is developed
in every textbook in multivariate data analysis, like [MKB79, CC80, Ize08]. [Wol87] is a survey of PCA tools with an introduction on main milestones in the development of the method. Classical textbooks in French literature are [CP79, LMF82]. A recent survey of PCA and its recent developments can be found in [JC16].

PCA is probably one of the most used tool in multivariate statistics. It is known as Karhunen-Loève decomposition in signal theory.

3.1 Setting the problem

Let us recall that the norm here and in the following sections is the Frobenius or \( \ell^2 \) norm unless otherwise stated:

\[
\|x\| = \left( \sum_i x_i^2 \right)^{1/2} \tag{3.1.1}
\]

for a vector and

\[
\|A\| = \left( \sum_{i,j} a_{ij}^2 \right)^{1/2} \tag{3.1.2}
\]

for a matrix.

The problem can be set as follows:

- **Algebraic approach:**

\[
\begin{aligned}
\text{Given} & \quad A \in \mathbb{R}^{n \times p} \\
& \quad 0 < r < p \\
\text{Find} & \quad A_r \in \mathbb{R}^{n \times p} \\
& \quad \text{with} \quad \text{rank} \ A_r = r \\
& \quad \text{such that} \quad \|A - A_r\| \text{ minimal}
\end{aligned}
\]

- **Geometric approach:**
Given a point cloud $A = (a_1, \ldots, a_n)$
$a_i \in \mathbb{R}^p$
$0 < r < p$

Find a subspace $E_r \subset \mathbb{R}^p$
with $\dim E_r = r$
such that $d(A, A_r)$ minimal
where $d(A, A_r) = \sum_i \|a_i - \tilde{a}_i\|^2$
and $\tilde{a}_i$ is the projection of $a_i$ on $E_r$

Let us note that a point cloud $A$ is equivalent to a matrix $A$ with the row $i$ of $A$ being the point $a_i \in A$. Knowing that, both approaches are equivalent. To see that, let us consider an orthonormal basis $(v_1, \ldots, v_r)$ of $E_r$, and let us complete it to have an orthonormal basis of $\mathbb{R}^p$ as $(v_1, \ldots, v_r, v_{r+1}, \ldots, v_p)$. If the projection is exact, i.e. if $\forall i, a_i = \tilde{a}_i$, the columns $r + 1$ to $p$ of $A$ in basis $V$ are zero, and $A$ is of rank $r$. The converse is true as well.

### 3.2 Solving the problem

The solution to this problem is well known (see any textbook mentioned in the introduction of this section). Finding the subspace $E_r$ is finding an orthonormal basis for it. Let us fix a subspace $E_r \subset \mathbb{R}^p$ of dimension $r$. If $\tilde{a}_i$ is the projection of $a_i$ on $E_r$, we have, by Pythagorean theorem

$$\forall i \in [1, n], \quad \|a_i\|^2 = \|\tilde{a}_i\|^2 + \|a_i - \tilde{a}_i\|^2$$

Then, setting $\sum_i \|a_i - \tilde{a}_i\|^2$ minimum is equivalent to setting $\sum_i \|\tilde{a}_i\|^2$ maximum. We then can set PCA as

| Given a point cloud $A = (a_1, \ldots, a_n)$
   $a_i \in \mathbb{R}^p$
   $0 < r < p$

| Find a subspace $E_r \subset \mathbb{R}^p$
   $\dim E_r = r$

| such that $\sum_i \|\tilde{a}_i\|^2$ maximal
   $\tilde{a}_i$ is the projection of $a_i$ on $E_r$

Let us consider the simple case $r = 1$ and $0 \in E_r$. If $u$ with $\|u\| = 1$ is a basis of $E_1$, we have $\tilde{a}_i = (a_i, u)u$, and the optimization problem can be stated as
Given a point cloud \( \mathcal{A} = (a_1, \ldots, a_n) \) with \( a_i \in \mathbb{R}^p \)

Find a vector \( u \in \mathbb{R}^p \) with \( \| u \| = 1 \) such that \( \| \mathcal{A} u \| \) is maximal \hspace{1cm} (3.2.2)

Indeed, \( \tilde{a}_i = \langle a_i, u \rangle u \). Then \( \sum_i \| \tilde{a}_i \|^2 = \sum_i (\langle a_i, u \rangle)^2 \). The elements \( \langle a_i, u \rangle \) are the coordinates of the vector \( y = \mathcal{A} u \). The solution of such a problem is classical: \( u \) is the eigenvector of \( \mathcal{A}^T \mathcal{A} \) associated with the largest eigenvalue

\[
\mathcal{A}^T \mathcal{A} v = \lambda v, \quad \lambda = \max \{ \lambda \in \text{Sp} \mathcal{A}^T \mathcal{A} \} \hspace{1cm} (3.2.3)
\]

The Eckart-Young theorem [EY36] extends this result to \( r > 1 \) and states that an orthonormal basis of \( E_r \) is the set \( (v_1, \ldots, v_r) \) with

\[
\mathcal{A}^T v_j = \lambda_j v_j \quad \text{with} \quad \lambda_1 \geq \ldots \geq \lambda_r \geq \lambda_{r+1} \geq \ldots \geq \lambda_p \geq 0 \hspace{1cm} (3.2.4)
\]

This is a direct application of the variational properties of the Rayleigh quotients (see [HJ12, sec. 4.2]). This can be written

\[
\mathcal{A}^T A V = V \Lambda \hspace{1cm} (3.2.5)
\]

where \( V \) is the \( p \times p \) matrix with column \( j \) being \( v_j \) and \( \Lambda \) is the diagonal matrix with terms \( (\lambda_i) \), in the diagonal (in decreasing order). The coordinates of the point cloud \( \mathcal{A} \) in new basis \( V \) are given by

\[
Y = AV \hspace{1cm} (3.2.6)
\]

Hence an first algorithm:

---

**Algorithm 1** PCA of a matrix with EVD: \( \text{PCA}_\text{EVD}(A) \)

1. **input** \( A \in \mathbb{R}^{n \times p} \)
2. **compute** \( C = A^T A \)
3. **compute** \( (\lambda_\alpha, v_\alpha) \) such that \( C v_\alpha = \lambda_\alpha v_\alpha \), or \( CV = \Lambda A \)
4. **compute** \( Y = AV \)
5. **return** \( Y, \Lambda, V \)

---

- The vectors in new basis are called *principal axis*, and the coordinates along the principal axis are called *principal components*.

- Here is a summary of the results:

| \( C \in \mathbb{R}^{p \times p} \) | matrix of correlations of columns of \( A \) |
| \( V \in \mathbb{R}^{p \times p} \) | new basis |
| \( \Lambda \in \mathbb{R}^p \) | eigenvalues of \( C \) in decreasing order |
| \( Y \in \mathbb{R}^{n \times p} \) | matrix of coordinates in new basis |

Inria
### 3.3 Link with SVD

Let \((U, \Sigma, V)\) be the SVD of \(A\).

\[
A = U \Sigma V^T
\]

Then

\[
C = A^T A = (V \Sigma U^T)(U \Sigma V^T) = V \Sigma^2 V^T \quad \text{as} \quad U^T U = I_n
\]

and

\[
CV = V \Sigma^2 \quad \text{as} \quad V^T V = I_p
\]

Hence, \(V\) as new basis for PCA of \(A\) is the matrix \(V\) in SVD of \(A\), and \(\Lambda = \Sigma^2\). We have

\[
Y = AV = U \Sigma V^T V = U \Sigma
\]

This yields a second algorithm for PCA:

---

**Algorithm 2** PCA of a matrix with SVD: \(\text{PCA } \text{sVD}(A)\)

1: **input** \(A \in \mathbb{R}^{n \times p}\)
2: **compute** \(U, \Sigma, V = \text{svd}(A)\)
3: **compute** \(\Lambda = \Sigma^2\)
4: **compute** \(Y = U \Sigma\)
5: **return** \(Y, \Lambda, V\)

---

Adopting the SVD viewpoint has some advantages:

- \(\rightarrow\) there exists efficient and stable algorithms for computing a SVD,
- \(\rightarrow\) it avoids a matrix \(\times\) matrix computation \((C = A^T A)\) which can be costly when \(n\) and \(p\) are large,
- \(\rightarrow\) it leads to easier generalization with instrumental variables or metrics on row or column space (see section 7),
- \(\rightarrow\) If the dimensions \(n\) and \(p\) are (very) large, the SVD can be computed with random projection (see [HMT11]). Such a calculation is presented below.

### 3.4 Randomized SVD

Let \(A \in \mathbb{R}^{n \times p}\) with \(n \geq p\). The complexity (number of operations) of the SVD of \(A\) is in \(O(n^2 p)\). SVD becomes untractable for large values of \(n\) and \(p\), say \(10^4\). Fortunately, there are some very efficient heuristics, with bounds on errors, to compute the first singular values and vectors, based
on randomized algorithms. The use of randomized algorithm to compute efficiently the SVD of a very large matrix is fully developed in [HMT11]. The idea behind is the following.

If \( Q \in \mathbb{R}^{n \times k} \) is columnwise orthonormal, the projection of the columns of \( A \) on the vector space spanned by the columns of \( Q \) is

\[
\tilde{A} = QQ^\top A
\]

Let us denote

\[
B = Q^\top A, \quad B \in \mathbb{R}^{k \times p}
\]

Then, \( \tilde{A} = QB \) is a rank \( k \) approximation of \( A \). The SVD of \( B \) (\( B = U_n \Sigma V \)) is in \( \mathcal{O}(p^2k) \) instead of \( \mathcal{O}(n^2p) \), and we have

\[
A \approx QB = Q(U_n \Sigma V) = (QU_n) \Sigma V = U \Sigma V
\]

So

\[
A \approx U \Sigma V \quad \text{with} \quad U = QU_n \quad (3.4.1)
\]

Next step is to show that (what we will not develop here), when \( n \) and \( p \) are large, \( A \approx QQ^\top A \) with high quality for any random matrix \( Q \). This comes from deep theorems in geometry of Banach spaces, and from Johnson-Lindenstrauss lemma which states that, for any \( \epsilon > 0 \), and any dimension \( n \), there exists a dimension \( k \) such that for any cloud \( X \) of \( n \) points, there exists an embedding

\[
f : \mathbb{R}^n \longrightarrow \mathbb{R}^k
\]

such that for any \( x, y \in X \)

\[
(1 - \epsilon)\|x - y\|^2 \leq \|f(x) - f(y)\|^2 \leq (1 + \epsilon)\|x - y\|^2 \quad (3.4.2)
\]

It is shown by showing that such an embedding exists with probability one. The dimension \( k \) must comply with

\[
k \geq \frac{8 \log n}{\epsilon^2} \quad (3.4.3)
\]

The bad news is that \( \epsilon^2 \) is at the denominator (so, \( k \) is large when \( \epsilon \) is small), but the good news is that \( k \) grows with \( \log n \) and not \( n \). This becomes efficient when \( n \) is large. So, \( Q \) as an orthonormal basis is build as the \( QR - \) decomposition of \( Y = A\Omega \) where \( \Omega \) is a random matrix \( (Y \) is in the span of \( A \)). There are several ways to chose \( \Omega \), and here we restrict ourselves to the Gaussian random projection, i.e. \( \Omega \) is a random Gaussian matrix with

\[
\Omega[i,j] \sim \mathcal{N}(0,1)
\]

Usually, for a good accuracy at rank \( k \), it is advised to select \( \Omega \) as \( n \times k' \) with \( k' = k + s \) where \( s \) is called the oversampling. Usually, taking \( s = 5 \) is said to be sufficient. The reader is encouraged to read [HMT11] for further details and explanations on randomized algorithms in matrix computations (what is presented here is the tip of the iceberg). The algorithm runs as follows:
Algorithm 3 SVD of a matrix with Gaussian Random projection \texttt{svd\_grp}(A, k)

1: \textbf{input} $A \in \mathbb{R}^{n \times p}$, $k$ as prescribed rank
2: \textbf{build} $\Omega \in \mathbb{R}^{p \times k}$, random ($\Omega[i,j] \sim \mathcal{N}(0,1)$)
3: \textbf{compute} $Y = A\Omega$
4: \textbf{compute} the QR-decomposition of $Y$: $Y = QR$
5: \textbf{build} $B = Q^tA$
6: \textbf{run} the SVD of $B$: $B = U_b\Sigma V$, or $(U_b, \Sigma, V) = \text{svd}(B)$
7: \textbf{compute} $U = QU_b$
8: \textbf{return} $U, \Sigma, V$

Here are the dimensions of the involved matrices:

| Matrix | dimensions | computation |
|--------|------------|-------------|
| $A$    | $n \times p$ | data       |
| $\Omega$ | $p \times k$ | Gaussian random matrix |
| $Y$    | $n \times k$ | $Y = A\Omega$ |
| $Q$    | $n \times k$ | $Y = QR$ |
| $B$    | $k \times p$ | $B = Q^tA$ |
| $U_b$  | $k \times k$ | $B = U_b\Sigma V$ |
| $\Sigma$ | $k \times k$ | idem |
| $V$    | $k \times p$ | idem |
| $U$    | $n \times k$ | $U = Q^tU_b$ |

3.5 Core algorithm for PCA

Wrapping all this together leads to a core algorithm for PCA, where the user can select which method to implement, presented hereafter in pseudocode:
Algorithm 4 PCA of a matrix: PCA_CORE(A, k = −1, meth=svd)

1: input \( A \in \mathbb{R}^{n \times p}, k \in \mathbb{N} \cup \{-1\}, \text{meth} \in \{\text{evd, svd, grp}\} \)
2: if \( \text{meth}==\text{evd} \) then
3: \hspace{1em} compute \( C = A^tA \)
4: \hspace{1em} compute \((\lambda_\alpha, v_\alpha)\) such that \( Cv_\alpha = \lambda_\alpha v_\alpha, \) or \( CV = V\Lambda \)
5: \hspace{1em} compute \( Y = AV \)
6: \hspace{1em} if \( k > 0 \) then
7: \hspace{2em} \( Y = Y[:,0:k]; \quad \Lambda = \Lambda[0:k]; \quad V = V[:,0:k] \)
8: \hspace{1em} end if
9:\hspace{1em} end if
10: if \( \text{meth}==\text{svd} \) then
11: \hspace{1em} compute \( U, \Sigma, V = \text{svd}(A) \)
12: \hspace{1em} compute \( \Lambda = \Sigma^2 \)
13: \hspace{1em} compute \( Y = U\Sigma \)
14: \hspace{1em} if \( k > 0 \) then
15: \hspace{2em} \( Y = Y[:,0:k]; \quad \Lambda = \Lambda[0:k]; \quad V = V[:,0:k] \)
16: \hspace{1em} end if
17:\hspace{1em} end if
18: if \( \text{meth}==\text{grp} \) then
19: \hspace{1em} compute \( U, \Sigma, V = \text{svd_grp}(A,k) \)
20: \hspace{1em} compute \( \Lambda = \Sigma^2 \)
21: \hspace{1em} compute \( Y = U\Sigma \)
22:\hspace{1em} end if
23: return \( Y, \Lambda, V \)

• Comments: Here are some comments:

Why \texttt{PCA}\_\texttt{CORE}? The reason for the name is the following: PCA is a method which very seldom runs dimension reduction or approximation by a low rank matrix directly on the data matrix. Most of the times, there is a pretreatment (see section 3.7), like centering and scaling columnwise, and dimension reduction is performed after pretreatment. The name PCA designates the whole analysis:

1. a pretreatment 
   
   \[
   A \xrightarrow{\text{pretreatment}} A' \]

2. the treatment itself 
   
   \[
   Y, \Lambda, V = \text{PCA}\_\texttt{CORE}(A')
   \]

   which is denoted \texttt{PCA}\_\texttt{CORE}().

Choice of the method: Three methods are described here: eigenvalues of the correlation matrix (evd), SVD of the data matrix (svd), and SVD with Gaussian Random Projection (grp). Here is an advice for selecting the right method

→ If the size of the matrix (number \( p \) of columns) is small to medium (say \( \approx 10^4 \)), then evd or svd can be used

→ If the size of the matrix is large to very large (\( 10^4 < p < 10^6 \)), gaussian random projection must be used.
Prescribed rank: \( k \) is the prescribed rank. Its default value is \( k = -1 \), which means that all the eigenvalues, or singular values, and components and axis will be computed. This is relevant for methods \textsc{evd} or \textsc{svd} only. A rank \( k > 0 \) must be prescribed for method \textsc{grp}. If a rank \( k > 0 \) is prescribed, the first \( k \) eigenvalues or singular values, components and axis only will be computed.

3.6 Interpretation and plotting

In this section, the geometrical viewpoint is adopted. Interpretation of the results of a \textsc{PCA} is about quantifying the fraction of inertia of the point cloud (i.e. variance of the associated variables, or norm of the associated matrix) which is preserved by projection either on one axis or on a space spanned by \( r \) first axis. When the point cloud is centered, \textsc{PCA} is finding a rotation in \( \mathbb{R}^p \) such that these quantities are maximal.

- Let \( A \in \mathbb{R}^{n \times p} \) be a matrix, \( A \) its associated point cloud in \( \mathbb{R}^p \), and \((Y, V, \Lambda)\) the \textsc{PCA} of \( A \). Then, the column \( y_j \) of \( Y \) with \( j \in [1, p] \) is the vector in \( \mathbb{R}^n \) of the coordinates of the points of \( A \) on principal axis \( j \).

Proof. Indeed, let us have a point cloud \( A \) made of \( n \) points \( a_i \in \mathbb{R}^p \) with \( a_i \) being row \( i \) of \( A \). \textsc{PCA} of \( A \) is performing a SVD of \( A \) as

\[
A = U \Sigma V^T
\]  

and yields a new orthogonal basis \((v_1, \ldots, v_p)\) of \( \mathbb{R}^p \), where \( v_j \) is the column \( j \) of \( V \). Let us recall that

\[
Y = U \Sigma
\]  

Then

\[
Y = U \Sigma = U \Sigma (V^T V) = (U \Sigma V^T) V = AV
\]

which means that the rows of \( Y \) are the coordinates of the points of \( A \) in new basis \( V \). \( \square \)

- If \( y \in \mathbb{R}^n \) and \( v \in \mathbb{R}^p \), let us recall the notation \( \otimes \) for tensor product

\[
y \otimes v \equiv y v^T
\]

(i.e. \((y \otimes v)_{ij} = y_i v_j \)). Then,

\[
A = \sum_{j=1}^{p} y_j \otimes v_j
\]  

from which

\[
\|A\|^2 = \sum_{j} ||y_j||^2
\]  

RR n° 9488
Proof. Indeed,

\[ \|A\|^2 = \left\| \sum_{j=1}^{p} y_j \otimes v_j \right\|^2 = \left\langle \sum_{j=1}^{p} y_j \otimes v_j , \sum_{k=1}^{p} y_k \otimes v_k \right\rangle = \sum_{j,k} \langle y_j , y_k \rangle \cdot \langle v_j , v_k \rangle = \sum_{j} \langle y_j , y_j \rangle = \sum_{j} \|y_j\|^2 \]

(another way to see this is to observe that \( Y \) is deduced from \( A \) by a rotation, which is an isometry, then \( \|A\| = \|Y\| \)). The tensor product \( \otimes \) has been introduced here because it will be used later. As \( Y = U\Sigma \) with \( U^T U = I_p \), we have

\[ \|y_j\| = \sigma_j \]

(3.6.6)

Hence, the norm of \( A \) can be partitioned as

\[ \|A\|^2 = \sum_{j=1}^{p} \sigma_j^2 \]

(3.6.7)

Let us recall that

\[ \lambda_j = \sigma_j^2, \quad A^T Av_j = \lambda_j v_j \]

Then

\[ \|A\|^2 = \sum_{i=1}^{p} \lambda_i \]

(3.6.8)

and the quality of the representation of \( A \) by its projection on the axis spanned by \( v_j \) is

\[ \varrho_j = \frac{\lambda_j}{\sum_i \lambda_i} \]

(3.6.9)

The quality of representation of the point cloud (i.e. of array \( A \)) by its projection \( A_r \) on the subspace spanned by vectors \( (v_1, \ldots, v_r) \) is

\[ \rho_r = \sum_{j=1}^{r} \varrho_j = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_i \lambda_i} \]

(3.6.10)
• The quality of representation of item $i$ on axis $j \in \{1, p\}$ is

$$
\psi(i, j) = \frac{y_{ij}^2}{\sum_{\ell=1}^{p} y_{i\ell}^2}
$$

and the quality of projection of item $i$ on the subspace spanned by vectors $(v_1, \ldots, v_r)$ is

$$
\theta(i, r) = \sum_{j=1}^{r} \psi(i, j) = \frac{\sum_{j=1}^{r} y_{ij}^2}{\sum_{\ell=1}^{p} y_{i\ell}^2}
$$

We have

$$
\begin{align*}
\varrho_j &= \sum_{i=1}^{n} \psi(i, j) \\
\rho_r &= \sum_{i=1}^{n} \theta(i, r)
\end{align*}
$$

This can be summarized as

| Quality of representation of | Notation | Calculation | Observation |
|-----------------------------|----------|-------------|-------------|
| item $i$ on axis $j$        | $\psi(i, j)$ | $\frac{y_{ij}^2}{\sum_{\ell=1}^{p} y_{i\ell}^2}$ | $\sum_{i=1}^{n} \psi(i, j)$ |
| item $i$ on subspace $E_r$  | $\theta(i, r)$ | $\sum_{j=1}^{r} \psi(i, j)$ | $\sum_{i=1}^{n} \theta(i, r)$ |
| point cloud on axis $j$    | $\varrho_j$ | $\frac{\lambda_j}{\sum_{i=1}^{n} \lambda_i}$ | $\sum_{i=1}^{n} \psi(i, j)$ |
| point cloud on subspace $E_r$ | $\rho_r$ | $\sum_{j=1}^{r} \varrho_j$ | $\sum_{i=1}^{n} \theta(i, r)$ |

• **Prescribed rank or accuracy:** In the algebraic framework, PCA is about the best low rank approximation of a matrix. This can be set in two guises:

\[ \rightarrow \text{select a rank } r, \text{ and deduce the quality } \rho \text{ of the approximation}, \]

\[ \rightarrow \text{select a quality of an approximation, and deduce the rank at which it should be done.} \]

The former is PCA at prescribed rank, whereas the latter is PCA at prescribed accuracy. The key tool for implementing the one or the other is the curve $r \mapsto \rho(r)$.
3.7 Classical analysis

Here, we denote $A \geq 0$ if all coefficients in $A$ are nonnegative. Let us recall that $a_i \in \mathbb{R}^p$, $a_{is}$ denotes the row $i$ of $A \in \mathbb{R}^{n \times p}$, and $a_{sj} \in \mathbb{R}^n$ its column $j$.

The mean and standard deviation of a distribution are often the best summary of it. If PCA yields the best rank one approximation, it is likely that first axis and components mirror this best summary and bring few information on the inner structure of matrix $A$. Hence a standard procedure is to center and scale a dataset, and run the PCA on the scaled and centered dataset to focus on the inner structure (e.g. correlations between columns).

- **Centering**: Centering a matrix $A$ is translating the attached point cloud $\mathcal{A}$ to its barycenter:

  \[
  a_i \xrightarrow{\text{centering}} \overline{a}_i = a_i - g,
  \]

  where $g \in \mathbb{R}^p$ is the barycenter of the point cloud, i.e.

  \[
  g_j = \frac{1}{n} \sum_i a_{ij}
  \]

  It is easy to check that $\sum_i \overline{a}_i = \sum_i (a_i - \frac{1}{n} \sum_i a_i) = \sum_i a_i - \sum_i a_i = 0$. Matrix $\overline{A}$ is centered columnwise:

  \[
  \forall j, \sum_i \overline{a}_{ij} = 0
  \]

- **Scaling**: Scaling a matrix columnwise is dividing each column vector $a_{sj}$ by its norm, aka its standard deviation if it is centered:

  \[
  a_{sj} \xrightarrow{\text{scaling}} \overline{a}_{sj} = \frac{a_{sj}}{\|a_{sj}\|}
  \]

  The centered-scaled matrix $A'$ is defined by

  \[
  a_{sj} \xrightarrow{\text{scaled-centered}} \overline{a}_{sj} = \frac{a_{sj} - g_j}{\|\overline{a}_{sj}\|}
  \]

  with $\overline{a}_{sj} = a_{sj} - g_j 1_n$ (3.7.5)

- **Scaled-centered PCA** of a matrix $A$ is defined as:

  **Algorithm 5** pca-sc($A$)

  1: input $A \in \mathbb{R}^{n \times p}$
  2: compute the barycenter of $A$: $g = \frac{1}{n} \sum_i a_{is}$
  3: center $A$: $\forall i, \ a_i \rightarrow \overline{a}_i = a_i - g$
  4: scale $\overline{A}$: $\forall j, \overline{a}_{sj} \rightarrow a'_{sj} = \frac{\overline{a}_{sj}}{\|\overline{a}_{sj}\|}$
  5: do $Y, \Lambda, V = \text{pca\_core}(A')$
  6: return $Y, \Lambda, V$

- There are a few elementary results for scaled-centered PCA. By definition, the coefficients $c_{j\ell}$ of $C = A'^\top A'$ are the correlations between centered scaled variables $a'_{sj}$ and $a'_{s\ell}$. Hence, we have

  \[
  -1 \leq c_{j\ell} \leq 1
  \] (3.7.6)

  Inria
We have as well

\[ \forall j, \quad c_{jj} = 1 \quad (3.7.7) \]

Hence

\[ \sum_j \lambda_j = \text{Tr} C = p \quad (3.7.8) \]

Hence, the quality of approximation at rank \( r \) of \( A' \) is

\[ \rho_r = \frac{1}{p} \sum_{j \leq r} \lambda_j \quad (3.7.9) \]

- **Double averaging or bicentering:** Let us have a matrix \( A \geq 0 \) which is for example an array of counts. A classical example is a contingency table (contingency tables can be analysed with Correspondence Analysis, see section 7). The structure of \( A \) is dominated by the property \( A \geq 0 \). If a PCA of \( A \) is run, this will be the main (trivial) information given by axis 1. This trivial information can be filtered out by setting the model

\[
a_{ij} = m + \underbrace{x_i}_{\text{global mean}} + \underbrace{y_j}_{\text{effect of row } i} + \underbrace{r_{ij}}_{\text{effect of column } j + \text{residuals}}
\]

with

\[
\begin{align*}
\sum_i x_i &= 0 \\
\sum_j y_j &= 0 \\
\forall j, \sum_i r_{ij} &= 0 \\
\forall i, \sum_j r_{ij} &= 0
\end{align*}
\]

Then, we have

\[
\begin{align*}
m &= \frac{1}{np} \sum_{i,j} a_{ij} \\
x_i &= \left( \frac{1}{p} \sum_j a_{ij} \right) - m \\
y_j &= \left( \frac{1}{n} \sum_i a_{ij} \right) - m
\end{align*}
\]

**Proof.** We have

\[ \sum_{i,j} a_{ij} = np \quad m \]

so

\[ m = \frac{1}{np} \sum_{i,j} a_{ij} \]

Then

\[ \sum_i a_{ij} = nm + ny_j \]

RR n° 9488
and
\[ y_j = \left( \frac{1}{n} \sum_i a_{ij} \right) - m \] (3.7.13)

Similarly
\[ \sum_j a_{ij} = pm + px_i \]

and
\[ x_i = \left( \frac{1}{p} \sum_j a_{ij} \right) - m \] (3.7.14)

So, we have
\[
a_{ij} = m + x_i + y_j + r_{ij}
\]
\[
= m + \left( \frac{1}{p} \sum_j a_{ij} \right) - m + \left( \frac{1}{n} \sum_i a_{ij} \right) - m + r_{ij}
\]
\[
= -\frac{1}{np} \sum_{i,j} a_{ij} + \left( \frac{1}{p} \sum_j a_{ij} \right) + \left( \frac{1}{n} \sum_i a_{ij} \right) + r_{ij}
\] (3.7.15)

which is denoted as well
\[ r_{ij} = a_{ij} - a_i - a_j + a.. \] (3.7.16)

• PCA with double averaging is

1. computing the global mean \( m \), each effect \( x_i \) and \( y_j \) and the matrix of residuals \( R \)

2. run the PCA of \( R \), which is already centered, without scaling.

4 Complements on PCA

4.1 Statistical approach

Let us have \( p \) random variables \( X_1, \ldots, X_p \) observed each on \( n \) independent items. The \( n \) observations for variable \( j \) are the column \( j \) of a matrix \( X \), which is itself the realization of a random variable, the joint law of the \( X_j \). The variables \( X_j \) are assumed to be centered: \( \mathbb{E}(X_j) = 0 \). PCA at rank \( r \) is about finding \( r \) independent linear combinations of the \( X_j \) which have maximum variance.

• Let us denote as in [And58] the variance-covariance of \( X \) as \( \Sigma \), which is standard in statistics\(^\text{4}\): \( \Sigma = X^\top X \) as \( X \) is centered. There is no need for the random variables \( X_j \) to be Gaussian,

\(^\text{4}\)\( \Sigma \) is the standard notation for the diagonal matrix of singular values of a given matrix as well. Confusion can easily be avoided from context.
although some more can be said if they are Gaussian. Let \( u \in \mathbb{R}^p \) and let us consider the r.v. 
\[ Xu = \sum_j u_j X_j. \]
It is centered as 
\[ E(Xu) = \sum_j u_j E(X_j) = 0, \]
and its variance is
\[ \text{var} Xu = E(Xu)^2 = \langle Xu, Xu \rangle \]
as \( X \) is centered.
\[ \langle Xu, Xu \rangle = \langle u, \Sigma u \rangle \]
Maximizing \( \text{var} Xu \) with the constraint \( \|u\| = 1 \) yields that \( u \) is the eigenvector of \( \Sigma \) associated to its largest eigenvalue.

- The general result is [And58, th. 11.2.1]: Let \( X \) be a random variable on \( \mathbb{R}^p \), with \( E(X) = 0 \). Let us denote \( \text{var} X = \Sigma \). Then, there exists an rotation 
\[ V = Xu \]
such that the covariance matrix of \( V \) is diagonal and the \( k \)th component of \( V \) has maximum variance among all normalized linear combinations uncorrelated with \( V_1, \ldots, V_{k-1} \).

- **Probabilistic PCA:** This statistical approach has been recently renewed by so called probabilistic PCA, presented in [Bis06] where observed data are realizations of a Gaussian random variable conditionally to Gaussian latent variables. Let us have unobserved Gaussian latent variable \( z \in \mathbb{R}^r \) with
\[ z \sim \mathcal{N}(0, \mathbb{I}_r) \]
The probability distribution of observed variables \( x \in \mathbb{R}^p \) is given conditionally to a choice for the latent variables, as
\[ P(x \mid z) \sim \mathcal{N}(Wz + \mu, \sigma^2 \mathbb{I}_p) \]
where \( W \in \mathbb{R}^{p \times r} \) and \( \mu \in \mathbb{R}^p \). Then, the marginal distribution of \( x \) is given by
\[ p(x) = \int_z p(x \mid z)p(z) \, dz \]
which can be shown to follow (see [Bis06, p.573])
\[ P(x) \sim \mathcal{N}(\mu, WW^T + \sigma^2 \mathbb{I}_p) \]
Next step is to estimate the parameters \((W, \mu, \sigma)\) of the model knowing the observations. This is not so easy, and is presented in [Bis06, sect. 12.2.1].

**Notes and references:** A standard textbook for statistical approach to **PCA** is [And58]. This section is adapted from [And58, chap. 11]. Another key reference is [Rao64].

### 4.2 Factor Analysis and PCA

The idea behind Factor Analysis (FA) is that \( p \) observed correlated variables can be explained by \( r < p \) unobserved uncorrelated variables (hence the need for a statistical model). The unobserved variables are called *latent variables* in the statistical model. Even if many authors (including [Jol02, chap. 7]) have insisted on the difference between PCA and FA, it is tempting to say that
FA is a statistical approach of PCA where principal axis are latent variables. In that sense, it is very close to probabilistic PCA.

- Factor analysis is presented as follows in [And58, sect. 14.7]. Let us have a set of \( n \) observations of a random variable \( x \in \mathbb{R}^p \). Let us suppose there are
  
  → a vector \( z \in \mathbb{R}^r \) of non-observable factors (the factor scores)
  
  → a matrix \( \Omega \in \mathbb{R}^{p \times r} \) (the factor loadings)
  
  → a fixed vector \( \mu \in \mathbb{R}^p \) of means
  
  → \( u \in \mathbb{R}^p \) a vector of non observable errors

Then, \( x \) is modelled as

\[
x = \Omega z + \mu + u
\]  

(4.2.1)

It is assumed that

→ \( \mathbb{E}(z) = 0 \)

→ \( \mathbb{E}(u) = 0 \)

→ \( \mathbb{E}(zz^T) = M \)

→ \( \mathbb{E}(uu^T) = \Psi \) (diagonal)

→ \( \mathbb{E}(zu^T) = 0 \)

Then, it can be shown that \( \mathbb{E}(x) = \mu \) and

\[
\text{var} \: x = \mathbb{E}((x - \mu)(x - \mu)^T) = \Omega M \Omega^T + \Psi
\]  

(4.2.2)

A set of observations being given, the parameters of the model \((\Omega, M, \Psi, \mu)\) can be estimated by maximum of likelihood using principal components.

- According to [Bis06], FA is very close to probabilistic PCA, differing only in equation (4.1.2) by permitting more flexibility in the error term, written as

\[
P(x \mid z) \sim \mathcal{N}(Wz + \mu, \Psi)
\]  

(4.2.3)

where \( \Psi \) is a diagonal matrix in \( \mathbb{R}^{p \times p} \). The variance-covariance structure of the observed variables is split into a variance structure for each variable given by \( \Psi \) and the covariance structure given by \( W \).

**Notes and references:** Factor analysis has been developed in the 50’ and is presented in [And58, sect. 4.7], with references. See as well [Bis06, sect. 12.2] for a clear presentation of statistical approach of PCA, probabilistic PCA, and factor analysis.
4.3 Unitarily invariant norms

The problem of PCA as set in section 3.1 can be set for any norm, and not Frobenius norm only. Most widely used norms in data analysis are $\ell^1$ and $\ell^\infty$ norms, on top of $\ell^2$ norms. However, there are very few norms for which exact solution and efficient algorithms to compute a solution are known. One exception is the family of unitarily invariant norms.

- **Unitarily Invariant norm (UIN):** There is a generalization of Eckart-Young theorem through unitarily invariant norms. The framework is that of vector spaces on $\mathbb{C}$, but it can be applied on vector spaces on $\mathbb{R}$ as well. $U(\mathbb{C}^n)$ denotes the set of unitary matrices in $\mathbb{C}^n$, i.e. matrices having the property $UU^* = U^*U = I_n$, where $U^* = \overline{U^T}$ (and the same for $\mathbb{C}^p$). The equivalent in $\mathbb{R}^n$ is the set $O(\mathbb{R}^n)$ of orthogonal matrices such that $U^TU = I_n$.

- Let $E, F$ be two complex vector spaces, $A \in E \otimes F \simeq \mathcal{L}(F, E)$. A norm $\|\cdot\|$ of $E \otimes F \rightarrow \mathbb{R}^+$ is said unitarily invariant if

\[
\forall \begin{cases} U \in U(\mathbb{C}^n) \\ V \in U(\mathbb{C}^p) \\ A \in \mathbb{C}^{n \times p} \end{cases} \quad \|UAV^*\| = \|A\| \quad (4.3.1)
\]

For example, spectral and Frobenius norms are unitarily invariant norms. If $E, F$ are real vector spaces, the norm is said invariant by orthogonal transformation if

\[
\forall \begin{cases} U \in O(\mathbb{R}^n) \\ V \in O(\mathbb{R}^p) \\ A \in \mathbb{R}^{n \times p} \end{cases} \quad \|UAV^T\| = \|A\| \quad (4.3.2)
\]

- **Symmetric gauge function (SGF):** A norm $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^+$ is called a symmetric gauge function if it is invariant by any permutation of the coordinates in $\mathbb{R}^n$, i.e., if $\mathcal{P}$ is the set of permutations in $\mathbb{R}^n$

\[
\forall P \in \mathcal{P}, \quad \Phi(Px) = \Phi(x) \quad (4.3.3)
\]

- There is a remarkable link between unitary invariant norms and symmetric gauge functions. Let $A \in \mathbb{C}^{n \times p}$ (or $\in \mathbb{R}^{n \times p}$) and $\Sigma = (\sigma_1, \ldots, \sigma_p)$ its singular values. Let $\Phi$ be a SGF. To each SGF $\Phi$, one associates the norm $\|\cdot\|_{\Phi}$ defined by

\[
\|A\|_{\Phi} = \Phi(\sigma_1, \ldots, \sigma_n) \quad (4.3.4)
\]
Then, $\|\cdot\|_\Phi$ is a unitarily invariant norm. Let us note that if $\|\cdot\|$ is a UIN, $A$ a matrix in $\mathbb{C}^{n \times p}$ and $A = U \Sigma V^*$ the SVD of $A$, then $\|A = U^* A V\|$ and, as $U^* A V = \Sigma$, $\|A\| = \|\Sigma\|$ i.e. is a function of its singular values.

- **Mirsky’s theorem:** Mirsky has shown: Let $E, F$ be two vector spaces on $\mathbb{C}$ or $\mathbb{R}$, and $A, B \in E \otimes F \simeq \mathcal{L}(F, E)$. Let $(\alpha_1, \ldots, \alpha_p)$ (resp. $(\beta_1, \ldots, \beta_p)$) be the singular values of $A$ (resp. $B$). Then, for any unitarily invariant norm $\|\cdot\|$

\[
\|\text{diag} (\alpha_1 - \beta_1, \ldots, \alpha_p - \beta_p)\| \leq \|A - B\|
\]

(4.3.5)

- **Schmidt-Mirsky theorem:** Let

$\rightarrow A \in E \otimes F \simeq \mathcal{L}(F, E)$

$\rightarrow (\sigma_1, \ldots, \sigma_p)$ the singular values of $A$ in non increasing order

$\rightarrow B \in E \otimes F$ with rank $B = r \leq p$

$\rightarrow \Phi$ a Symmetric Gauge Function with $\|\cdot\|_\Phi$ as associated unitarily invariant norm

Then

\[
\|A - B\|_\Phi \geq \Phi(0, \ldots, 0, \sigma_{r+1}, \ldots, \alpha_p)
\]

(4.3.6)

Moreover, if $A = U \Sigma V^*$ is the SVD of $A$, the equality is reached for matrix $A_r$ defined as

\[
A_r = U \Sigma_r V^*
\]

(4.3.7)

where $\Sigma_r$ is the diagonal matrix obtained from $\Sigma$ be setting to 0 all singular values beyond $r$.

Then, $A_r$ is the best rank $r$ approximation of $A$ for norm $\|\cdot\|_\Phi$ as well.

**Notes and references:** The link between UIN and SGF has been shown in J. von Neumann (1937), Some Matrix Inequalities and Metrization of Matrix Spaces, *Tomsk Univ. Rev.*, 1286-300. The extension of PCA to UIN and SGF is nicely presented with many references in [Cla87]. See as well [Sch60] for an algebraic survey.

5 **PCA with Instrumental Variables**

Here, we use tensor notation for PCA. For sake of clarity for those readers not familiar with those notations, we set the problem with classical notations like $xy^T$ for a rank one matrices as a starting point. Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$. We can accept as a definition that $\otimes$ is a bilinear form

$\mathbb{R}^n \times \mathbb{R}^p \longrightarrow \mathbb{R}^{n \times p}$

\[
(x, y) \longmapsto x \otimes y
\]

defined by

\[
x \otimes y := xy^T
\]

which can be illustrated as
If \( x = (x_1, \ldots, x_n) \in \mathbb{R}^n \) and \( y = (y_1, \ldots, y_p) \in \mathbb{R}^p \), then \( x \otimes y \in \mathbb{R}^{n \times p} \) and 
\[
(x \otimes y)_{ij} = x_i y_j
\]

- Let \( A \in \mathbb{R}^{n \times p} \). A rank-\( r \) best approximation of \( A \) can be written as
  \[
  A_r = \sum_{a=1}^{r} y_a v_a^T, \quad y_a = A v_a
  \]
or, equivalently
  \[
  A_r = \sum_{a} y_a \otimes v_a, \quad \text{with } \begin{cases} 
    y_a & \in \mathbb{R}^n \\
    v_a & \in \mathbb{R}^p 
  \end{cases}
  \]

PCA with instrumental variables (PCAiv) is setting some constraints on \((v_a)_a\) or \((y_a)_a\), i.e. that
they live in given subspaces respectively \( F \subset \mathbb{R}^p \) and \( E \subset \mathbb{R}^n \).

Usually, those spaces are given as spanned by a set of vectors in respectively \( \mathbb{R}^p \) (for \( F \), constraint
on \( v \)) and \( \mathbb{R}^n \) (for \( E \), constraints on \( y \)). A classical situation is when there is no constraint on \( v \),
and only on \( y \), and \( E \) is spanned by a \( n \times q \) matrix denoted \( B \).

5.1 Setting the problem

Let us suppose that \( \dim E = m \) and \( \dim F = q \). PCAiv can be stated as

Technically, it is possible to specify this problem with basis of \( E \) and \( F \) having been selected. We assume here that they are orthonormal. If they are not orthonormal, it is possible to build
orthonormal basis by QR decomposition, or Gram-Schmidt orthogonalization procedure.
5.2 Solving the problem

If $E \subset \mathbb{R}^n$ and $F \subset \mathbb{R}^p$, then $E \otimes F \subset \mathbb{R}^n \otimes \mathbb{R}^p$, where $\subset$ means "is a vector subspace of".

- Before stating the main result, we need to recall some elementary results on linear projectors.
  Let $E \subset \mathbb{R}^n$. Then, the projector on $E$ is denoted $P_E$. Let $U = (u_1, \ldots, u_m)$ be an orthonormal basis of $E$. Then
  $$P_E = UU^T$$  \hfill (5.2.1)

  Indeed, we have
  $$P_E x = \sum_i (u_i, x) u_i$$
  $$= \sum_i (u_i \otimes u_i) x$$
  $$= \left(\sum_i u_i \otimes u_i\right) x$$

  Then
  $$P_E = \sum_i u_i \otimes u_i = UU^T$$  \hfill (5.2.2)

- Let $P$ be the projector on $E$, $P_E$ be the projector on $F$, and $A \in \mathbb{R}^{n \times p}$. Then, the projector $P_{E \otimes F}$ of $E \otimes F$ is defined by
  $$P_{E \otimes F} A = UU^T AVV^T$$  \hfill (5.2.3)

Main result: Let $A_{E \otimes F}$ be the projection of $A$ on $E \otimes F$. Then, the solution $A_r$ of $\text{PCAiv}$ is the PCA of $A_{E \otimes F}$.

Proof. Let us denote by $P$ the projection from $\mathbb{R}^n \otimes \mathbb{R}^p$ on $E \otimes F$, and by $P^\perp$ the projection on $(E \otimes F)^\perp$. Let us recall that $A_r = \sum_j y_j \otimes v_j$. We have $I = P + P^\perp$ and

- $\|A - A_r\|^2 = \| (P + P^\perp)(A - A_r) \|^2$ as $I = P + P^\perp$
  $$= \|P(A - A_r) + P^\perp(A - A_r)\|^2$$
  $$= \|P(A - A_r)\|^2 + \|P^\perp(A - A_r)\|^2$$ by Pythagore
  $$= \|P(A - A_r)\|^2 + \|P^\perp(A)\|^2$$ as $P^\perp A_r = 0$
  $$= \|P(A) - A_r\|^2 + \|P^\perp(A)\|^2$$ as $P A_r = A_r$

Let us recall that $A, E, F$ are fixed, hence so are $P, P^\perp$, and $A_r$ only can vary. Hence $\|A - A_r\|^2$ is minimum when $\|P(A) - A_r\|^2$ is minimum, and $A_r$ is the PCA of $P(A) = A_{E \otimes F}$.

- This leads to a solution for $\text{PCAiv}$. Let $(u_1, \ldots, u_m)$ be an orthonormal basis for $E \subset \mathbb{R}^n$, and $(w_1, \ldots, w_q)$ be for $F \subset \mathbb{R}^p$, with $m < n$ and $q < p$. Let $U$ be the $n \times m$ matrix with columns $u_j$ and $W$ be the $p \times q$ matrix with columns $w_j$. Then, the orthogonal projector $P_r$ from $\mathbb{R}^p$ to $F$ is given by matrix
  $$P_r = WW^T$$

  and the orthogonal projector from $\mathbb{R}^n$ to $E$ is given by matrix
  $$P_E = UU^T$$
Let $A \in \mathbb{R}^{n \times p}$. Then

$$P(A) = UU^\top AWW^\top$$

(5.2.4)

Hence the algorithm

**Algorithm 6 Pseudocode for PCAiv$(A,U,V)$**

1. **input**: $A \in \mathbb{R}^{n \times p}$
2. **input**: $E = \text{span } U \subset \mathbb{R}^n$, $U$ orthonormal
3. **input**: $F = \text{span } W \subset \mathbb{R}^p$, $W$ orthonormal
4. **input** $r < p$
5. **compute** $R = UU^\top$
6. **compute** $S = WW^\top$
7. **compute** $M = RAS$
8. **compute** $(Y,\Lambda,V) = \text{pca\_core}(M)$
9. **return** $Y,\Lambda,V$

**Notes and references:** Apparently, the term "PCA with Instrumental Variables" appeared first in [Rao64]. It has been studied with a double set of constraints in [Sab84].

### 5.3 Interpretation of PCAiv

PCAiv is a two steps procedures:

1. project the variables into the space spanned by the instrumental variables
2. run the PCA of the projected variables

This can be sketched by the following diagram:

$$A \xrightarrow{P} RAS \xrightarrow{\text{PCA\_iv}} Y,\Lambda,V$$

There is an estimate of the quality of each step.

- For the PCA, classical estimators of the quality of the PCA can be used. Let us denote

$$\|Y_r\| = \rho_r\|RAS\|$$

(5.3.1)

i.e. the quality of the PCA is denoted by $\rho_r$ at rank $r$.

- The quality of the projection can be quantified by

$$\|RAS\| = \theta\|A\|$$

(5.3.2)
(θ is the cosine of the angle between $A$ and $RAS$).

- We then have

$$
\|Y_r\| = \rho_r \theta \|A\|
$$

(5.3.3)

and the quality of the PCA can be poor for two reasons:

→ the quality $\theta$ of the projection is poor
→ the quality $\rho_r$ of the PCA at rank $r$ is poor.

- It is essential to distinguish between the quality of the projection and the quality of the PCA. Let us see it on an example. We assume that the matrix $A$ is built as a low rank matrix plus some important noise. It can be expected that the noise is poorly projected (there is no specific subspace where the noise is better represented), whereas there is some specific low dimensional subspace where the low rank component of $A$ is well represented. Then, projection will filter out the noise, whereas the PCA of the projected matrix will find the low rank property of the structure of $A$. $\theta$ will be low, but $\rho_r$ close to 1. Even if the quality $\rho_r \theta$ is poor because of the poor quality $\theta$ of the projection, PCA is a success as it has filtered out the noise and exhibited the low rank of the data set.

5.4 Non orthonormal basis

In section 5.1, the subspaces $E$ and $F$ are given by their basis $U$ and $V$ respectively. Let us now suppose that $E$ and $F$ are respectively spanned by the columns of $U$ and $W$, which are no longer assumed to be orthonormal. Then

$$
R = U(U^TU)^{-1}U^T, \quad S = W(W^TW)^{-1}W^T
$$

and equation (5.2.3) reads

$$
M = RAS
= U(U^TU)^{-1}U^TAW(W^TW)^{-1}W^T
$$

- If there are no constraints on the components $(y_i)_i$, which can be set by selecting $U = I_n$. Then

$$
M = AW(W^TW)^{-1}W^T
$$

Symetrically, the absence of constrains on $(x_i)_i$ can be set by selecting $W = I_p$, and

$$
M = U(U^TU)^{-1}U^TA
$$

- There is another way to run PCA when the basis $U$ spanning $E$ and $W$ spanning $F$ are given as non orthogonal:

1. build an orthogonal basis $U'$ of $E$ and an orthogonal basis $V'$ of $F$.
2. build the projectors $R = U'U'^T$ and $S = WW'^T$

Building $U'$ and $W'$ can be achieved by QR decomposition of $U$ and $V$. 

Inria
6 PCA with metrics on rows and columns

A matrix $A \in \mathbb{R}^{n \times p}$ can be considered as an element of space $\mathbb{R}^n \otimes \mathbb{R}^p$. This space is implicitly embedded with the canonical inner product

$$\forall A, B \in \mathbb{R}^{n \times p}, \quad \langle A, B \rangle = \sum_{i,j} \alpha_{ij} \beta_{ij} = \text{Tr} A^T B = \text{Tr} B^T A$$

Any $np \times np$ symmetric definite positive matrix defines an inner product on $\mathbb{R}^{n \times p}$, embedded with a Euclidean structure. We restrict ourselves here to inner products which establish a link with possible inner products in $\mathbb{R}^n$ and $\mathbb{R}^p$.

6.1 Metrics and weights on row and column spaces

Let us define first an inner product in $\mathbb{R}^p$, by a SDP matrix $P$, i.e.

$$\langle x, y \rangle_p = \langle x, Py \rangle$$ (6.1.1)

As $P$ is symmetric, we have $P = P^T$, and $\langle x, y \rangle_p = \langle Px, y \rangle$ as well. This means, component-wise in a given basis, that

$$\langle x, y \rangle_p = \sum_{i,j=1}^p p_{ij} x_i y_j, \quad p_{ji} = p_{ij}$$ (6.1.2)

if $x = (x_i)_i$, $y = (y_j)_j$ and $P = (p_{ij})_{i,j}$. If $P = I_p$, canonical inner product is recovered, as $p_{ij} = \delta_{ij}$. A case worth being studied in details is when $P$ is diagonal, i.e. $P = \text{diag} w$ with $w = (w_1, \ldots, w_p)$, or

$$p_{ij} = \begin{cases} w_i & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

In such a case

$$\langle x, y \rangle_p = \sum_{i=1}^p w_i x_i y_i$$ (6.1.3)

- As $P$ is SDP, there exists a unique SDP matrix $Q$ such that $P = Q^2$. Then

$$\langle x, y \rangle_p = \langle Qx, Qy \rangle$$ (6.1.4)

and

$$\|x\|_p = \|Qx\|$$ (6.1.5)

One may wonder whether the metric should be given by $P$ or by $Q$. It is tempting to give it by $Q$, because $Q$ establishes an isometry $\iota_Q$ between $(\mathbb{R}^p, Q)$ and $(\mathbb{R}^p, I)$ by

$$(\mathbb{R}^p, Q) \xrightarrow{\iota_Q} (\mathbb{R}^p, I)$$

$$x \mapsto Qx$$

RR n° 9488
as
\[ \|x\|_Q = \|Qx\| = \|\iota_Q x\| \]

However, in data analysis, it is classical to use weights, i.e., define metrics with diagonal matrices. Let
\[ w = (w_1, \ldots, w_p) \in \mathbb{R}^{p^+} \]

Then, a distance between \( x, y \in \mathbb{R}^p \) with weights \( w \) is given by
\[ d_w(x, y) = \|x - y\|_w = \sqrt{\sum_i w_i(x_i - y_i)^2} \]

This is consistent with an isometry defined by
\[ Q = \text{diag} \left( \sqrt{w_1}, \ldots, \sqrt{w_p} \right) \]
as
\[ d_w(x, y) = \sqrt{\sum_i \left( \sqrt{w_i} x_i - \sqrt{w_i} y_i \right)^2} \]

It is customary to define the weights by \( w \), which are the diagonal elements of \( P \), and not \( \sqrt{w} \), hence the metric by \( P \), hence denote \( \langle x, y \rangle_P \). We will use \( P \) or \( Q \) indifferently, knowing that \( P = Q^2 \).

\[ \bullet \quad \text{A metric can be defined in the same way in } \mathbb{R}^n, \text{ the column space. It is given by a SDP matrix } N \in \mathbb{R}^{n \times n}. \text{ We denote } N = M^2, \text{ and } \]
\[ \langle y, y' \rangle_n = \langle My, My' \rangle, \quad \|y\|_n = \|My\| \]

\[ \bullet \quad \text{Let } P \text{ define a metric on } \mathbb{R}^p \text{ and } N \text{ a metric on } \mathbb{R}^n. \text{ An inner product on } \mathbb{R}^{n \times p} \text{ will be defined, which is canonically associated to } N \text{ and } P \text{ and such that the map} \]
\[ (\mathbb{R}^{n \times p}, T) \rightarrow (\mathbb{R}^{n \times p}, \|\|) \]
is an isometry. Therefore, let \( a, x \in \mathbb{R}^n \) and \( b, y \in \mathbb{R}^p \). Let \( T = Z^2 \in \mathbb{R}^{np \times np} \) be an inner product on \( \mathbb{R}^n \otimes \mathbb{R}^p \). We wish to have on elementary (= rank one) matrices
\[ \langle a \otimes b, x \otimes y \rangle_T = \langle a, x \rangle_n \langle b, y \rangle_p \]
with
\[ \langle a \otimes b, x \otimes y \rangle_T = \langle Z(a \otimes b), Z(x \otimes y) \rangle \]

As
\[ \langle a \otimes b, x \otimes y \rangle_T = \langle Ma, Mx \rangle(Qb, Qy) \]
\[ = \langle Ma \otimes Qb, Mx \otimes Qy \rangle \]

\[ (6.1.6) \]

(6.1.6) is fulfilled by selecting
\[ Z(a \otimes b) = Ma \otimes Qb \]
\[ = M(a \otimes b)Q \]

and, for any matrix \( A \in \mathbb{R}^{n \times p} \)
\[ ZA = MAQ \]

\[ (6.1.9) \]
by linearity.

- So, we have

\[ \|A\|_T = \|MAQ\| \]  \hspace{1cm} (6.1.10)

This permits to solve PCA of a matrix \( A \in \mathbb{R}^{n \times p} \) with metrics \( N \) on \( \mathbb{R}^n \) and \( P \) on \( \mathbb{R}^p \) by transporting the problem by the isometry \( \iota(A) = MAQ \) to PCA of the image \( \iota(A) \) and transporting back the solution into initial space by the inverse \( \iota^{-1} \) of the isometry.

- **Remark:** We might stop this section here by saying that all what has been said about PCA makes no assumption about the inner product defining a Euclidean structure in \( \mathbb{R}^p, \mathbb{R}^n \) or \( \mathbb{R}^n \otimes \mathbb{R}^p = \mathbb{R}^{n \times p} \), and anything can be transported by the isometry mentioned above, period. But it may be not useless to develop this in more details. Even if in such a compact form it is exact and provides all needed information and procedures.

### 6.2 Setting the problem

Here, we use the algebraic approach of PCA

| Given | a matrix \( A \in \mathbb{R}^p \otimes \mathbb{R}^n \) |
|-------|-----------------------------------------------|
|       | a rank \( r < p \) |
| find  | a matrix \( A_r \) of rank \( r \) |
| such that | \( \|A - A_r\| \) is minimum |

to set the problem of PCA with metrics on rows and columns as

| Given | a matrix \( A \in \mathbb{R}^{n \times p} \) |
|-------|-----------------------------------------------|
|       | an inner product in \( \mathbb{R}^n \) defined by \( N \in \mathbb{R}^{n \times n} \) |
|       | an inner product in \( \mathbb{R}^p \) defined by \( P \in \mathbb{R}^{p \times p} \) |
|       | with a rank |
| Define | the inner product \( T \) on \( \mathbb{R}^{n \times p} \) associated to \( (N, P) \) |
|       | \( T = Z^2 \) |
|       | \( ZA = MAQ \) |
| Find  | a matrix \( A_r \in \mathbb{R}^{n \times p} \) |
| with  | rank \( A_r = r \) |
| such that | \( \|A - A_r\|_T \) minimal |
6.3 Solving the problem
We first give a direct solution, without using the associated isometry.

- A matrix of rank \( r \) can be written as
  \[
  A_r = \sum_{i=1}^{r} y_i \otimes v_i,
  \]
  \[
  \begin{cases}
  \|v_i\| = 1 \\
  y_i = A_r v_i
  \end{cases}
  \]
  Then
  \[
  \|A - A_r\| = \left\|A - \sum y_i \otimes v_i\right\|
  \]

- Let us select the inner product \( T = N \times P \) on \( \mathbb{R}^{n \times p} \) as
  \[
  \langle A, B \rangle_{N, P} = \langle MAQ, MBQ \rangle
  \]
  We recall that \( M(y \otimes v)Q = My \otimes Q^Tv = My \otimes Qv \) as \( Q \) is symmetric. Then
  \[
  \begin{align*}
  \|A - A_r\|_{N, P} &= \left\|M \left(A_r - \sum y_i \otimes v_i\right)Q\right\| \\
  &= \left\|MAQ - \sum M y_i \otimes Qv_i\right\|
  \end{align*}
  \]

- Let us denote
  \[
  \begin{cases}
  My_i = w_i \\
  Qv_i = x_i
  \end{cases}
  \]
  As \( \|v_i\|_p = 1 \) and \( \|v_i\|_p = \|Qv_i\| = v_i \), we have \( \|x_i\| = 1 \). Then, this can be written as
  \[
  \text{find } (w_i)_i, (x_i)_i \\
  \text{with } \|x_i\| = 1 \\
  \text{such that } \left\|MAQ - \sum w_i \otimes x_i\right\| \text{ minimal}
  \]
  Then, \( \{(w_i)_i, (x_i)_i\} \) are the solution of the PCA of \( MAQ \), and
  \[
  \begin{cases}
  v_i = Q^{-1}x_i \\
  y_i = M^{-1}w_i
  \end{cases}
  \]
  Hence the algorithm:
Algorithm 7 PCA of a matrix with double metrics: \texttt{PCA\_MET}(A,M,Q)

1: \textbf{input} $A \in \mathbb{R}^{n \times p}$; $M \in \mathbb{R}^{p \times p}$, SDP ; $Q \in \mathbb{R}^{n \times n}$, SDP
2: \textbf{compute} $B = MAQ$
3: \textbf{compute} $W, \Lambda, X = \texttt{PCA\_CORE}(B)$
4: \textbf{compute} $Y = M^{-1}W$
5: \textbf{compute} $V = Q^{-1}X$
6: \textbf{return} $Y, \Lambda, V$

6.4 Isometry

This result can be derived without calculation by transportation of \texttt{PCA} by isometry. Let $\mathbb{R}^n$ (resp. $\mathbb{R}^p$) be embedded with a Euclidean structure defined by SDP matrix $N = M^2$ (resp. $P = Q^2$). Then, the maps

$$(\mathbb{R}^n, N) \longrightarrow (\mathbb{R}^n, I_n)$$

$$(\mathbb{R}^p, P) \longrightarrow (\mathbb{R}^p, I_p)$$

and

$$(y) \longrightarrow My$$

$$(v) \longrightarrow Qv$$

are isometries, as

$$\begin{align*}
\langle Qv, Qv' \rangle &= \langle v, v' \rangle_p \\
\langle My, My' \rangle &= \langle y, y' \rangle_n
\end{align*}$$

This induces an isometry on $\mathbb{R}^n \otimes \mathbb{R}^p$ by

$$\psi : (\mathbb{R}^n \otimes \mathbb{R}^p, N \otimes P) \longrightarrow (\mathbb{R}^n \otimes \mathbb{R}^p, I_n \otimes I_p)$$

$$A \longrightarrow MAQ$$

as

$$\langle MAQ, MBQ \rangle = \langle A, B \rangle_{N \otimes P}$$

\texttt{PCA} of $A$ with metric $P$ on $\mathbb{R}^p$ and $N$ on $\mathbb{R}^n$ is finding the best rank $r$ approximation of a matrix $A$, i.e. finding $(y_j \otimes v_j)_{1 \leq j \leq r}$ such that

$$\Delta = \left\| A - \sum_{j=1}^{r} y_j \otimes v_j \right\|_{N \otimes P}$$

is minimum. Then,

$$\Delta_\psi = \left\| \psi(A) - \psi \left( \sum_{j} y_j \otimes v_j \right) \right\|$$

is minimum ($\Delta_\psi = \Delta$ because $\psi$ is an isometry). We have

$$\psi(A) - \psi \left( \sum_{j} y_j \otimes v_j \right) = MAQ - M \left( \sum_{j} y_j \otimes v_j \right) Q$$

$$= MAQ - \sum_{j} My_j \otimes Qv_j$$

RR n° 9488
Then, \( \sum_{j=1}^{r} M y_j \otimes Q v_j \) with \( \|Qv_j\| = 1 \) \( \forall j \) is the best rank \( r \) approximation of \( MAQ \) which can be solved by a PCA of \( MAQ \). Let \( \sum_{j} w_j \otimes x_j \) be the best rank \( r \) approximation of \( MAQ \). Then, by applying isometry \( \psi^{-1} \), \( \sum_{j} M^{-1} w_j \otimes Q^{-1} x_j \) is the best rank \( k \) approximation of \( A \) for metric defined by \( N \otimes P \), and the solution is \( (Y, V, \Lambda) \) with \( Y = M^{-1} W \) and \( V = Q^{-1} X \).

### 6.5 Interpretation and plotting

- A common situation is when metrics are given as weights on the columns only. Then, \( M = I_n \) and \( MAQ = AQ \)

Hence

\[
B^T B = (MAQ)^\top (MAQ) = (AQ)^\top (AQ) = Q^\top A^\top AQ
\]  
(6.5.1)

Similarly, if the metrics are weights on the rows only, \( Q = I_p \) and \( MAQ = MA \)

Hence

\[
B^T B = (MAQ)^\top (MAQ) = (MA)^\top (MA) = A^\top MA \quad (6.5.2)
\]

If metrics are given on both rows and columns, we have

\[
B^T B = QA^\top NAQ \quad (6.5.3)
\]

- A remark about the calculation: Principal components \( (y_i)_i \) and principal axis \( (v_i)_i \) are solution of

\[
\begin{cases}
B = MAQ \\
B^T Bw_i = \lambda_i w_i \\
x_i = Bw_i \\
v_i = Q^{-1} w_i \\
y_i = M^{-1} x_i
\end{cases} \quad (6.5.4)
\]

with \( x_i, y_i \in \mathbb{R}^n \), \( v_i, w_i \in \mathbb{R}^p \)

We have

\[
B^T B = QA^\top NAQ \\
B^T Bw_i = \lambda_i w_i \\
w_i = Qv_i
\]

\[
\implies QA^\top NAQ v_i = \lambda_i Q v_i 
\]  
(6.5.5)

and, as \( Q \) is invertible

\[
A^\top NAQ v_i = \lambda_i v_i 
\]  
(6.5.6)

This might lead to a way of computing the principal axis \( (v_i)_i \) directly without computing the \( (w_i)_i \) before. However, the matrix \( B^T B \) is symmetric, whereas matrix \( A^\top NA^p \) is not. It is known that numerical computation of eigenvectors and eigenvalues of symmetric matrices is more
accurate and robust than of non-symmetric matrices. Hence, it is recommended to compute first $(w_i)_i$ as solutions of $B^tBw_i = \lambda_iw_i$ and then the principal axis as $v_i = Q^{-1}w_i$.

- **Centering the cloud:** As for PCA, it is advised to center the cloud before analysing it when there are some weights on rows. Let us recall that if each point $a_i \in \mathbb{R}^p$ is given a weight $w_i$, the barycenter $g \in \mathbb{R}^p$ is given by
  \[
  g = \frac{1}{w} \sum_i w_i a_i \quad \text{or} \quad g = \frac{1}{\sum_i w_i} \sum_i w_i a_i.
  \]
  The centered cloud is the cloud with points $\tilde{a}_i = a_i - g$.

- **Geometric approach: attached point cloud:** Let $A$ be the point cloud of $n$ points in $\mathbb{R}^p$ attached to matrix $A$. Distances between points do not reflect the distances induced by the inner products $(M,Q)$. Let us denote by $B$ the point cloud in $\mathbb{R}^p$ attached to matrix $B = MAQ$. Points $b_i, b_k$ have as coordinates respectively the rows $i$ and $k$ of $B$. If $M, Q$ are diagonal matrices with weights $(\sqrt{\nu}_i, \sqrt{\pi}_j)$ respectively, then
  \[
  MAQ = \begin{bmatrix} \sqrt{\nu}_i \pi_j \alpha_{ij} \end{bmatrix}_{i,j}
  \]
  and, in $\mathbb{R}^p$
  \[
  d^2(b_i, b_k) = \sum_j (\sqrt{\nu}_i \pi_j \alpha_{ij} - \sqrt{\nu}_k \pi_j \alpha_{kj})^2
  = \sum_j \pi_j (\sqrt{\nu}_i \alpha_{ij} - \sqrt{\nu}_k \alpha_{kj})^2
  \]
  This is the distance between points of the point cloud in $\mathbb{R}^p$ attached to matrix $MA$ with the inner product in $\mathbb{R}^p$ defined by weight matrix $P$. So, PCA of matrix $A \in \mathbb{R}^{n \times p}$ with inner product defined by $N$ in $\mathbb{R}^n$ and $P$ in $\mathbb{R}^p$ is PCA of point cloud $A_d$ in $\mathbb{R}^p$ attached to matrix $MA$ with inner product defined by $P$ in $\mathbb{R}^p$ for computing distances. This will be useful for Correspondance Analysis (see section 7).

- **Scaled PCA:** A straightforward and standard application of PCA with metrics is scaled PCA. Let $A \in \mathbb{R}^{n \times p}$ be a columnwise centered matrix, i.e.
  \[
  \sum_i a_i = 0
  \]
The variances (or norms) of columns of $A$ can vary significantly. In such a case, the variance/covariance matrix $\Sigma = A^t A$ can be dominated by rows and columns corresponding to the variable with largest variance. Scaled PCA is clipping this uninteresting result off, by giving equal weights to each variable. The technical trick is to equalize variances between columns, by dividing each column $j$ by its standard deviation (or norm). If $a_{\bullet j} \in \mathbb{R}^n$ is column $j$ of $A$, this reads

$$a_{\bullet j} \longrightarrow a'_{\bullet j} = \frac{a_{\bullet j}}{\|a_{\bullet j}\|}$$

(6.5.12)

Hence

$$\|a'_{\bullet j}\| = 1$$

(6.5.13)

This can be read as a PCA with inner product $N = I_n$ in $\mathbb{R}^n$ and $P = \text{diag} \left( \frac{1}{\|a_{\bullet j}\|} \right)$ in $\mathbb{R}^p$. So $MAQ = A'$, and PCA of $A'$ is run.

**Notes and references:** The problem (and solution) of PCA with weights on rows and columns can be found in [Rao64] or [Gre84]. It is presented in [Jol02, sect. 14.2]. The algebraic approach with generalization to metrics in Euclidean spaces has been proposed as a general method with the notion of duality diagram in [CP79] which has been at the root of many works (see [PCY79]). The formalism presented here can be found in [Fra92, Fra21].

### 6.6 PCA with metrics and instrumental variables

Those methods, PCAmet and PCAiv can be associated like pieces of puzzle to build a chain of treatments.

- Let us recall (see section 5) that PCAiv is running the PCA of $A$ with constraints on principal axis and components which must belong to subspaces of respectively $\mathbb{R}^p$ and $\mathbb{R}^n$:

$$\begin{cases} y_j \in E \subset \mathbb{R}^n \\ v_j \in F \subset \mathbb{R}^p \end{cases}$$

(6.6.1)

If $U$ (resp. $V$) is an orthonormal matrix with a basis of $E$ (resp. $F$) as column vectors, this is done by building the projectors

$$\mathbb{R}^n \xrightarrow{R=UU^t} E, \quad \mathbb{R}^p \xrightarrow{S=VV^t} F$$

(6.6.2)

and running the PCA of $A' = RAS$, the projection of $A$ on $E \otimes F$.

- If $\mathbb{R}^n$ and $\mathbb{R}^p$ are endowed with metrics given by $N$ and $P$ respectively, running the PCA of $A'$ with those metrics is running the PCA of $MAQ$ with $M = N^{1/2}$ and $Q = P^{1/2}$. However, the projectors $R$ and $S$ depend on the metrics $N$ and $P$.

- Let us write the projector on $F \subset \mathbb{R}^p$ with the inner product defined by $P$ first. Let $x \in \mathbb{R}^p$ and $v \in F$ with $\|v\|_v = 1$. The projection $x'$ of $x$ on $F = \mathbb{R}v$ is given by

$$x' = \langle x, v \rangle_v v = \langle x, P v \rangle v = \langle P v, x \rangle v = \langle v \otimes P v \rangle x$$

(6.6.3)
Hence, the projector on $Rv$ with the inner product defined by $P$ is $v \otimes Pv$. If $F = \text{span}(v_1, \ldots, v_r)$, we have
\[
x' = \sum_a \langle x, Pv_a \rangle v_a
= \sum_a (v_a \otimes Pv_a)x
\] (6.6.4)
and the projector is
\[
S = \sum_a v_a \otimes Pv_a
= V(PV)^\tau
= VV^\tau P
= PVV^\tau
\] (6.6.5)
if $V \in \mathbb{R}^{p \times r}$ is the matrix with $v_a$ in column $a$. The last equality comes from the observation that $P$ and $VV^\tau$ are symmetric, hence $(VV^\tau)P$ is symmetric and $VV^\tau P = (VV^\tau)P^\tau = PVV^\tau$.

- Similarly, we have
\[
R = UU^\tau N
\] (6.6.6)
and
\[
A' = RAS
= UU^\tau NAPVV^\tau
\] (6.6.7)

- Let us now write the PCA of $A'$ with inner products defined by $N$ on $\mathbb{R}^n$ and $P$ on $\mathbb{R}^p$. It is the PCA of

$A'' = MA'Q$, with $N = M^2$, $P = Q^2$

7 Correspondence Analysis

A remarkable application of this approach is a presentation of Correspondence Analysis as the analysis of a contingency table with metrics associated to its margins.

Let us adopt here some standard notations for contingency tables. A contingency table $T$ is a table of counts of $n$ items allocated to categories of two variables. Indices of the values of the first variable are denoted $i$, and of the second variable $j$. The value $n_{ij}$ in row $i$ and column $j$ of $T$ is the number of items in category $i$ for the first variable and $j$ for the second. It is standard to denote that $i \in [1, I]$ and $j \in [1, J]$. Then
\[
T \in \mathbb{R}^{I \times J} \simeq \mathbb{R}^I \otimes \mathbb{R}^J
\]

7.1 Link with $\chi^2$ distance

We first establish a link between the norm of a contingency table with metrics associated to margins on rows and columns on one hand and the $\chi^2$ of the table on the other.

- Let $T$ be a contingency table of two discrete variables observed on $n$ individuals, with $T_{ij}$ being the number of individuals with observation $i$ for first variable and $j$ for the second. Then $T \in \mathbb{R}^{I \times J}$ if first variable has $I$ values and second $J$. 
Let us denote
\[ A = \frac{T}{T_{++}} \quad T_{++} = \sum_{i,j} T_{ij} \] (7.1.1)

The general term in A is denoted $\alpha_{ij}$ and we have
\[ A \in \mathbb{R}^{I \times J} \simeq \mathbb{R}^I \otimes \mathbb{R}^J \] (7.1.2)

Let us denote respectively by $r \in \mathbb{R}^I$ and $c \in \mathbb{R}^J$ the marginal sums of $A$ on rows and columns
\[
\begin{aligned}
    r_i &= \sum_j \alpha_{ij} \\
    c_j &= \sum_i \alpha_{ij}
\end{aligned}
\] (7.1.3)

Let us denote by $D_r$ and $D_c$ the square diagonal matrices with diagonal respectively $r$ and $c$:
\[ D_r = \text{diag } r, \quad D_c = \text{diag } c \] (7.1.4)

- In case of independence between both variables, the expectation for $A$ is
\[ \tilde{A} = r \otimes c \] (7.1.5)

Indeed, we have, ignoring the value of the other variable
\[ P(X_r = i) = r_i, \quad P(X_c = j) = c_j \]

Then, in case of independence
\[ P(X_r = i; X_c = j) = r_i c_j \]

Then, a first observation is that
\[ \chi^2(A) = \| A - \tilde{A} \|_{D_r^{-1} \otimes D_c^{-1}} \] (7.1.6)

**Proof.** Indeed, we have
\[
\begin{align*}
\chi^2(A) &= \sum_{i,j} \frac{(\alpha_{ij} - r_i c_j)^2}{r_i c_j} \\
&= \sum_{i,j} \left( \frac{\alpha_{ij} - r_i c_j}{\sqrt{r_i c_j}} \right)^2 \\
&= \sum_{i,j} \left( \frac{1}{\sqrt{r_i}}(\alpha_{ij} - r_i c_j) \frac{1}{\sqrt{c_j}} \right)^2 \\
&= \left\| D_r^{-1/2} (A - r \otimes c) D_c^{-1/2} \right\|^2 \\
&= \left\| A - r \otimes c \right\|^2_{D_r^{-1} \otimes D_c^{-1}}
\end{align*}
\] (7.1.7)
7.2 Description of the method

Then, Correspondence Analysis is a partition of the variance \( \| A - r \otimes c \|_{D_r^{-1} \otimes D_c^{-1}} \) concentrated on the first axis. It is henceforth a PCA of \( A - r \otimes c \) with metrics defined by \( D_r^{-1} \) on rows and \( D_c^{-1} \) on columns, i.e. with weights \( 1/r_i \) on row \( i \) and \( 1/c_j \) on column \( j \).

We have

\[
\begin{align*}
M &= \text{diag } 1/\sqrt{r_i} \\
Q &= \text{diag } 1/\sqrt{c_j}
\end{align*}
\]

Then

\[
A_{m,q} = M(A - r \otimes c)Q = \left[ \frac{\alpha_{ij} - r_i c_j}{\sqrt{r_i c_j}} \right]_{i,j}
\]  

(7.2.1)

This yields the following algorithm:

**Algorithm 8 Correspondence Analysis of a contingency table: COA(T)**

1. input \( T \in \mathbb{R}^{I \times J} \), a contingency table
2. compute \( A = T/T_{++} \), with \( T_{++} = \sum_{i,j} T_{ij} \)
3. compute \( r_i = \sum_j \alpha_{ij} \), \( D_r = \text{diag } r \)
4. compute \( c_j = \sum_i \alpha_{ij} \), \( D_c = \text{diag } c \)
5. compute \( M = \text{diag } (1/\sqrt{r_i}) \)
6. compute \( Q = \text{diag } (1/\sqrt{c_j}) \)
7. compute \( A_{m,q} = M(A - r \otimes c)Q = \left[ \frac{\alpha_{ij} - r_i c_j}{\sqrt{r_i c_j}} \right]_{i,j} \)
8. compute \( Z, X, \Lambda = \text{PCA\_CORE}(A_{m,q}) \)
9. compute \( Y_r = M^{-1}Z \)
10. compute \( Y_c = Q^{-1}X \)
11. return \( Y_r, Y_c, \Lambda \)
7.3 CoA and geometry of point clouds

Let us consider the point cloud \( X \) of \( I \) points in \( \mathbb{R}^j \) of coordinates

\[
X_{ij} = \frac{\alpha_{ij}}{\sqrt{r_i c_j}}
\]

(7.3.1)

in a Euclidean space with standard inner product (i.e. \( I \)). Then, if \( x_i, x_i' \) are two points in \( X \), we have

\[
d^2(x_i, x_i')^2 = \sum_j \left( \frac{\alpha_{ij}}{r_i c_j} - \frac{\alpha_{ij}'}{r_i' c_j} \right)^2
\]

(7.3.2)

It is standard to say that point cloud \( X = (x_i)_i \) is embedded with weights \( 1/c_j \) for column \( = \) variable \( j \) and metrics

7.4 Classical presentation

There is a classical presentation of Correspondance Analysis as an analysis of two point clouds associated to a contingency table, one for the rows and one for the columns (which play a symmetric role), each with weights and metrics. It may be slightly confusing, because a metric is defined in CA as weights in relevant space. This is to emphasize that two point clouds can be built: one for rows, and one for columns, and CA can be seen as a simultaneous analysis of both.

- Let us recall that if \( T \) is a contingency table, and \( A \) is built from \( T \) by dividing all elements by the total of the elements in \( T \):

\[
T \longrightarrow T_{++} = \sum_{i,j} T_{ij} \longrightarrow A: \alpha_{ij} = \frac{T_{ij}}{T_{++}}
\]

- Two point clouds are classically attached to \( A \)

  \( \rightarrow \) a cloud of row profiles, as \( I \) points \( x_i \) in \( \mathbb{R}^j \), with point \( x_i \) of coordinates

  \[
x_i = \left[ \frac{\alpha_{ij}}{r_i} \right]_j, \quad r_i = \sum_j \alpha_{ij}
\]

  (7.4.1)

  \( \rightarrow \) a cloud of column profiles, as \( J \) points \( x^T_j \) in \( \mathbb{R}^i \), with point \( x^T_j \) of coordinates

  \[
x^T_j = \left[ \frac{\alpha_{ij}}{c_j} \right]_i, \quad c_i = \sum_j \alpha_{ij}
\]

  (7.4.2)

- Then, metrics with diagonal matrices respectively \( D_{c}^{-1} \) for \( R^i \) and \( D_{r}^{-1} \) for \( R^i \) are selected. Hence, distances between points \( x_i \) and \( x'_j \) in \( \mathbb{R}^i \) are computed as

  \[
d^2(x_i, x'_j) = \sum_j \frac{1}{c_j} \left( \frac{\alpha_{ij}}{r_i} - \frac{\alpha_{ij}'}{r_i'} \right)^2
\]

(7.4.3)
and between points \( x_{jt} \) and \( x'_{jt} \) in \( \mathbb{R}^1 \) as

\[
d^2(x_{jt}, x'_{jt}) = \sum_i \frac{1}{r_i} \left( \frac{\alpha_{ij}(x_t - x_t')}{c_j - c_j'} \right)^2
\]

(7.4.4)

These weights tend to give an equal importance to all modalities of a variable, whatever their size. It is analogous to weighting by the inverse of the variance in scaled PCA. A further property often invoked is that, if two categories of a same variable have the same profile (say \( i \) and \( i' \) for which \( \forall j, \alpha_{ij}/r_i = \alpha_{ij'}/r_i' \)), then it is logical to lump them together into a single category, and this must not modify the distances between row profiles. It is then standard to define both point clouds as (see [Gre84, sect. 4.1], [LMF82, sect. IV.5.], [Sap90, sect. 10.1])

| Point cloud     | \( R: \) row profiles in \( \mathbb{R}^J \) | \( C: \) column profiles in \( \mathbb{R}^I \) |
|-----------------|-------------------------------------------|-------------------------------------------|
| Metric          | \( D_r^{-1} A \)                           | \( D_c^{-1} A' \)                        |
| Weights         | \( D_r^{-1} \)                             | \( D_c^{-1} \)                           |

The centroids of \( R \) and \( C \) are respectively \( r \) and \( c \). Then, the inertia \( I_r \) of centered point cloud \( R \), i.e. of \( R - 1_I \otimes c \) with weights \( r \) on rows and metric \( D_r^{-1} \) in \( \mathbb{R}^J \) is, (step by step ...)

\[
I_r^2 = \sum_i r_i \left\| (D_r^{-1} A)_i - c \right\|_P^2
= \sum_i r_i \left\| \frac{a_i}{r_i} - c \right\|_P^2
= \sum_i r_i \left( \frac{1}{c_j} \sum_j \left( \frac{a_{ij}}{r_i} - c_j \right)^2 \right)
= \sum r_i \left( \sum_j \left( \frac{a_{ij}}{r_i \sqrt{c_j}} - \frac{\alpha_{ij}}{\sqrt{c_j}} \right)^2 \right)
= \sum_{i,j} \left( \frac{a_{ij} - r_i c_j}{r_i \sqrt{c_j}} \right)^2
= \sum_{i,j} \left( \frac{a_{ij} - r_i c_j}{r_i \sqrt{c_j}} \right)^2
\]

(7.4.5)

\[
I_r = \sum_i r_i \left\| (D_r^{-1} A)_i - c \right\|_P = \sum_i \frac{a_i}{r_i}
= \sum r_i \left( \frac{1}{c_j} \sum_j \left( \frac{a_{ij}}{r_i} - c_j \right)^2 \right)
= P = \text{diag} \left( \frac{1}{c_j} \right)
= \sum_i \left( \sum_j \left( \frac{a_{ij}}{r_i \sqrt{c_j}} - \frac{\alpha_{ij}}{\sqrt{c_j}} \right)^2 \right)
= Q = \text{diag} \left( \frac{1}{\sqrt{c_j}} \right)
\]

- This can be done as well with notations presented in section 6.1. Let \( X \) be a \( n \times p \) matrix.
  The inertia \( I \) of the associated point cloud in \( \mathbb{R}^p \) is its norm \( \| X \| \). If \( \mathbb{R}^p \) is embedded with a Euclidean structure defined by matrix \( P = Q^2 \), then \( I = \| X \|_P = \| X Q \| \). If rows of \( X \) are given weights, defined by coordinates of vector \( w_r \in \mathbb{R}^n \), then it is equivalent to defining a metric by diagonal matrix \( N = \text{diag} w_r = M^2 \) in \( \mathbb{R}^n \), and \( I = \| X \|_N = \| M X \| \). If we have weights on rows and metric in \( \mathbb{R}^p \), we have \( I = \| X \|_{P \otimes N} = \| M X Q \| \). Let \( X = D_r^{-1} A - c \otimes 1_n \), \( P = D_c^{-1} \) hence
\( Q = D^{-1/2}_c \), and \( w_r = r \) hence \( M = D_r^{-1/2} \). Then, knowing that \( A(x \otimes y)B = Bx \otimes Ay \)

\[
I_n = \|MXQ\| \\
= \|D_r^{1/2}(D_r^{-1}A - c \otimes 1_n)D_c^{-1/2}\| \\
= \|D_r^{-1/2}AD_c^{-1/2} - D_r^{1/2}(c \otimes 1_n)D_c^{-1/2}\| \\
= \|D_r^{-1/2}AD_c^{-1/2} - D_c^{-1/2}c \otimes D_r^{1/2}1_n\| \\
= \|D_r^{-1/2}AD_c^{-1/2} - D_c^{-1/2}c \otimes D_r^{-1/2}r\| \\
= \|D_r^{-1/2}(A - c \otimes r)D_c^{-1/2}\|
\]

If \( I_c^2 \) is the inertia of centered point cloud \( C \) in \( \mathbb{R}^n \), i.e. of \( C - 1_p \otimes r \) with weights \( c \) on rows and metric \( D_r^{-1} \) in \( \mathbb{R}^n \), a similar calculation yields

\[
I_c^2 = \sum_{i,j} \left( \frac{a_{ij} - r_ic_j}{\sqrt{r_ir_j}} \right)^2 = I_n^2 \quad (7.4.6)
\]

Both inertia are equal.

Notes and references: Correspondence Analysis has a long history, and has been object of long debates, renaming and rediscoveries between different schools. Correspondence Analysis has been proposed first by Hirshfeld in 1935 (Hirschfeld, M. O. - 1935 - A connection between correlation and contingency - Proc. Camb. Phil. Soc., 31:520-524). It has been rediscovered by Guttman in 1959 (Guttman, L. - 1959 - Metricizing rank ordered and unordered data for a linear factor analysis. Sankhyā, 21:257-268). The link between CA and reciprocal averaging has been presented in [Hil74]. Correspondence Analysis has been rediscovered and studied independently by several researchers, as J.-P. Benzecri, in 1962 in the context of mathematical linguistics inspired by the works of Chomsky, J. de Leeuw in Netherlands and C. Hayashi in Japan. His type III Quantification methods, published in the 50’s (Hayashi, C. (1954). Multidimensional quantification with applications to analysis of social phenomena. Annals of the Institute of Statistical Mathematics, 5(2):121–143.), is equivalent to Correspondence Analysis. A historical background and synthesis is given in [TY85]. One of the early work in the French school is Cordier, B. - Sur l’analyse Factorielle des Correspondances. PhD, Rennes, 1965. This approach has been developed by Greenacre in Pretoria who has studied with J.-P. Benzecri [Gre84]. The algorithm presented here is the one presented in [NG87]. We have used as well [Sap90, chapter 10] for the geometric interpretation.

8 Canonical Correlation Analysis

Let us have two data sets as two sets of variables on the same set of items, as \( A, B \), with \( A \in \mathbb{R}^{n \times p} \) and \( B \in \mathbb{R}^{n \times q} \) We assume that \( p, q < n \). The set of columns of each matrix spans a subspace in \( \mathbb{R}^n \). If a column of \( A \) belongs to the space spanned by the columns of \( B \), than there exists a linear regression on the columns of \( B \) which explains this column of \( A \), and both sets of columns are correlated in \( \mathbb{R}^n \). Canonical Correlation Analysis (CCA) is about finding sets of vectors (= components) in the spaces spanned by the columns of each matrix with greatest correlation.
8.1 Stating the problem
We will denote by $u_a$ a vector in $\mathbb{R}^p$ and by $u_b$ in $\mathbb{R}^q$. A vector $y_a \in \text{span} A$ (resp. $y_n \in \text{span} B$) can be written as $Au_a$ (resp. $Bu_n$). The correlation between $y_a$ and $y_n$ is

$$\text{corr}(y_a, y_n) = \frac{\langle y_a, y_n \rangle}{\|y_a\| \|y_n\|}$$

Then, Canonical Correlation Analysis of $(A, B)$ for first canonical components can be stated as

Given $A \in \mathbb{R}^{n \times p}, B \in \mathbb{R}^{n \times q}$

Find $u_a \in \mathbb{R}^p, u_n \in \mathbb{R}^q$

such that $\frac{\langle Au_a, Bu_n \rangle}{\|Au_a\| \|Bu_n\|}$ maximal

As such the problem is difficult to solve. One reason is that $\|u_a\|$ and $\|u_n\|$ can take any non zero value (the correlation remains unchanged by a rescaling of $\|u_a\|$ or $\|u_n\|$). One could add a constraint like $\|u_a\| = \|u_n\| = 1$, but the problem still is difficult to solve. There is an equivalent formulation leading to easier calculations for the solution, by setting a constraint on $y_a = Au_a$ (resp. $y_n = Bu_n$):

Given $A \in \mathbb{R}^{n \times p}, B \in \mathbb{R}^{n \times q}$

Find $u_a \in \mathbb{R}^p, u_n \in \mathbb{R}^q$

with $\|Au_a\| = 1, \|Bu_n\| = 1$

such that $\langle Au_a, Bu_n \rangle$ is maximal

8.2 Solving the problem
This is an optimization problem with constraints, which can be solved by Lagrange multipliers. Let us recall that if

$$\mathbb{R}^n \xrightarrow{f,g} \mathbb{R}$$

an optimum of $f(x)$ under the constraint $g(x) = 0$ is obtained at some points $x$ satisfying

$$\nabla f - \lambda \nabla g = 0$$

where

$$\nabla f = \left( \frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_1} \right)$$

It remains to check that such a solution is a maximum (it can be a minimum or a saddle point).

- Here, the unknowns are $(u_a, u_n)$, the function $f$ is $f(u_a, u_n) = \langle Au_a, Bu_n \rangle$ and $g$ is $\|Au_a\| = \|Bu_n\| = 1$. One computes seperately the partial derivatives with respect to $u_a$ and $u_n$, denoting
them $\nabla u_a$ and $\nabla u_b$. One has
\[
\begin{align*}
\nabla u_a \langle Au_a, Bu_b \rangle &= A^T Bu_b \\
\nabla u_b \langle Au_a, Bu_b \rangle &= A^T Au_a \\
\n\nabla u_a \|Au_a\|^2 &= 2 A^T Au_a \\
\n\nabla u_b \|Bu_b\|^2 &= 2 B^T Bu_b
\end{align*}
\]
(8.2.1)

Then, the solution satisfies to
\[
\begin{align*}
\nabla u_a : A^T Bu_b &= \lambda A^T Au_a \\
\nabla u_b : B^T Au_a &= \mu B^T Bu_b
\end{align*}
\]
(8.2.2)

- We first show that $\lambda = \mu$.

**Proof.** Therefore, we observe that
\[
\begin{align*}
\langle A^T Bu_b, u_a \rangle &= \lambda \langle A^T Au_a, u_a \rangle \\
\langle B^T Au_a, u_b \rangle &= \mu \langle B^T Bu_b, u_b \rangle
\end{align*}
\]
and that
\[
\begin{align*}
\langle A^T Au_a, u_a \rangle &= \langle Au_a, Au_a \rangle = 1 \\
\langle B^T Bu_b, u_b \rangle &= \langle Bu_b, Bu_b \rangle = 1
\end{align*}
\]
Then
\[
\lambda = \langle A^T Bu_b, u_a \rangle = \langle B^T Au_a, u_b \rangle = \mu
\]

- Thus, equation (8.2.2) reads
\[
\begin{align*}
A^T Bu_b &= \lambda A^T Au_a \\
B^T Au_a &= \lambda B^T Bu_b
\end{align*}
\]
(8.2.3)

Multiplying leftwise the first equation by $(A^T A)^{-1}$ and the second by $(B^T B)^{-1}$ yields
\[
\begin{align*}
(A^T A)^{-1} A^T Bu_b &= \lambda u_a \\
(B^T B)^{-1} B^T Au_a &= \lambda u_b
\end{align*}
\]
(8.2.4)

and, having in mind that $Au_a = y_a$ and $Bu_b = y_n$
\[
\begin{align*}
A(A^T A)^{-1} A^T y_n &= \lambda y_a \\
B(B^T B)^{-1} B^T y_n &= \lambda y_n
\end{align*}
\]
(8.2.5)

- One recognizes in the l.h.s. of (8.2.5)
\[
\begin{align*}
A(A^T A)^{-1} A^T &= P_a \\
B(B^T B)^{-1} B^T &= P_n
\end{align*}
\]
i.e. the projectors on the spaces spanned by the columns of $A$ and of $B$ respectively. This leads to
\[
\begin{align*}
P_a y_n &= \lambda y_a \\
P_n y_a &= \lambda y_n
\end{align*}
\]
Linear Dimensionality Reduction

or

\[
\begin{align*}
\mathcal{P}_a \mathcal{P}_b y_a &= \lambda^2 y_a \\
\mathcal{P}_b \mathcal{P}_a y_a &= \lambda^2 y_a
\end{align*}
\] (8.2.6)

- **Interpretation:** The interpretation is quite natural. \(\text{span } A\) and \(\text{span } B\) are two vector subspaces in \(\mathbb{R}^n\) of dimension \(p\) and \(q\) respectively. Let us have \(y_a \in \text{span } A\). It is projected as \(y'_a \in \text{span } B\) by \(y'_a = \mathcal{P}_b y_a\). \(y'_b\) itself is projected as \(y''_b \in \text{span } A\) by \(y''_b = \mathcal{P}_a y'_b\). One has

\[
\text{span } A \xrightarrow{\mathcal{P}_a} \text{span } B \xrightarrow{\mathcal{P}_b} \text{span } A
\]

\[
y_a \quad \longmapsto \quad y'_a \quad \longmapsto \quad y''_a
\]

The same can be written for \(y_b\):

\[
\text{span } B \xrightarrow{\mathcal{P}_b} \text{span } A \xrightarrow{\mathcal{P}_a} \text{span } B
\]

\[
y_a \quad \longmapsto \quad y'_a \quad \longmapsto \quad y''_a
\]

Equation (8.2.6) says that when the correlation between \(y_a\) and \(y_b\) is maximal, then \(y_a\) (resp. \(y_b\)) and \(y''_a\) (resp. \(y''_b\)) are colinear, and eigenvectors of \(\mathcal{P}_b \mathcal{P}_a\) (resp. \(\mathcal{P}_a \mathcal{P}_b\)).

**Notes and references:** The computation of the solution in this section is classical, and has been borrowed from [LMF82, section IV.6.4].

### 8.3 Computation of the solution

This solution is geometrically speaking very intuitive, but it does not lead to the most efficient way to compute a solution. We show here how it is possible to derive the solution as the result of an EVD of a symmetric matrix. We start from

\[
\begin{align*}
\mathcal{P}_a \mathcal{P}_b y_a &= \lambda^2 y_a \\
\mathcal{P}_b \mathcal{P}_a y_a &= \lambda^2 y_a
\end{align*}
\]

Let us recall that

\[
\begin{align*}
\mathcal{P}_a &= (A^t A)^{-1} A^t \\
\mathcal{P}_b &= (B^t B)^{-1} B^t
\end{align*}
\]

Then

\[
\begin{align*}
A(A^t A)^{-1} A^t B (B^t B)^{-1} B^t y_a &= \lambda^2 y_a \\
B (B^t B)^{-1} B^t A (A^t A)^{-1} A^t y_a &= \lambda^2 y_a
\end{align*}
\] (8.3.1)

Without loss of generality, we assume that \(p \geq q\). Let us denote

\[
\begin{align*}
T &= A^t B \\
N &= (A^t A)^{-1} \\
P &= (B^t B)^{-1}
\end{align*}
\] (8.3.2)

Then

\[
\begin{align*}
ANTPB y_a &= \lambda^2 y_a \\
BPT^t NA^t y_a &= \lambda^2 y_a
\end{align*}
\]

Let us recall that

\[
y_a = A u_a, \quad y_b = B u_b
\]

RR n° 9488
Then
\[
\begin{aligned}
ANTPT^\top u_\lambda &= \lambda^2 A u_\lambda \\
BPT^\top NT u_n &= \lambda^2 B u_n
\end{aligned}
\]

We can "simplify" by \( A \) and \( B \) by left multiplication by \((A^\top A)^{-1}A^\top\) and \((B^\top B)^{-1}B^\top\). We have
\[
\begin{aligned}
NTPT^\top u_\lambda &= \lambda^2 u_\lambda \\
PT^\top NT u_n &= \lambda^2 u_n
\end{aligned}
\]  \hfill (8.3.3)

This reminds of the type of equation of a PCA with metrics (see equation (6.5.6)).

- Here, we show how the solution of CCA can be read as the solution of a PCA with metrics. Therefore, let us denote, as in section 6,

\[
M = N^{1/2}, \quad Q = P^{1/2}, \quad R = MTQ \in \mathbb{R}^{p \times q}
\]

Then, equation (8.3.3) reads
\[
\begin{aligned}
MRR^\top M^{-1} u_\lambda &= \lambda^2 u_\lambda \\
Q R^\top RQ^{-1} u_n &= \lambda^2 u_n
\end{aligned}
\]  \hfill (8.3.4)

Let us denote
\[
w_\lambda = M^{-1} u_\lambda, \quad w_n = Q^{-1} u_n
\]

Then, by left multiplication by \( M^{-1} \) of the first equation and by \( Q^{-1} \) of the second, we have
\[
\begin{aligned}
RR^\top w_\lambda &= \lambda^2 w_\lambda \\
R^\top R w_n &= \lambda^2 w_n
\end{aligned}
\]  \hfill (8.3.5)

where we recognize the PCA of \( R = MTQ \), which is the PCA of \( T \) with metrics defined by \( N \) on \( \mathbb{R}^p \) and \( P \) on \( \mathbb{R}^q \). Let us note as well that \( w_\lambda \) is a principal axis of the PCA of \( R^\top \), and \( w_n \) is a principal axis of the PCA of \( R \). As \( R \in \mathbb{R}^{p \times q} \) and as we have assumed that \( p \geq q \), it is natural to run the PCA of \( R \), hence compute \( w_n \) as a principal axis of \( R \). Then, \( w_\lambda \) as a principal axis of \( R^\top \) is a principal component of \( R \) and related to \( w_n \) by
\[
w_\lambda = Rw_n, \quad \text{with} \quad \begin{cases}
R & \in \mathbb{R}^{p \times q} \\
w_\lambda & \in \mathbb{R}^p \\
w_n & \in \mathbb{R}^q
\end{cases}
\]  \hfill (8.3.6)

Hence the SVD or search for eigenvalues and eigenvectors will be done once only.

- **Interpretation:** Hence the result: the solution of the CCA of two arrays \( A \) and \( B \) is the solution of the PCA of \( T = A^\top B \) with an inner product defined by \( N \) on rows and by \( P \) on columns where \( N = (A^\top A)^{-1} \) and \( P = (B^\top B)^{-1} \). We then have

\[
u_\lambda = Mw_\lambda, \quad u_n = Qw_n
\]

and
\[
y_\lambda = Au_\lambda, \quad y_n = Bu_n
\]

- **Algorithm:** There are several ways to write an algorithm for this calculation. Here is a direct one, without calling PCA or PCA-MET.
Algorithm 9 Canonical Correlation Analysis: $\text{CCA}(A, B)$

1: input $A \in \mathbb{R}^{n \times p}, B \in \mathbb{R}^{n \times q}$ with $p \geq q$
2: compute $T = A^T B$
3: compute $N = (A^T A)^{-1}$
4: compute $P = (B^T B)^{-1}$
5: compute $M = N^{1/2}$
6: compute $Q = P^{1/2}$
7: compute $R = MTQ$
8: compute $W_a, W_b, \Psi = \text{pca_core}(R)$
9: compute $\Lambda = \sqrt{\Psi}$
10: compute $U_a = MW_a$
11: compute $U_n = QW_n$
12: compute $Y_a = AU_a$
13: compute $Y_n = BU_n$
14: return $Y_a, Y_n, U_a, U_n, \Lambda$

Notes and references: Canonical Analysis seems to have been proposed by Hotelling in 1936 in Hotelling, H. - Relation between two sets of variables, Biometrika, 28:361-377. It is presented with a statistical approach in [And58, chap. 12] or [Rao73, Section 8f]. A review of Canonical Analysis with (debated) applications in ecology can be found in [Git85]. It is presented with a more algebraic approach in classical textbooks of the French school of data analysis, e.g. [LMT77, LMF82, EP90]. It is presented in [Sap90] too. Canonical Analysis is often referred to as Canonical Correlation Analysis (CCA). Both denominations will be used here.

9 Multiple Correspondence Analysis

Here, we develop a way to extend CCA with more than 2 arrays $A$ and $B$. There are different ways to do it, which are not equivalent. Indeed, let us have 3 arrays, $A$, $B$ and $C$. Let us select one component per array, respectively $y_a = Au_a$, $y_b = Bu_a$ and $y_c = Cu_c$ with $\|y_a\| = \|y_b\| = \|y_c\| = 1$. One can define three-ways CCA as finding a triplet $(y_a, y_b, y_c)$ such that their correlation is maximal. There is however no canonical way to define the correlation between 3 vectors. It can be defined as $\|y_a \wedge y_b \wedge y_c\|$, or $\|y_a + y_b + y_c\|$, or other ways. Here, we extend CCA to more than two arrays along a way which passes through an equivalence between CCA and CoA. We then extend CoA to more than two variables, and come back to CCA through the equivalence between CoA and CCA.

9.1 A tight link between Canonical Analysis and Correspondence Analysis

Let us consider a set of two qualitative variables $A$ and $B$ on a same set of items $[1, n]$. We build the so called indicator array of each variable. If there are $n$ items, $p$ modalities for $A$ and $q$ for $B$, it is a $n \times p$ array for $A$ and $n \times q$ for $B$, with, in each row $i$, zero in each column but 1 in column $j$ if the modality $j$ of the variable has been observed for item $i$.

$$\alpha_{ij} = \begin{cases} 1 & \text{if modality } j \text{ has been observed for item } i \\ 0 & \text{otherwise} \end{cases}$$
\[ \beta_{ik} = \begin{cases} 1 & \text{if modality } k \text{ has been observed for item } i \\ 0 & \text{otherwise} \end{cases} \]

Let us do the Canonical Analysis of both arrays \( A \) and \( B \). The solution is given by

\[ w = M^T M w, \quad v = D_n^{1/2} w \]

with

\[ M = D_n^{1/2} T D_n^{1/2}, \quad T = A^T B, \quad D_n = (A^T A)^{-1}, \quad D_n = (B^T B)^{-1} \]

Let us note that \((A^T A)^{-1} \in \mathbb{R}^{p \times p}\) (resp. \((B^T B)^{-1} \in \mathbb{R}^{q \times q}\)) is the diagonal matrix with in position \( j \) (resp. \( k \)) the inverse of the number of rows in \( A \) (resp. \( B \)) where the modality \( j \) (resp. \( k \)) of the variable has been observed. One recognizes in \( T \) the contingency table between \( A \) and \( B \), and in the PCA of \( M \) the PCA of \( T \) with inner product defined by \( D_n^{1/2} \in \mathbb{R}^p \) and by \( D_n^{1/2} \in \mathbb{R}^q \), i.e. correspondence analysis of \( T \). This leads to an intimate link between Correspondence Analysis and Canonical Analysis, which permits further an extension of Correspondence Analysis to more than two variables.

### 9.2 Link between Canonical Analysis and PCA with metric on rows

Let \( A \in \mathbb{R}^{n \times p} \) and \( B \in \mathbb{R}^{n \times q} \) with \( q \leq p \) and \( p + q \leq n \). Let us consider the data set

\[ X = [A|B] \in \mathbb{R}^{n \times (p+q)} \]

built by columnwise concatenation of \( A \) and \( B \). Let us define

\[ D_n = \begin{pmatrix} D_A & 0 \\ 0 & D_B \end{pmatrix} \in \mathbb{R}^{(p+q) \times (p+q)} \]

Let us perform the PCA of \( X \) with metrics on rows given by \( D \) (a row of \( X \) belongs to \( \mathbb{R}^{p+q} \)). Let us denote

\[ T = A^T B, \quad R = D_n^{1/2} T D_n^{1/2} \]

Then, performing the PCA of \( X \) with row distances given by \( D \) is performing the PCA of \( R \) with

\[ R = (A^T A)^{-1/2} A^T B (B^T B)^{-1/2} \]

where we recognise the CCA of \((A,B)\) (see equation 8.3.5 and algorithm 9). This leads to a possibility to perform Canonical Analysis on more than two arrays (see section 9.3).

- If \( A \) and \( B \) are indicative arrays, this is performing the CoA of \( T \), which is the contingency table of \( A \) and \( B \). When \( T \) (which is \((p+q) \times (p+q)\) is decomposed blockwise (see in the proof), it is called the Burt table of \( A \) and \( B \).

**Proof.** The solution is performing a PCA of

\[ S = XD_n^{1/2}, \quad S \in \mathbb{R}^{n \times p+q} \]
Linear Dimensionality Reduction

i.e. performing an EVD of \( R^T R \). We will denote the matrix dimensions under each block. We have

\[
S = \begin{bmatrix}
A D_{\frac{1}{2}}^{1/2}_n & B D_{\frac{1}{2}}^{1/2}_m
\end{bmatrix}_{n \times (p+q)}, \quad S^T = \begin{bmatrix}
D_{\frac{1}{2}}^{1/2}_n A^T & D_{\frac{1}{2}}^{1/2}_m B^T
\end{bmatrix}_{(p+q) \times n}
\]

So

\[
S^T S = \begin{bmatrix}
D_{\frac{1}{2}}^{1/2} A^T A D_{\frac{1}{2}}^{1/2} & D_{\frac{1}{2}}^{1/2} A^T B D_{\frac{1}{2}}^{1/2} \\
D_{\frac{1}{2}}^{1/2} B^T A D_{\frac{1}{2}}^{1/2} & D_{\frac{1}{2}}^{1/2} B^T B D_{\frac{1}{2}}^{1/2}
\end{bmatrix}
\]

Let us observe that

\[
A^T A = D^{-1}_n, \quad B^T B = D^{-1}_m
\]

Hence

\[
S^T S = \begin{bmatrix}
I_p & D_{\frac{1}{2}}^{1/2} A^T B D_{\frac{1}{2}}^{1/2} \\
D_{\frac{1}{2}}^{1/2} B^T A D_{\frac{1}{2}}^{1/2} & I_q
\end{bmatrix}
\]

Let

\[
x = \begin{bmatrix}
u \\
v
\end{bmatrix}
\]

be such that

\[
S^T S x = \lambda x
\]

This yields

\[
\begin{cases}
D_{\frac{1}{2}}^{1/2} A^T B D_{\frac{1}{2}}^{1/2} v = (\lambda - 1) u \\
D_{\frac{1}{2}}^{1/2} B^T A D_{\frac{1}{2}}^{1/2} u = (\lambda - 1) v
\end{cases}
\]

Let us denote

\[
T = A^T B, \quad R = D^{1/2}_n T D^{1/2}_n
\]

Then

\[
\begin{cases}
R v = (\lambda - 1) u \\
R^T u = (\lambda - 1) v
\end{cases}
\]

or

\[
\begin{cases}
R^T R v = (\lambda - 1)^2 v \\
R^T R u = (\lambda - 1)^2 u
\end{cases}
\]

where we recognize the PCA of \( R \).

9.3 Multiple Correspondence or Canonical Analysis

Knowing that, the extension to an analysis of more than two qualitative or quantitative variables is straightforward.

- Let us have \( m \) qualitative variables on \( n \) items, each with indicator matrix \( A_\ell (\ell \in [1, m]) \), or quantitative variables, with

\[
A_\ell \in \mathbb{R}^{n \times p_\ell}, \quad \sum_\ell p_\ell \leq n
\]
Let us define

\[ D_\ell = (A_\ell^T A_\ell)^{-1}, \in \mathbb{R}^{p_\ell \times p_\ell} \]

Let us build

\[ X = [A_1 \ldots A_m] \]

and

\[
D = \begin{bmatrix}
D_1 & 0 & \ldots & 0 \\
0 & D_2 & 0 & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
0 & \ldots & 0 & D_m
\end{bmatrix}
\]

Then, the MCoA or MCA of \((A_1, \ldots, A_m)\) is the PCA of \(X\) with distances on rows given by \(D\). It is the PCA of

\[ R = [A_1 D_1 | \ldots | A_m D_m] \]  

(9.3.1)

We have

\[ R^T = \begin{bmatrix}
D_1 A_1^T \\
\vdots \\
D_m A_m^T
\end{bmatrix} \]

and \(Z = R^T R\) can be written blockwise as

\[
\begin{cases}
Z_{\ell\ell} = I_{p_\ell} & \text{if } \ell = \ell' \\
Z_{\ell\ell'} = D_\ell A_\ell^T A_{\ell'} D_{\ell'} & \text{if } \ell \neq \ell'
\end{cases}
\]

Notes and references: These links between different treatments of a same dataset which establish some dependencies between methods have been subject to thorough studies and presentations by the French school of multivariate analysis, more algebraic and geometrical than statistical, in the 70’s. with the names of B. Escofier, J.-P. Fénelon, L. Lebart, A. Morineau, J. Pagès, among others. It is presented in all textbooks of this school in multivariate data analysis, under chapters called ”other methods and complements”, like [LMT77, LMF82]

10 Multidimensional Scaling

Multidimensional Scaling is a technique to map a discrete metric space into a Euclidean space. Let \((M, d)\) be a metric space with \(|M| = n\). It is given by a pairwise distance matrix

\[ D = [d_{ij}]_{1 \leq i,j \leq n} \]

such that

\[ d_{ij} = d(i, j) \]

MDS at dimension \(r\) is finding a point cloud

\[ X = (x_i)_{1 \leq i \leq n}, \quad x_i \in \mathbb{R}^r \]

such that the distance between points \(x_i\) and \(x_j\) is as close as possible from \(d_{ij}\) or, loosely speaking

\[ \|x_i - x_j\| \approx d_{ij} \]
A matrix $X \in \mathbb{R}^{n \times r}$ is attached to the point cloud $\mathcal{X}$, with $x_i$ being row $i$ of $X$.

- Then, two situations may occur

  1. either the distances $d_{ij}$ come from a Euclidean distance between (unknown) points, and the problem is to recover them, i.e. produce an isometry, and a best approximation of it at dimension $r$

  2. or they do not come from a Euclidean distance, and a best approximation is sought for, knowing there exists no exact isometry

Problem 1 is known as classical MDS and problem 2 as Least Square Scaling. In this note, we address the problem of classical MDS only. Least square scaling is a delicate and non trivial optimization problem. It appears however that, in practical cases, one never knows whether the distances come from a Euclidean point cloud or not, and the procedure is "as if".

- One thing to understand is that classical MDS assumes that the distances in metric space $(M, d)$ are $\ell^2$-norm distance in an (unknown) Euclidean space. If it is not the case, a numerical problem occurs (but a ad hoc standard solution is provided). Three points with pairwise distances can be arranged as a triangle in $\mathbb{R}^2$, and four points with pairwise distances can be arranged as a tetrahedron in $\mathbb{R}^3$. More generally, $n$ points with pairwise Euclidean distances can be isometrically embedded in $\mathbb{R}^{n-1}$ at most. Then, MDS at rank $r$ is made in two steps:

  1. find a point cloud $X$ in $\mathbb{R}^{n-1}$ with distances matching exactly the values of $d(i, j)$ for each pair

  2. reduce the dimension $r < n$ of the space where to build the point cloud $X_r$ as close as possible to $X$. This can be solved by PCA of $X$. It will appear that principal axis and components of the PCA of $X$ will be given by MDS without further calculations.

The procedure to build matrix $X$ attached to point cloud $\mathcal{X}$ is in four steps, presented hereafter

**Algorithm 10** General procedure for classical MDS

1: **given** a distance matrix $D \in \mathbb{R}^{n \times n}$ and a dimension $r < n$
2: **compute** the Gram matrix $G$ associated to $D$
3: **compute** the EVD or the SVD of $G$ with singular values $\Sigma$
4: **compute** the coordinates $X \in \mathbb{R}^{n \times n}$ of point cloud $\mathcal{X}$
5: **compute** the best low rank approximation $X_r \in \mathbb{R}^{n \times r}$ of $X$
6: **return** $X_r, \Sigma_r$

**Notes and references:** There exists many excellent textbooks presenting MDS. We can recommend [CC01] or [Ize08, chap. 13]. A comprehensive reference is [BG05]. A classical and rigorous reference with many results, their demonstration and history is [MKB79] which we highly recommend for those enjoying a mathematical based approach. Classical MDS has been proposed by Torgerson in 1952 [Tor52]. Here, we have followed [CC01, chap. 2].
10.1 The Gram matrix

Let \( X = (x_i) \) be a cloud on \( n \) points in \( \mathbb{R}^r \), such that the pairwise distances only are known.

\[ \| x_i - x_j \| = d_{ij} \]

and not the coordinates of the \( x_i \). The Gram matrix \( G \in \mathbb{R}^{n \times n} \) of \( X \) is the matrix with elements

\[ G_{ij} = \langle x_i, x_j \rangle \]

- There is a well known correspondence between the Gram Matrices \( G \) of inner products \( \langle x_i, x_j \rangle \) and the Euclidean Distance Matrix of quantities \( \| x_i - x_j \| \), both \( n \times n \). If

\[ \begin{bmatrix}
    g_{ij} & d_{ij}^2 \\
    d_{ij}^2 & \| x_i - x_j \|^2
\end{bmatrix} \]

Then

\[
\begin{align*}
    d_{ij}^2 &= g_{ii} + g_{jj} - 2g_{ij} \\
    g_{ij} &= -\frac{1}{2} (d_{ij}^2 - d_{i*}^2 - d_{j*}^2 + d_{**}^2)
\end{align*}
\]

(10.1.1)

with

\[
\begin{align*}
    d_{i*}^2 &= \frac{1}{n} \sum_j d_{ij}^2 \\
    d_{j*}^2 &= \frac{1}{n^2} \sum_{i,j} d_{ij}^2 = \frac{1}{n} \sum_i d_{i*}^2
\end{align*}
\]

(10.1.2)

Such a correspondence has been studied for decades (see e.g. [Sch38, Lau98]). We then have the scheme (with an arrow meaning "built from"):

\[
\begin{align*}
    X & \quad \overset{\text{G}}{\rightarrow} \quad G \\
    D & \quad \overset{\text{X}}{\rightarrow} \quad X
\end{align*}
\]

- A key question is to know whether one-way arrows \( X \rightarrow G \) and \( X \rightarrow D \) can be reversed, i.e. whether \( X \) can be computed knowing \( G \) or knowing \( D \). This amounts to answering to the question: a pairwise distance matrix \( D \) being given, is there a dimension \( m \) and a point cloud \( X \) in \( \mathbb{R}^m \) such that the distance between \( x_i \) and \( x_j \) is precisely \( d_{ij} \)?

- A matrix \( D \) being given, it is always possible to compute a matrix \( G \) by equation (10.1.1). But \( G \) is not necessarily positive, i.e. it is not necessarily the Gram matrix of a point cloud \( X \). The conditions on \( D \) for \( G \) to be positive, i.e. a Gram matrix have been thoroughly studied. In most of the case, when the Gram matrix is not positive, the negative eigenvalues are just ignored. This leads to some subtleties when connecting the EVD and the SVD of the Gram matrix to compute the coordinates.

- The coordinates of the point cloud \( X \) can be computed from the eigenvectors and eigenvalues, or the Singular Value Decomposition of the Gram matrix. The recipe is given here.
10.2 Eigendecomposition of the Gram Matrix

Let $G$ be the Gram matrix. If (this is an hypothesis) there exists a set of $n$ points in $\mathbb{R}^m$ such that

$$\forall i, j, \|x_i - x_j\| = d_{ij}$$

then

$$G_{ij} = \langle x_i, x_j \rangle$$ \hspace{1cm} (10.2.1)

and $G$ is positive. We assume here that $G$ as computed from equation (10.1.1) is positive.

- Let $X \in \mathbb{R}^{n \times m}$ be the matrix with row $i$ being $x_i$. Then

$$G = XX^\top$$ \hspace{1cm} (10.2.2)

Next step is about computing $X$ knowing $XX^\top$.

- Let $(u_\alpha, \sigma_\alpha)_\alpha$ be the set of eigenpairs of $G$

$$Gu_\alpha = \sigma_\alpha u_\alpha$$ \hspace{1cm} (10.2.3)

with

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$$

As $G$ is symmetric, the eigenvectors if normed form an orthonormal family. If $U = [u_1 | \ldots | u_n]$ is the matrix with $u_\alpha$ in column $\alpha$ and $\Sigma$ the diagonal matrix with $\sigma_\alpha$ in its diagonal, we have

$$GU = U\Sigma$$ \hspace{1cm} (10.2.4)

As $U$ is orthonormal, $UU^\top = \mathbb{I}_n$, we have by right multiplication by $U^\top$

$$G = U\Sigma U^\top$$ \hspace{1cm} (10.2.5)

We recognize here the SVD of $G$ (hence the choice of $\Sigma$ for the eigenvalue). The eigenvalues of $G$ are its singular values. This is a general property of symmetric matrices.

- As $G$ is definite positive, let

$$\Sigma = \Lambda^2$$

Then

$$G = (U\Lambda)(U\Lambda)^\top$$ \hspace{1cm} (10.2.6)

and we can select

$$X = UA, \quad \Lambda = \Sigma^{1/2}$$ \hspace{1cm} (10.2.7)

It is not the only solution, because any matrix $X' = X\Omega$ where $\Omega$ is a rotation in $\mathbb{R}^m$ is a solution too.

- Hence the algorithm:
Algorithm 11 MDS with eigendecomposition or SVD of Gram matrix

1: input: Gram matrix $G$
2: compute eigenpairs $(u_\alpha, \lambda_\alpha)$ such that $Gu_\alpha = \sigma_\alpha u_\alpha$
3: or compute the SVD of $G$ as $G = U\Sigma U^\top$
4: compute $\Lambda = \Sigma^{1/2}$
5: compute $X = U\Lambda$
6: return $X, \Lambda$

10.3 Dimension reduction

Once matrix $X$ has been computed, finding a point cloud $X_r$ of $n$ points in $\mathbb{R}^r$ as close as possible to $X$ is done by the PCA of $X$.

- Let us recall that
  \[ X = U\Sigma^{1/2} \quad \text{with} \quad U^\top U = I_n \] (10.3.1)

  We recognize here the SVD of $X$ as $X = U\Lambda V^\top$, with $\Lambda = \Sigma^{1/2}$ and $V = I_n$. Hence, $X = U\Sigma^{1/2}$ is the development of $X$ in the basis of its first axis, and its first columns are the first components.

- There is still one point to look at. In all these developments, we have assumed that the Gram matrix $G$ built from distance matrix $D$ with receipe in equation (10.1.1) is positive, i.e. that all eigenvalues of $G$ (the $\sigma_\alpha$) are non-negative. This is not always the case with real data. Without going too much into details, if one eigenvalue at least is negative, there is no isometry between $(M,d)$ and a Euclidean space, whatever its dimension. However, it can be shown that there exists a quadratic embedding between $(M,d)$ and a pseudo-euclidean space, i.e. a vector space with a quadratic form with a signature $(p,q)$. Such an embedding is rarely done, and most of the time the axis (and components) associated to negative eigenvalues of $G$ simply are ignored, or clipped to zero.

10.4 MDS algorithm

Wrapping all this together yields the following algorithm for MDS

Algorithm 12 Classical MDS: $X, \Lambda = \text{mds}(D,r)$

1: input: a distance matrix $D \in \mathbb{R}^{n \times n}$; a dimension $r < n$
2: compute the Gram matrix of $D$: $G = \text{gram}(D)$
3: compute the eigenpairs $(u_\alpha, \lambda_\alpha)$ such that $Gu_\alpha = \sigma_\alpha u_\alpha$
4: or compute the SVD of $G$ as $G = U\Sigma U^\top$
5: keep in $U$ the columns associated to non-negative eigenvalues of $G$; clip them off in $\Sigma$
6: compute $\Lambda = \Sigma^{1/2}$
7: compute $X = U\Lambda$
8: keep in $X_r$ the $r$ first columns of $X$ only
9: return $X_r, \Lambda$

- Notes: Here are some notes on this algorithm:
We have not specified the choice of the \( r \) first columns of \( X \) in algorithm summarized in section (11).

It is better to have as an output the whole set of non-negative eigenvalues of \( G \). Keeping the \( r \) first if useful is very simple.

If the user selects the computation of the eigenvalues of \( G \), it is simple to detect those which are negative, and clip the corresponding columns in \( U \) (and rows and columns in \( \Sigma \)). If the SVD is selected, we have \( G = U \Sigma U^T \). We have the choice of the sign in columns of \( U \). If the choice is the same in \( U \) and \( V^T = U^T \), then the sign of \( \sigma_\alpha \) in the SVD is the same than the corresponding eigenvalue of \( G \) (SVD algorithms keep \( \sigma_\alpha \geq 0 \) by changing the sign of corresponding column in \( V \)).
## 11 Summary

Here is a summary of the methods with

- the name of the function
- the call of the function
- the calculations to produce the result

| Method | Call | Computation |
|--------|------|-------------|
| PCA    | \((Y, V, \Lambda) = \text{PCA\_CORE}(A)\) | \(C = A^t A\)  
\((\Lambda, V) = \text{eig}(C)\)  
\(Y = AV\)  
or  
\((U, \Sigma, V) = \text{svd}(A)\)  
\(\Lambda = \Sigma^2\)  
\(Y = U\Sigma\) |
| PCAiv  | \((Y, V, \Lambda) = \text{PCA\_IV}(A, U, V)\) | \(P_E = U(U^t U)^{-1} U^t\)  
\(P_F = V(V^t V)^{-1} V^t\)  
\(A_{u,v} = P_E A P_F\)  
\((Y, \Lambda, V) = \text{PCA\_CORE}(A_{u,v})\) |
| PCAmet | \((Y, V, \Lambda) = \text{PCA\_MET}(A, M, Q)\) | \(A_{m,q} = MAQ\)  
\((Z, \Lambda, X) = \text{PCA\_CORE}(A_{m,q})\)  
\(Y = M^{-1} Z\)  
\(V = Q^{-1} X\) |
| CoA    | \((Y_r, Y_c, \Lambda) = \text{CoA}(T)\) | \(A = T/T_{++}\)  
\(T_{++} = \sum_{i,j} T_{ij}\)  
\(r_i = \sum_j \alpha_{ij}\)  
\(c_j = \sum_i \alpha_{ij}\)  
\(M = \text{diag}(1/\sqrt{r_i})\)  
\(Q = \text{diag}(1/\sqrt{c_j})\)  
\(A_{m,q} = M(A - r \otimes c)Q = \alpha_{i,j-s,q}/\sqrt{r_i c_j}\)  
\(Z, \Lambda, X = \text{PCA\_CORE}(A_{m,q})\)  
\(Y_r = M^{-1} Z, \quad \bar{Y}_c = Q^{-1} X\) |
| CCA    | \((Y_A, Y_B, U_A, U_B, \Lambda) = \text{CCA}(A, B)\) | \(T = A^t B\)  
\(M = (A^t A)^{-1/2}\)  
\(Q = (B^t B)^{-1/2}\)  
\(R = MTQ\)  
\(W_A, \Psi, W_B = \text{PCA\_CORE}(R)\)  
\(\Lambda = \sqrt{\Psi}\)  
\(U_A = MW_A, \quad U_B = QW_B\)  
\(Y_A = AU_A, \quad Y_B = BU_B\) |
| MDS    | \(Y, \Lambda = \text{MDS}(D, r)\) | \(G = \text{gram}(D)\)  
\((U, \Sigma) = \text{svd}(G): G = U\Sigma U^t\)  
\(Y = U\Sigma^{1/2}\) |
12 References in textbooks

In this section, we indicate where, and under which name, some techniques are presented in most classical textbooks.

| Technique                          | Reference | Location       |
|------------------------------------|-----------|----------------|
| Canonical Correlation Analysis     | [And58]   | chapter 12    |
| [Rao73]                            | section 8f1|
| [MKB79]                            | chapter 10 |
| [Jol02]                            | section 9.3|
| [Ize08]                            | section 7.3|
| [HTF09]                            | section 14.5.1|
| [Mur12]                            | section 12.5.3|
| Correspondence Analysis            | [Gre84]   | the book      |
| [Ize08]                            | chapter 17 |
| Factor Analysis                    | [And58]   | section 14.7   |
| [Rao73]                            | section 8f4|
| [MKB79]                            | chapter 9  |
| [Ize08]                            | section 15.4|
| [HTF09]                            | section 14.7.1|
| [Mur12]                            | section 12.1|
| Independent Component Analysis     | [Ize08]   | section 15.3   |
| [HTF09]                            | section 560|
| Karhunen Loeve transform           | [Mur12]   | section 12.2, p. 387|
| Latent variables                   | [Bis06]   | chapter 12    |
| [Ize08]                            | chapter 15 |
| [HTF09]                            | section 14.7.1|
| [Mur12]                            | chapter 12 |
| Multiple Correspondence Analysis   | [Ize08]   | section 17.4   |
| Non metric scaling                 | [CC01]    | chapter 3     |
| [Ize08]                            | section 13.9|
| Principal Component Analysis       | [And58]   | chapter 11    |
| [MKB79]                            | chapter 8  |
| [Jol02]                            | the book   |
| [Bis06]                            | section 12.1|
| [Ize08]                            | section 7.2 |
| [Mur12]                            | section 12.2|
| Probabilistic PCA                  | [Bis06]   |                |
| [Mur12]                            | section 12.2.4|
| Sensible PCA                       | [Mur12]   | section 12.2, p. 387|
| Sparse PCA                         | [HTF09]   | section 14.5.5|
| Supervised PCA                     | [Mur12]   | section 12.5.1|

RR n° 9488
Abbreviations

| Abbreviation | Description                        |
|--------------|------------------------------------|
| CCA          | Canonical Correlation Analysis     |
| CA           | Canonical Analysis                 |
| CoA          | Correspondance Analysis            |
| EVD          | Eigenvalue Decomposition           |
| IV           | Instrumental Variables             |
| MCA          | Multiple Canonical Analysis        |
| MCoA         | Multiple Correspondence Analysis   |
| MDS          | Multidimensional Scaling           |
| PCA          | Principal Component Analysis       |
| SDP          | Symmetric Definite Positive        |
| SGF          | Symmetric Gauge Function           |
| SVD          | Singular Value Decomposition       |
| UIN          | Unitarily Invariant Norm           |

Classical Scaling  
[CC01] section 2.2  
[Ize08] section 13.6

Multidimensional Scaling  
[CC01] the book  
[MKB79] chapter 14  
[Ize08] chapter 13  
[HTF09] section 14.8

Non metric scaling  
[CC01] chapter 3  
[Ize08] section 13.9
References

[ACD+22] E. Agullo, O. Coulaud, A. Denis, M. Faverge, A. Franc, J.-M. Frigerio, N. Furmento, A. Guilbaud, E. Jeannot, R. Peressonì, F. Pruvost, and S. Thibault. Task-based randomized singular value decomposition and multidimensional scaling. Research Report RR-9482, Inria Bordeaux - Sud Ouest ; Inrae - BioGeCo, September 2022.

[And58] T. W. Anderson. An introduction to Multivariate Statistical Analysis. John Wiley & Sons, 1958.

[BCF+18] P. Blanchard, P. Chaumeil, J.-Marc. Frigerio, F. Rimet, F. Salin, S. Thérond, O. Coulaud, and A. Franc. A geometric view of Biodiversity: scaling to metagenomics. Research Report RR-9144, INRIA ; INRA, January 2018.

[Ben73a] J.-P. Benzecri. L’Analyse des Données ; tome 2: l’analyse des correspondances. Dunod, 1973.

[Ben73b] J.-P. Benzecri. L’Analyse des Données, tome 1: la taxinomie. Dunod, 1973.

[BG05] I. Borg and P. J. F. Groenen. Modern Multidimensional Scaling. Springer Series in Statistics. Springer, second edition, 2005.

[Bis06] C. M. Bishop. Pattern recognition and machine learning. Springer, Berlin, 2006.

[CC80] C. Chatfield and A. J. Collins. Introduction to Multivariate Analysis. Chapman & Hall, 1980.

[CC01] T.F. Cox and M. A. A. Cox. Multidimensional Scaling - Second edition, volume 88 of Monographs on Statistics and Applied Probability. Chapman & al., 2001.

[Cla87] A. Claret. Contribution au problème de l’approximation factorielle d’un tableau de données. PhD thesis, Université des Sciences et techniques du Languedoc, 1987.

[CP79] F. Cailliez and J.-P. Pagès. Introduction à l’Analyse des Données. S.M.A.S.H. Editions, 1979.

[CST00] N. Cristianini and J. Shawe-Taylor. An introduction to Support Vector Machines and other Kernel-based learning methods. Cambridge University Press, 2000.

[EP90] B. Escofier and J. Pagès. Analyse Factorielles simple et multiples. Dunod, Paris, 1990.

[EY36] K. Eckart and G. Young. The approximation of a matrix by another of lower rank. Psychometrika, 1(3):211–218, 1936.

[Fra92] A. Franc. Etude Algébrique des multitableaux: apports de l’algèbre tensorielle - PhD Thesis. PhD thesis, Université Montpellier 2, 1992.

[Fra21] A. Franc. Tensor Ranks for the Pedestrian for Dimension Reduction and Disentangling Interactions. Research Report RR-9445, Inrae - BioGeCo ; Inria Bordeaux Sud-Ouest, December 2021.

[Git85] R. Gittins. Canonical Analysis: A Review with Applications in Ecology, volume 12 of Biomathematics. Springer, 1985.
[Gro84] M. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic Press, 1984.

[Hil74] M. O. Hill. Correspondence analysis: A neglected multivariate method. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 23(3):320–354, 1974.

[HJ12] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge, second edition, 2012.

[HMT11] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[HTF09] T. Hastie, R. Tishibani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, second edition, 2009.

[Ize08] A. J. Izenman. *Modern Multivariate Statistical Techniques*. Springer, NY, 2008.

[Jac91] J. E. Jackson. *A user’s guide to principal components*. Wiley, 1991.

[JC16] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A*, page 20150202, 2016.

[Jol02] I. T. Jolliffe. *Principal Component Analysis*. Sprin, second edition, 2002.

[Lau98] M. Laurent. A connection between positive semidefinite and euclidean distance matrix completion problems. *Linear Algebra and its Applications*, 273(9-22), 1998.

[LMF82] L. Lebart, A. Morineau, and J.-P. Fénelon. *Traitement des données statistiques*. Dunod, Paris, 1982.

[LMT77] L. Lebart, A. Morineau, and N. Tabard. *Techniques de la description statistique*. Bordas - Dunod, 1977.

[LV07] J. A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer, NY, 2007.

[MKB79] K. V. Mardia, J.T. Kent, and J. M. Bibby. *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press, 1979.

[Mur12] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[NG07] O. Nenadić and M. Greenacre. Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. *Journal of Statistical Software*, 20(3):1–12, 2007.

[PCY79] J.-P. Pagès, F. Cailliez, and Escoufier Y. Analyse factorielle : un peu d’histoire et de géométrie. *Revue de Statistiques Appliquées*, 27(1):5–28, 1979.

[Rao64] C. R. Rao. The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya*, 26(4):329–368, 1964.

[Rao73] C. R. Rao. *Linear Statistical Infernece and its Applications*. Wiley Series in Probability and Mathematical Statistics. Wiley, second edition, 1973.

[Sab84] R. Sabatier. Quelques généralisations de l’Analyse en Composantes Principales de Variables Instrumentales. *Stat. & Ann. Donn.*, 9(3):75–103, 1984.
[Sap90] G. Saporta. Probabilités, Analyse de Données et Statistique. Editions Technip, 1990.

[Sch38] I. J. Schoenberg. Metric Spaces and Positive Definite Functions. Transactions of the American Mathematical Society, 44(3):522–536, 1938.

[Sch60] N. Schatten. Norms ideals of completely continuous operators. Springer Verlag, 1960.

[SS04] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. Statistics and Computing, 14:199–222, 2004.

[SSBD14] S. Shalev-Shwartz and S. Ben-David. Understanding Machine Learning - from theory to algorithms. Cambridge University Press, 2014.

[Tor52] W. S. Torgerson. Multidimensional Scaling: I. Theory and Method. Psychometrika, 17(4):401–419, 1952.

[TY85] 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 multivariate categorical data. Psychometrika, 50(1):91–119, 1985.

[Wan12] J. Wang. Geometric structure if high-dimensional data and dimensionality reduction. Springer & Higher Education Press, 2012.

[Wol87] S. Wold. Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems, 2:37–52, 1987.
