Linear Dimensionality Reduction

Alain Franc

Linear Dimensionality Reduction

Alain Franc

RESEARCH REPORT
N° 9488
September 2022
Project-Team Pleiade

ISSN 0249-6399 ISRN INRIA/RR--9488--FR+ENG

Linear Dimensionality Reduction

Alain Franc*†

Project-Team Pleiade

Research Report n° 9488 — September 2022 — 69 pages

Abstract: These notes are an overview of some classical linear methods in Multivariate Data Analysis. This is an good old 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`

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

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

Avant-propos

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 faCcon 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 franCcaise, 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'Analyse Canonique, 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 franCcais 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'apprentissge 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

où chaque méthode peut être lue 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, voir [HMT11]),
- 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])

→ 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 mat riciel, 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

Foreword

The writing of this document was motivated by the convergence of several observations.

◇ The diversity of methods described as Data Analysis in the 1970s, nowadays attached to Statistical Learning, can be organised in a circular way where one is a variation of the other via certain choices. This global approach is essentially algebraic, based on matrix calculation, and has been highly developed by a French school, in parallel with more statistical approaches in Anglo-Saxon countries.

◇ The main methods are PCA (Principal Component Analysis), normalized PCA, Correspondence Analysis (The term "Factorial" has been dropped in the English translation) of contingency tables, analysis of two tables by Canonical Analysis, MDS (Multidimensional Scaling, a method of constructing a point cloud from a table of distances, term for which there is no well accepted French term). These methods had fallen into disuse as descriptive methods, but have experienced a revival of interest in exploring structures in massive data (which is sometimes called "pattern discovery" and is related to unsupervised learning).

◇ The starting point for each of these methods is a data array, which can be a cross-tabulation between objects and variables, a contingency table, a distance or dissimilarity matrix, etc. ... A key observation is that each of the methods is organised according to the triptych:

where each method can be read as a diabolo, where a pre-processing of the data constructs a matrix, which is itself decomposed via an SVD (Singular Value Decomposition), the outputs of which are post-processed to produce the desired result. The SVD step is generally the limiting step for scaling up, i.e. processing massive data from very large arrays: the complexity is cubic with the size of the arrays.

However, significant progress has been made recently in scaling up the SVD by combining three elements:

- an evolution of the SVD computation algorithm by including the Gaussian Random Projection (rSVD, see [HMT11]),
- an implementation of distributed memory calculations for the elementary operations of matrix calculation,
- the use of a task-based programming paradigm for the assembly of these steps.

This progress has been made in particular,

→ for the integration of the rSVD in the MDS algorithm, by a collaboration between HiePACS, Pleiade, IDRIS and INRAE (see [BCF+18])

→ for the numerical implementation (distributed memory, task graph programming), through a collaboration between the HiePacs, Tadaam, Storm and Pleiade teams during an ADT called ADT Gordon, carried out in 2019-2020 (see [ACD+22])

By providing this simplified presentation of the diversity of linear dimension reduction methods, making explicit the successive steps that call for matrix computation, emphasising the common ford of scaling up to SVD to comfortably handle massive data, building on the success of ADT Gordon in scaling up to the SVD, we hope to facilitate the extension of this approach with the same success into the zoological garden of linear dimension reduction methods, to infer certain patterns by revealing low rank structure in the matrices under study.

A.F., at Pierroton, September 22, 2022

Contents

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 $f(x) = \beta + \sum_i w_i x_i$, 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 $O(\mathrm{Log}\, p)$) 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 indviduals 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

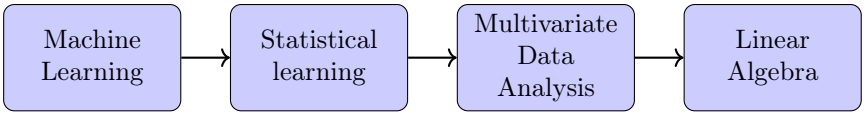
Beyond this dictionary, there is genuine innovation in machine learning, which is about inferring meaningful pattern in data in a very elaborate way, far beyond data analysis (see e.g. [SSBD14]). A grail is to mimic the way a brain learns from experience. This is beyond the scope of those

[1] Which deserves to be called a theorem, but is 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:



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 a 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 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 α_{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 *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 observatons x can be modled as $\mathbb{P}(x \mid z)$.

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: $A = U\Sigma V^{\mathrm{T}}$. These methods have experienced[3] two revolutions

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

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 [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).

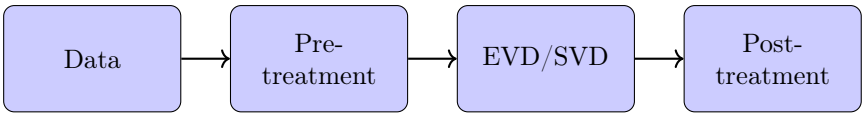
The high diversity of existing method can be organized into a small set of iconic methods, knowing that such a classification is far from being either unique or universally adopted. However, most of textbooks presenting these methods progressively reach an agreement on the poles around which to organize their variety and similarities/dissimilarities. We will not discuss this

[3] at least ...

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.

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 [Jol02] XXX and a current research with "Principal Component Analysis" on Google Scholar yielded more than 2 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:

$\|.\|$	Frobenius norm (unless otherwise stated)
$[\![a, b]\!]$	the set of integers i with $a \leq i \leq b$
a_{ij}	coefficient in row i nd colimn j of matrix A
a_{i*}	row i of matrix A $\quad (\in \mathbb{R}^p)$
a_{*j}	column j of matrix A $\quad (\in \mathbb{R}^n)$
A	a matrix in $\mathbb{R}^{n \times p}$
$A \geq 0$	a non-negative matrix
$\mathcal{A}$	a point cloud of n points in $\mathbb{R}^p$
$\mathbb{I}_n$	the identity matrix in $\mathbb{R}^{n \times n}$
$\mathcal{L}(E, F)$	the space of linear functions from E to F
$\mathbb{R}^{n \times p}$	The space of matrices with n rows and p columns

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 $\mathcal{A}$ 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 $\mathcal{A}_{\text{E}}$ of $\mathcal{A}$ on E is as close as possible to $\mathcal{A}$.

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 Froenius (ℓ^2) norm.

statistical: A set of p random variables being observed on n items independently, find r independent linear combination 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 ℓ^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:**

Given	$A \in \mathbb{R}^{n \times p}$	
	$0 < r < p$	
Find	$A_r \in \mathbb{R}^{n \times p}$	
with	$\text{rank } A_r = r$	
such that	$\|A - A_r\|$	minimal

- **Geometric approach:**

Given a point cloud	$\mathcal{A} = (a_1, \dots, a_n)$ $a_i \in \mathbb{R}^p$ $0 < r < p$
Find	a subspace $E_r \subset \mathbb{R}^p$
with	$\dim E_r = r$
such that	$d(\mathcal{A}, \mathcal{A}_r)$ minimal
where	$d(\mathcal{A}, \mathcal{A}_r) = \sum_i \|a_i - \widetilde{a_i}\|^2$
and	$\widetilde{a_i}$ is the projection of a_i on E_r

• Let us note that a point cloud $\mathcal{A}$ is equivalent to a matrix A with the row i of A being the point $a_i \in \mathcal{A}$. Knowing that, both approaches are equivalent. To see that, let us consider an orthonormal basis $(v_1, \dots, v_r)$ of E_r, and let us complete it to have an orthonormal basis of $\mathbb{R}^p$ as $(v_1, \dots, v_r, v_{r+1}, \dots, v_p)$. If the projection is exact, i.e. if $\forall\, i,\ a_i = \widetilde{a_i}$, 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 $\widetilde{a_i}$ is the projection of a_i on E_r, we have, by Pythagore theorem

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

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

Given a point cloud	$\mathcal{A} = (a_1, \dots, a_n)$ $a_i \in \mathbb{R}^p$ $0 < r < p$	
Find	a subspace $E_r \subset \mathbb{R}^p$	(3.2.1)
with	$\dim E_r = r$	
such that	$\sum_i \|\widetilde{a_i}\|^2$ maximal	
where	$\widetilde{a_i}$ 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 $\widetilde{a_i} = \langle a_i, u \rangle u$, and the optimization problem can be stated as

$$\left|\begin{array}{lll} \text{Given a point cloud} & \mathcal{A} = (a_1, \ldots, a_n) \\ & a_i \in \mathbb{R}^p \\ \\ \text{Find} & \text{a vector } u \in \mathbb{R}^p \\ \text{with} & \|u\| = 1 \\ \\ \text{such that} & \|Au\| \quad \text{maximal} \end{array}\right. \tag{3.2.2}$$

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

$$A^{\mathrm{T}}Av = \lambda v, \qquad \lambda = \max\{\lambda \in \mathrm{Sp}\, A^{\mathrm{T}}A\} \tag{3.2.3}$$

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

$$A^{\mathrm{T}}Av_j = \lambda_j v_j \qquad \text{with} \quad \lambda_1 \geq \ldots \geq \lambda_r \geq \lambda_{r+1} \geq \ldots \geq \lambda_p \geq 0 \tag{3.2.4}$$

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

$$A^{\mathrm{T}}AV = V\Lambda \tag{3.2.5}$$

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

$$Y = AV \tag{3.2.6}$$

Hence an first algorithm:

Algorithm 1 PCA of a matrix with EVD: PCA_EVD(A)

1: **input** $A \in \mathbb{R}^{n\times p}$
2: **compute** $C = A^{\mathrm{T}}A$
3: **compute** $(\lambda_\alpha, v_\alpha)$ such that $Cv_\alpha = \lambda_\alpha v_\alpha$, or $CV = V\Lambda$
4: **compute** $Y = AV$
5: **return** Y, Λ, 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	$C = A^{\mathrm{T}}A$
V	$\in \mathbb{R}^{p\times p}$	new basis	$Cv = \lambda v$
Λ	$\in \mathbb{R}^{p}$	eigenvalues of C in decreasing order	
Y	$\in \mathbb{R}^{n\times p}$	matrix of coordinates in new basis	$Y = AV$

3.3 Link with SVD

Let (U, Σ, V) be the SVD of A.

$$A = U\Sigma V^{\mathrm{T}} \tag{3.3.1}$$

Then

$$\begin{aligned} C &= A^{\mathrm{T}}A \\ &= (V\Sigma U^{\mathrm{T}})(U\Sigma V^{\mathrm{T}}) \\ &= V\Sigma^2 V^{\mathrm{T}} \qquad \text{as} \quad U^{\mathrm{T}}U = \mathbb{I}_n \end{aligned} \tag{3.3.2}$$

and

$$CV = V\Sigma^2 \qquad \text{as} \quad V^{\mathrm{T}}V = \mathbb{I}_p \tag{3.3.3}$$

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

$$\begin{aligned} Y &= AV \\ &= U\Sigma V^{\mathrm{T}}V \\ &= U\Sigma \end{aligned} \tag{3.3.4}$$

This yields a second algorithm for PCA:

Algorithm 2 PCA of a matrix with SVD: PCA_SVD(A))

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

Adopting the SVD viewpoint has some advantages:

→ there exists efficient and stable algorithms for computing a SVD,

→ it avoids a matrix × matrix computation ($C = A^{\mathrm{T}}A$) which can be costly when n and p are large,

→ it leads to easier generalization with instrumental variables or metrics on row or column space (see section 7),

→ 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 $\mathcal{O}(n^2 p)$. 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 randomzed 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

$$\widetilde{A} = QQ^{\mathrm{T}}A$$

Let us denote

$$B = Q^{\mathrm{T}}A, \qquad B \in \mathbb{R}^{k\times p}$$

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

$$A \approx QB = Q(U_{\mathrm{B}}\Sigma V) = (QU_{\mathrm{B}})\Sigma V = U\Sigma V$$

So

$$A \approx U\Sigma V \qquad \text{with} \quad U = QU_{\mathrm{B}} \tag{3.4.1}$$

Next step is to show that (what we will not develop here), when n and p are large, $A \approx QQ^{\mathrm{T}}A$ 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 $x, y \in X$

$$(1-\epsilon)\|x-y\|^2 \leq \|f(x)-f(y)\|^2 \leq (1+\epsilon)\|x-y\|^2 \tag{3.4.2}$$

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

$$k \geq \frac{8\ \mathrm{Log}\, n}{\epsilon^2} \tag{3.4.3}$$

The bad news is that ϵ^2 is at the denominator (so, k is large when ϵ is small), but the good news is that k grows with $\mathrm{Log}\, n$ 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 Ω is a random matrix (Y is in the span of A). There are several ways to chose Ω, and here we restrict ourselves to the Gaussian random projection, i.e. Ω 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 Ω 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 $\text{SVD\_GRP}(A, k)$)

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

- Here are the dimensions of the involved matrices:

Matrix	dimensions	computation
A	$n \times p$	data
Ω	$p \times k$	Gaussian random matrix
Y	$n \times k$	$Y = A\Omega$
Q	$n \times k$	$Y = QR$
B	$k \times p$	$B = Q^{\text{T}} A$
U_{B}	$k \times k$	$B = U_{\text{B}} \Sigma V$
Σ	$k \times k$	idem
V	$k \times p$	idem
U	$n \times k$	$U = Q^{\text{T}} U_{\text{B}}$

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)

input $A \in \mathbb{R}^{n \times p}$, $k \in \mathbb{N} \cup \{-1\}$, `meth` $\in \{$EVD, SVD, GRP$\}$
if `meth`==EVD **then**
  compute $C = A^{\mathrm{T}} A$
  compute $(\lambda_\alpha, v_\alpha)$ such that $Cv_\alpha = \lambda_\alpha v_\alpha$, or $CV = V\Lambda$
  compute $Y = AV$
  if $k > 0$ **then**
    $Y = Y$`[:,0:k]`; $\Lambda = \Lambda$`[0:k]`; $V = V$`[:,0:k]`
  end if
end if
if `meth`==SVD **then**
  compute $U, \Sigma, V =$ SVD(A)
  compute $\Lambda = \Sigma^2$
  compute $Y = U\Sigma$
  if $k > 0$ **then**
    $Y = Y$`[:,0:k]`; $\Lambda = \Lambda$`[0:k]`; $V = V$`[:,0:k]`
  end if
end if
if `meth`==GRP **then**
  compute $U, \Sigma, V =$ SVD_GRP(A, k)
  compute $\Lambda = \Sigma^2$
  compute $Y = U\Sigma$
end if
return Y, Λ, V

- **Comments:** Here are some comments:

Why PCA_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 preatreatment (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
$$Y, \Lambda, V = \text{PCA\_CORE(A')}$$
which is denoted PCA_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 $\approx 10^3$), 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 EVD or SVD only. A rank $k > 0$ must be prescribed for method 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 `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, `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, $\mathcal{A}$ its associated point cloud in $\mathbb{R}^p$, and (Y, V, Λ) the `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 $\mathcal{A}$ on principal axis j.

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

$$A = U\Sigma V^{\mathrm{T}} \tag{3.6.1}$$

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

$$Y = U\Sigma \tag{3.6.2}$$

Then

$$Y = U\Sigma = U\Sigma(V^{\mathrm{T}}V) = (U\Sigma V^{\mathrm{T}})V = AV \tag{3.6.3}$$

which means that the rows of Y are the coordinates of the points of $\mathcal{A}$ in new basis V. □

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

$$y \otimes v \equiv yv^{\mathrm{T}}$$

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

$$A = \sum_{j=1}^{p} y_j \otimes v_j \tag{3.6.4}$$

from which

$$\|A\|^2 = \sum_j \|y_j\|^2 \tag{3.6.5}$$

Proof. Indeed,

$$\begin{aligned}
\|A\|^2 &= \left\| \sum_{j=1}^{p} y_j \otimes v_j \right\|^2 \\
&= \left\langle \sum_{j=1}^{p} y_j \otimes v_j \ , \ \sum_{k=1}^{p} y_k \otimes v_k \right\rangle \\
&= \sum_{j,k} \langle y_j \otimes v_j \ , \ y_k \otimes v_k \rangle \\
&= \sum_{j,k} \langle y_j \ , \ y_k \rangle . \ \langle v_j \ , \ v_k \rangle \\
&= \sum_{j} \langle y_j \ , \ y_j \rangle \\
&= \sum_{j} \|y_j\|^2
\end{aligned}$$

□

(another way to see this is to observe that Y is deduced from A by a rotation, which is an isometry, then $\|A\| = \|Y\|$). As $Y = U\Sigma$ with $U^{\mathrm{T}}U = \mathbb{I}_p$, we have

$$\|y_j\| = \sigma_j \tag{3.6.6}$$

Hence, the norm of A can be partitioned as

$$\|A\|^2 = \sum_j \sigma_j^2 \tag{3.6.7}$$

Let us recall that

$$\lambda_j = \sigma_j^2, \qquad A^{\mathrm{T}} A v_j = \lambda_j v_j$$

Then

$$\|A\|^2 = \sum_{i=1}^{p} \lambda_i \tag{3.6.8}$$

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

$$\varrho_j = \frac{\lambda_j}{\sum_i \lambda_i} \tag{3.6.9}$$

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

$$\begin{aligned}
\rho_r &= \sum_{j=1}^{r} \varrho_j \\
&= \frac{\sum_{j=1}^{r} \lambda_j}{\sum_i \lambda_i}
\end{aligned} \tag{3.6.10}$$

- The quality of representation of item i on axis $j \in \{1, p\}$ is

$$\psi(i,j) = \frac{y_{ij}^2}{\sum_{\ell=1}^{p} y_{i\ell}^2} \tag{3.6.11}$$

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

$$\begin{aligned} \theta(i,r) &= \sum_{j=1}^{r} \psi(i,j) \\ &= \frac{\sum_{j=1}^{r} y_{ij}^2}{\sum_{\ell=1}^{p} y_{i\ell}^2} \end{aligned} \tag{3.6.12}$$

We have

$$\begin{cases} \varrho_j = \sum_{i=1}^{n} \psi(i,j) \\ \rho_r = \sum_{i=1}^{n} \theta(i,r) \end{cases} \tag{3.6.13}$$

This can be summarized as

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

  → select a rank r, and deduce the quality ρ of the approximation,

  → 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_{i*} denotes the row i of $A \in \mathbb{R}^{n \times p}$, and $a_{*j} \in \mathbb{R}^n$ 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:

$$\text{in } \mathbb{R}^p, \qquad a_i \xrightarrow{\text{centering}} \overline{a_i} = a_i - g, \tag{3.7.1}$$

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

$$g_j = \frac{1}{n} \sum_i a_{ij} \tag{3.7.2}$$

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

$$\forall\, j, \quad \sum_i \overline{a}_{ij} = 0 \tag{3.7.3}$$

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

$$a_{*j} \xrightarrow{\text{scaling}} \frac{a_{*j}}{\|a_{*j}\|} \tag{3.7.4}$$

The centered-scaled matrix A' is defined by

$$a_{*j} \longrightarrow \frac{\overline{a}_{*j}}{\|\overline{a}_{*j}\|} \qquad \text{with} \quad \overline{a}_{*j} = a_{*j} - g_j \mathbf{1}_n \tag{3.7.5}$$

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

Algorithm 5 PCA-SC(A)

input $A \in \mathbb{R}^{n \times p}$
compute the barycenter of A: $g = \frac{1}{n} \sum_i a_{i*}$
center A : $\forall\, i, \quad a_i \longrightarrow \overline{a}_i = a_i - g$
scale $\overline{A}$: $\forall\, j, \quad \overline{a}_{*j} \longrightarrow a'_{*j} = \frac{\overline{a}_{*j}}{\|\overline{a}_{*j}\|}$
do $Y, \Lambda, V =$ PCA_CORE(A')
return Y, Λ, V

- There are a few elementary results for scaled-centered `PCA` . By definition, the coefficients $c_{j\ell}$ of $C = {A'}^{\mathrm{T}} A'$ are the correlations between centered scaled variables a'_{*j} and $a'_{*\ell}$. Hence, we have

$$-1 \leq c_{j\ell} \leq 1 \tag{3.7.6}$$

We have as well

$$\forall\, j, \qquad c_{jj} = 1 \tag{3.7.7}$$

Hence

$$\sum_j \lambda_j = \operatorname{Tr} C = p \tag{3.7.8}$$

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

$$\rho_r = \frac{1}{p} \sum_{j \leq r} \lambda_j \tag{3.7.9}$$

• **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 Correspondecne 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

$$a_{ij} = \underbrace{m}_{\text{global mean}} + \underbrace{x_i}_{\text{effect of row } i} + \underbrace{y_j}_{\text{effect of column } j} + \underbrace{r_{ij}}_{residuals} \tag{3.7.10}$$

with

$$\left\{ \begin{array}{rrcl} & \displaystyle\sum_i x_i & = & 0 \\ & \displaystyle\sum_j y_j & = & 0 \\ \forall\, j, & \displaystyle\sum_i r_{ij} & = & 0 \\ \forall\, i, & \displaystyle\sum_j r_{ij} & = & 0 \end{array} \right. \tag{3.7.11}$$

Then, we have

$$\left\{ \begin{array}{rcl} m & = & \displaystyle\frac{1}{np} \sum_{i,j} a_{ij} \\ x_i & = & \displaystyle\left(\frac{1}{p} \sum_j a_{ij} \right) - m \\ y_j & = & \displaystyle\left(\frac{1}{n} \sum_i a_{ij} \right) - m \end{array} \right. \tag{3.7.12}$$

Proof. We have

$$\sum_{i,j} a_{ij} = np\, m$$

so

$$m = \frac{1}{np} \sum_{ij} a_{ij}$$

Then

$$\sum_i a_{ij} = n\, m + n y_j$$

and

$$y_j = \left(\frac{1}{n} \sum_i a_{ij} \right) - m \tag{3.7.13}$$

Similarly

$$\sum_j a_{ij} = p\, m + p x_i$$

and

$$x_i = \left(\frac{1}{p} \sum_j a_{ij} \right) - m \tag{3.7.14}$$

□

So, we have

$$\begin{aligned} a_{ij} &= m + x_i + y_i + r_{ij} \\ &= m + \left(\frac{1}{p} \sum_j a_{ij} \right) - m + \left(\frac{1}{n} \sum_i a_{ij} \right) - m + r_{ij} \\ &= -\frac{1}{np} \sum_{i,j} a_{ij} + \left(\frac{1}{p} \sum_j a_{ij} \right) + \left(\frac{1}{n} \sum_i a_{ij} \right) + r_{ij} \end{aligned} \tag{3.7.15}$$

which is denoted as well

$$r_{ij} = a_{ij} - a_{i.} - a_{.j} + a_{..} \tag{3.7.16}$$

- `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 Σ, which is standard in statistics[4]: $\Sigma = X^{\mathrm{T}} X$ as X is centered. There is no need for the random variables X_j to be Gaussian,

[4] Σ is the standard notation for the diagonal matrix of singular values of a givan 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 $\mathbb{E}(Xu) = \sum_j u_j \mathbb{E}(X_j) = 0$, and its variance is

$$\begin{aligned} \operatorname{var} Xu &= \mathbb{E}(Xu)^2 \\ &= \langle Xu\,,\, Xu \rangle \qquad \text{as } X \text{ is centered} \\ &= \langle u, X^{\mathrm{T}} X u \rangle \\ &= \langle u, \Sigma u \rangle \end{aligned}$$

Maximizing $\operatorname{var} Xu$ with the constraint $\|u\| = 1$ yields that u is the eigenvector of Σ 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 $\mathbb{E}(X) = 0$. Let us denote $\operatorname{var} X = \Sigma$. Then, there exists an rotation

$$V = Xu$$

such that the covariance matrix of V is diagonal and the kth 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

$$z \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_r) \tag{4.1.1}$$

The probability distribution of observed variables $x \in \mathbb{R}^p$ is given conditionally to a choice for the latent variables, as

$$\mathbb{P}(x \mid z) \sim \mathcal{N}(Wz + \mu\,,\, \sigma^2 \mathbb{I}_p) \tag{4.1.2}$$

where $W \in \mathbb{R}^{p \times r}$ and $\mu \in \mathbb{R}^p$. Then, the marginal distribution of x is given by

$$p(x) = \int_z p(x \mid z) p(z)\, dz \tag{4.1.3}$$

which can be shown to follow (see [Bis06, p.573])

$$\mathbb{P}(x) \sim \mathcal{N}(\mu\,,\, WW^{\mathrm{T}} + \sigma^2 \mathbb{I}_p) \tag{4.1.4}$$

Next step is to estimate the parameters (W, μ, σ) of the model kowing the observations. This is not so easy, and is presented in [Bis06, sect. 12.2.1].

Notes and references: A standard textbook for statistical appproach 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

$$x = \Omega z + \mu + u \tag{4.2.1}$$

It is assumed that

→ $\mathbb{E}(z) = \mathbf{0}$

→ $\mathbb{E}(u) = \mathbf{0}$

→ $\mathbb{E}(zz^{\mathrm{T}}) = M$

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

→ $\mathbb{E}(zu^{\mathrm{T}}) = \mathbf{0}$

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

$$\operatorname{var} x = \mathbb{E}((x-\mu)(x-\mu)^{\mathrm{T}}) = \Omega M \Omega^{\mathrm{T}} + \Psi \tag{4.2.2}$$

A set of observatons being given, the parameters of the model (Ω, M, Ψ, μ) 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

$$\mathbb{P}(x \mid z) \sim \mathcal{N}(Wz + \mu\,,\ \Psi) \tag{4.2.3}$$

where Ψ 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 Ψ 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 ℓ^1 and ℓ^∞ norms, on top of ℓ^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. $\mathbb{U}(\mathbb{C}^n)$ denotes the set of unitary matrices in $\mathbb{C}^n$, i.e. matrices having the property $UU^* = U^*U = \mathbb{I}_n$, where $U^* = \overline{U^{\mathrm{T}}}$ (and the same for $\mathbb{C}^p$). The equivalent in $\mathbb{R}^n$ is the set $\mathbb{O}(\mathbb{R}^n)$ of orthogonal matrices such that $U^{\mathrm{T}}U = \mathbb{I}_n$.

- Let E, F be two complex vector spaces, $A \in E \otimes F \simeq \mathcal{L}(F, E)$. A norm $\|.\|$

$$E \otimes F \xrightarrow{\|.\|} \mathbb{R}^+$$

is said unitarily invariant if

$$\forall \begin{cases} U & \in \mathbb{U}(\mathbb{C}^n) \\ V & \in \mathbb{U}(\mathbb{C}^p) \\ A & \in \mathbb{C}^{n\times p} \end{cases}, \qquad \|UAV^*\| = \|A\| \tag{4.3.1}$$

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

$$\forall \begin{cases} U & \in \mathbb{O}(\mathbb{R}^n) \\ V & \in \mathbb{O}(\mathbb{R}^p) \\ A & \in \mathbb{R}^{n\times p} \end{cases}, \qquad \|UAV^{\mathrm{T}}\| = \|A\| \tag{4.3.2}$$

- **Symmetric gauge function (SGF):** A norm

$$\mathbb{R}^n \xrightarrow{\Phi} \mathbb{R}^+$$

is called a symmetric gauge function if it is invariant by any permutation of the coordinates in $\mathbb{R}^n$, i.e., if $\mathscr{P}$ is the set of permutations in $\mathbb{R}^n$

$$\forall\, P \in \mathscr{P}, \quad \Phi(Px) = \Phi(x) \tag{4.3.3}$$

- 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 Φ be a SGF. To each SGF Φ, one associates the norm $\|.\|_\Phi$ defined by

$$\|A\|_\Phi = \Phi(\sigma_1, \ldots, \sigma_n) \tag{4.3.4}$$

Then, $\|.\|_\Phi$ is a unitarily invariant norm. Let us note that if $\|.\|$ is a UIN, A a matrix in $\mathbb{C}^{n\times p}$ and $A = U\Sigma V^*$ the SVD of A, then $\|A = U^*AV\|$ and, as $U^*AV = \Sigma$, $\|A\| = \|\Sigma\|$ 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 $\|.\|$

$$\|\text{diag}\,(\alpha_1 - \beta_1, \ldots, \alpha_p - \beta_p)\| \leq \|A - B\| \tag{4.3.5}$$

- **Schmidt-Mirsky theorem:** Let

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

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

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

  → Φ a Symmetric Gauge Function with $\|.\|_\Phi$ as associated unitarily invariant norm

Then

$$\|A - B\|_\Phi \geq \Phi(0, \ldots, 0, \sigma_{r+1}, \ldots, \alpha_p) \tag{4.3.6}$$

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

$$A_r = U\Sigma_r V^* \tag{4.3.7}$$

where Σ_r is the diagonal matrix obtained from Σ be setting to 0 all singular values beyond r.

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

Notes and references: The link between UIN and SGF has been shown in J. von Neumann (1937), Some Matrix Inequalitues and Metrization of Matrix Spaces, *Tomsk Univ. Rev.*, **1**286-300. 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

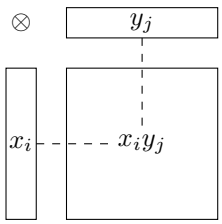
Here, we use tensor notation for `PCA` . For sake of clarity for those readers not familiar with those notations, we set the problem with classical notations like $xy^\top$ for a rank one matrices as a starting point. Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$. We can accept as a definition that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

$$\begin{array}{ccc} \mathbb{R}^n \times \mathbb{R}^p & \xrightarrow{\otimes} & \mathbb{R}^{n\times p} \\ (x, y) & \longrightarrow & x \otimes y \end{array}$$

defined by

$$x \otimes y := xy^\top$$

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

$$A_r = \sum_{a=1}^{r} y_a v_a^{\mathrm{T}}, \qquad y_a = A v_a$$

or, equivalently

$$A_r = \sum_{a} y_a \otimes v_a, \qquad \text{with } \begin{cases} y_a & \in \mathbb{R}^n \\ v_a & \in \mathbb{R}^p \end{cases}$$

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 respectivily $\mathbb{R}^p$ (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

given	$A \in \mathbb{R}^{n \times p}$
	$E \subset \mathbb{R}^n, \quad F \subset \mathbb{R}^p$
	$\dim E = m$, $\dim F = q$
	$0 < r < \min(m, q)$
find	$A_r \in E \otimes F$
such that	$\|A - A_r\|$ minimum

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 built 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 $\mathcal{P}_{\text{E}}$. Let $U = (u_1, \ldots, u_m)$ be an orthonormal basis of E. Then

$$\mathcal{P}_{\text{E}} = UU^{\text{T}} \tag{5.2.1}$$

Indeed, we have

$$\begin{aligned}\mathcal{P}_{\text{E}}x &= \sum_i \langle u_i, x\rangle\, u_i \\ &= \sum_i (u_i \otimes u_i).x \\ &= \left(\sum_i u_i \otimes u_i\right).x\end{aligned}$$

Then

$$\mathcal{P}_{\text{E}} = \sum_i u_i \otimes u_i = UU^{\text{T}} \tag{5.2.2}$$

• Let $\mathcal{P}_{\text{E}}$ be the projector on E, $\mathcal{P}_{\text{F}}$ be the projector on F, and $A \in \mathbb{R}^{n\times p}$. Then, the projector $\mathcal{P}_{E\otimes F}$ of $E \otimes F$ is defined by

$$\mathcal{P}_{E\otimes F}A = UU^{\text{T}}AVV^{\text{T}} \tag{5.2.3}$$

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

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

$$\begin{aligned}\|A - A_r\|^2 &= \left\|\left(\mathcal{P} + \mathcal{P}^\perp\right)(A - A_r)\right\|^2 && \text{as} \quad \mathbb{I} = \mathcal{P} + \mathcal{P}^\perp \\ &= \left\|\mathcal{P}(A - A_r) + \mathcal{P}^\perp(A - A_r)\right\|^2 \\ &= \|\mathcal{P}(A - A_r)\|^2 + \left\|\mathcal{P}^\perp(A - A_r)\right\|^2 && \text{by} \quad \text{Pythagore} \\ &= \|\mathcal{P}(A - A_r)\|^2 + \left\|\mathcal{P}^\perp(A)\right\|^2 && \text{as} \quad \mathcal{P}^\perp A_r = 0 \\ &= \|\mathcal{P}(A) - A_r\|^2 + \left\|\mathcal{P}^\perp(A)\right\|^2 && \text{as} \quad \mathcal{P}A_r = A_r\end{aligned}$$

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

• This leads to a solution for `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_i and W be the $p \times q$ matrix with columns w_j. Then, the orthogonal projector $\mathcal{P}_{\text{F}}$ from $\mathbb{R}^p$ to F is given by matrix

$$P_{\text{F}} = WW^{\text{T}}$$

and the orthogonal projector from $\mathbb{R}^n$ to E is given by matrix

$$P_{\text{E}} = UU^{\text{T}}$$

Let $A \in \mathbb{R}^{n\times p}$. Then

$$\mathcal{P}(A) = UU^{\mathrm{T}}AWW^{\mathrm{T}} \tag{5.2.4}$$

Hence the algorithm

Algorithm 6 Pseudocode for `PCAiv`(A, U, V)

input: $A \in \mathbb{R}^{n\times p}$
input: $E = \operatorname{span} U \subset \mathbb{R}^n$, U orthonormal
input: $F = \operatorname{span} W \subset \mathbb{R}^p$, W orthonormal
input $r < p$
compute $R = UU^{\mathrm{T}}$
compute $S = WW^{\mathrm{T}}$
compute $M = RAS$
compute $(Y, \Lambda, V) = \text{PCA\_CORE}(M)$
return Y, Λ, V

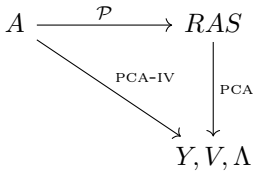
Notes and references: Apparently, the term "PCA with Instrumental Variables" appeared first in [Rao64]. It has been studied with a double set of constraints in [Sab84].

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:



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

$$\|Y_r\| = \rho_r \|RAS\| \tag{5.3.1}$$

i.e. the quality of the `PCA` is denoted by ρ_r at rank r.

- The quality of the projection can be quantified by

$$\|RAS\| = \theta \|A\| \tag{5.3.2}$$

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

- We then have
$$\|Y_r\| = \rho_r \theta \|A\| \tag{5.3.3}$$

and the quality of the `PCAiv` can be poor for two reasons:

→ the quality θ of the projection is poor

→ the quality ρ_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. θ will be low, but ρ_r close to 1. Even if the quality $\rho_r\theta$ is poor because of the poor quality θ of the projection, `PCAiv` 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

$$R = U(U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}, \qquad S = W(W^{\mathrm{T}}W)^{-1}W^{\mathrm{T}}$$

and equation (5.2.3) reads

$$\begin{aligned} M &= RAS \\ &= U(U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}AW(W^{\mathrm{T}}W)^{-1}W^{\mathrm{T}} \end{aligned}$$

- If there are no constraints on the components $(y_i)_i$, which can be set by selecting $U = \mathbb{I}_n$. Then
$$M = AW(W^{\mathrm{T}}W)^{-1}W^{\mathrm{T}}$$

Symetrically, the absence of constrains on $(x_i)_i$ can be set by selecting $W = \mathbb{I}_p$, and

$$M = U(U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}A$$

- There is another way to run `PCAiv` 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 $R = U'U'^{\mathrm{T}}$ and $S = WW'^{\mathrm{T}}$

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

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

$$\begin{aligned}\forall\, A, B \in \mathbb{R}^{n\times p}, \quad \langle A, B\rangle &= \sum_{i,j} \alpha_{ij}\beta_{ij} \\ &= \operatorname{Tr} A^{\mathrm{T}} B \\ &= \operatorname{Tr} B^{\mathrm{T}} A\end{aligned}$$

Any $np \times np$ symmetric definite positive matrix T defines an inner product on $\mathbb{R}^{n\times p}$. Componentwise, it is defined as

$$\langle A\,,\, B\rangle_{\mathrm{T}} = \sum_{ij,k\ell} T_{ij,k\ell}\, \alpha_{ij}\beta_{k\ell}$$

Then, $\mathbb{R}^{n\times p}$ is endowed with a Euclidean structure, which induces a norm, which induces a metric structure with

$$d_{\mathrm{T}}(A, B) = \|A - B\|_{\mathrm{T}}$$

Then, linear dimension reduction can be performed, e.g. on a matrix A. In the algebraic framework, it consists in finding the rank r matrix in $\mathbb{R}^{n\times p}$ which is the closest to A with the distance induced by T. The geometric framework consists in finding an affine subspace of $\mathbb{R}^{n\times p}$ of dimension r on which the projection of the point clod associated to A is optimal. The statistical viewpoint will not be developed here.

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.

$$\langle x, y\rangle_{\mathrm{P}} = \langle x\,,\, Py\rangle \tag{6.1.1}$$

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

$$\langle x, y\rangle_{\mathrm{P}} = \sum_{i,j=1}^{p} p_{ij}\, x_i\, y_j, \qquad p_{ji} = p_{ij} \tag{6.1.2}$$

if $x = (x_i)_i$, $y = (y_j)_j$ and $P = (p_{ij})_{i,j}$. If $P = \mathbb{I}_p$, canonical inner product is recovered, as $p_{ij} = \delta_i^j$. A case worth being studied in details is when P is diagonal, i.e. $P = \operatorname{diag} w$ 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

$$\langle x, y\rangle_{\mathrm{P}} = \sum_{i=1}^{p} w_i\, x_i\, y_i \tag{6.1.3}$$

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

$$\langle x, y \rangle_{\mathrm{P}} = \langle Qx \, , \, Qy \rangle \tag{6.1.4}$$

and

$$\|x\|_{\mathrm{P}} = \|Qx\| \tag{6.1.5}$$

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 ι_{Q} between $(\mathbb{R}^p, Q)$ and $(\mathbb{R}^p, \mathbb{I})$ by

$$\begin{array}{ccc} (\mathbb{R}^p, Q) & \xrightarrow{\iota_{\mathrm{Q}}} & (\mathbb{R}^p, \mathbb{I}) \\ x & \longrightarrow & Qx \end{array}$$

as

$$\|x\|_{\mathrm{Q}} = \|Qx\| = \|\iota_{\mathrm{Q}} x\|$$

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 $x, x' \in \mathbb{R}^p$ with weights w is given by

$$d_w(x, x') = \|x - x'\|_w = \sqrt{\|x - x'\|_w^2} = \sqrt{\sum_i w_i (x_i - x'_i)^2}$$

This is consistent with an isometry defined by

$$Q = \mathrm{diag} \, \left(\sqrt{w_1}, \ldots, \sqrt{w_p} \right)$$

as

$$d_w(x, x') = \sqrt{\sum_i \left(\sqrt{w_i} x_i - \sqrt{w_i} x'_i \right)^2}$$

It is customary to define the weights by w, which are the diagonal elements of P, and not $\sqrt{w}$. Hence, it is consistent with this classical approach to define the metric by P, hence denote $\langle x, x' \rangle_P$. We will use P or Q indifferently, knowing that $P = Q^2$.

- A metric can be defined in the same way in $\mathbb{R}^n$, the column space. It is given by a SDP matrix $N \in \mathbb{R}^{n \times n}$. We denote $N = M^2$, and

$$\langle y, y' \rangle_{\mathrm{N}} = \langle My, My' \rangle, \qquad \|y\|_{\mathrm{N}} = \|My\|$$

- Let P define a metric on $\mathbb{R}^p$ and N a metric on $\mathbb{R}^n$. An inner product on $\mathbb{R}^{n \times p}$ will be defined, which is canonically associated to N and P and such that the map

$$(\mathbb{R}^{n \times p}, T) \longrightarrow (\mathbb{R}^{n \times p}, \mathbb{I})$$

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

$$\langle a \otimes b\,,\, x \otimes y \rangle_{\mathrm{T}} = \langle a, x \rangle_{\mathrm{N}} \langle b, y \rangle_{\mathrm{P}}$$

with

$$\langle a \otimes b\,,\, x \otimes y \rangle_{\mathrm{T}} = \langle Z(a \otimes b)\,,\, Z(x \otimes y) \rangle \tag{6.1.6}$$

As

$$\begin{array}{rcl} \langle a \otimes b\,,\, x \otimes y \rangle_{\mathrm{T}} & = & \langle Ma, Mx \rangle \langle Qb, Qy \rangle \\ & = & \langle Ma \otimes Qb\,,\, Mx \otimes Qy \rangle \end{array} \tag{6.1.7}$$

(6.1.6) is fulfilled by selecting

$$\begin{array}{rcl} Z(a \otimes b) & = & Ma \otimes Qb \\ & = & M(a \otimes b)Q \end{array} \tag{6.1.8}$$

and, for any matrix $A \in \mathbb{R}^{n \times p}$

$$ZA = MAQ \tag{6.1.9}$$

by linearity.

- So, we have

$$\|A\|_{\mathrm{T}} = \|MAQ\| \tag{6.1.10}$$

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 ι^{-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 mentionned 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 $A \in \mathbb{R}^n \otimes \mathbb{R}^p$
	a rank $r < p$
find	a matrix $A_r \in \mathbb{R}^n \otimes \mathbb{R}^p$ 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	$N = M^2, \quad P = Q^2$
	a rank	$0 < r < p$
Define	the inner product T on $\mathbb{R}^{n\times p}$	$T = Z^2$
	associated to (N, P)	$ZA = MAQ$
Find	a matrix	$A_r \in \mathbb{R}^{n\times p}$
with		$\operatorname{rank} A_r = r$
such that		$\|A - A_r\|_{\mathrm{T}}$ 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

$$A_r = \sum_{i=1}^{r} y_i \otimes v_i$$

with $(v_i)_i$ being an orthonormal family for the inner product in $\mathbb{R}^p$ induced by P

$$\langle v_i \, , \, v_j \rangle_{\mathrm{P}} = \delta_i^j$$

Then

$$\|A - A_r\| = \left\| A - \sum_i y_i \otimes v_i \right\|$$

- Let us select the inner product $T = N \otimes P$ on $\mathbb{R}^{n\times p}$ as

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

Then

$$\|A\|_{\mathrm{N,P}} = \|MAQ\|$$

We recall that $M(y \otimes v)Q = My \otimes Q^{\mathrm{T}}v = My \otimes Qv$ as Q is symmetric. Then

$$\begin{aligned} \|A - A_r\|_{\mathrm{N,P}} &= \left\| M \left(A_r - \sum_i y_i \otimes v_i \right) Q \right\| \\ &= \left\| MAQ - \sum_i My_i \otimes Qv_i \right\| \end{aligned} \tag{6.3.1}$$

- Let us denote

$$\begin{cases} My_i &= w_i \\ Qv_i &= x_i \end{cases} \tag{6.3.2}$$

We have $\|x_i\| = \|Qv_i\| = 1$ as $\|v_i\|_{\text{P}} = 1$ and similarily $\langle x_i, x_j \rangle = \langle Qv_i, Qv_j \rangle = \delta_i^j$. Then, the $(x_i)_i$ are an orthonormal family for the canonical inner product. The problem can be formulated as

$$\left| \begin{array}{ll} \text{find} & (x_i)_i, \quad (w_i)_i \\ \text{with} & (x_i)_i \text{ an orthonormal family} \\ \text{such that} & \left\| MAQ - \sum_i w_i \otimes x_i \right\| \quad \text{minimal} \end{array} \right.$$

Then, $\{(w_i)_i, (x_i)_i\}$ are the solution of the PCA of MAQ. Th components $(y_i)_i$ and new basis vector $(v_i)_i$ of the PCA with metrics can be recovered simply by

$$\begin{cases} v_i &= Q^{-1}x_i \\ y_i &= M^{-1}w_i \end{cases} \tag{6.3.3}$$

Hence the algorithm:

Algorithm 7 PCA of a matrix with double metrics: PCA_MET(A, M, Q)

1: **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: **compute** $B = MAQ$
3: **compute** $W, \Lambda, X =$ PCA_CORE(B)
4: **compute** $Y = M^{-1}W$
5: **compute** $V = Q^{-1}X$
6: **return** Y, Λ, V

Remark: The metrics on $\mathbb{R}^n$ and $\mathbb{R}^p$ are given respectively by N and P, which are symmetric, definite and positive (SDP). The matrices involved in this algorithm are respectively M and Q, with $M = N^{1/2}$ and $Q = P^{1/2}$. They can be computed from a SVD of respectively N and P. As N is symmetric, its SVD reads

$$N = U\Sigma U^{\text{T}}$$

Then

$$M = U\Sigma^{1/2}U^{\text{T}}$$

Indeed, $M^2 = (U\Sigma^{1/2}U^{\text{T}})(U\Sigma^{1/2}U^{\text{T}}) = U\Sigma^{1/2}U^{\text{T}}U\Sigma^{1/2}U^{\text{T}} = U\Sigma U^{\text{T}} = N$.

6.4 Isometry

This result can be derived without calculation by transportation of `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

$$\begin{array}{ccc} (\mathbb{R}^n, N) & \longrightarrow & (\mathbb{R}^n, \mathbb{I}_n) \\ y & \longrightarrow & My \end{array}$$

and

$$\begin{array}{ccc} (\mathbb{R}^p, P) & \longrightarrow & (\mathbb{R}^p, \mathbb{I}_p) \\ v & \longrightarrow & Qv \end{array}$$

are isometries, as

$$\left\{ \begin{array}{rcl} \langle Qv\,,\, Qv' \rangle & = & \langle v, v' \rangle_{\mathrm{P}} \\ \langle My\,,\, My' \rangle & = & \langle y, y' \rangle_{\mathrm{N}} \end{array} \right.$$

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

$$\begin{array}{ccccc} \psi & : & (\mathbb{R}^n \otimes \mathbb{R}^p, N \otimes P) & \longrightarrow & (\mathbb{R}^n \otimes \mathbb{R}^p, \mathbb{I}_n \otimes \mathbb{I}_p) \\ & & A & \longrightarrow & MAQ \end{array}$$

as

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

`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 $(y_j \otimes v_j)_{1 \leq j \leq k}$ such that

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

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 ψ is an isometry). We have

$$\begin{aligned} \psi(A) - \psi\left(\sum_j y_j \otimes v_j \right) &= MAQ - M\left(\sum_j y_j \otimes v_j \right) Q \\ &= MAQ - \sum_j My_j \otimes Qv_j \end{aligned}$$

Then, $\sum_{j \leq r} My_j \otimes Qv_j$ 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 ψ^{-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, Λ) 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, $M = \mathbb{I}_n$ and

$$MAQ = AQ$$

Hence

$$\begin{array}{rcl} B^{\mathrm{T}}B & = & (MAQ)^{\mathrm{T}}(MAQ) \\ & = & (AQ)^{\mathrm{T}}(AQ) \\ & = & Q^{\mathrm{T}}A^{\mathrm{T}}AQ \end{array} \tag{6.5.1}$$

Similarly, if the metrics are weights on the rows only, $Q = \mathbb{I}_p$ and

$$MAQ = MA$$

Hence

$$\begin{array}{rcl} B^{\mathrm{T}}B & = & (MAQ)^{\mathrm{T}}(MAQ) \\ & = & (MA)^{\mathrm{T}}(MA) \\ & = & A^{\mathrm{T}}MMA \\ & = & A^{\mathrm{T}}NA \end{array} \tag{6.5.2}$$

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

$$B^{\mathrm{T}}B = QA^{\mathrm{T}}NAQ \tag{6.5.3}$$

• **A remark about the calculation:** Principal components $(y_i)_i$ and principal axis (v_i) are solution of

$$\left\{ \begin{array}{rcl} B & = & MAQ \\ B^{\mathrm{T}}Bw_i & = & \lambda_i w_i \\ x_i & = & Bw_i \\ v_i & = & Q^{-1}w_i \\ y_i & = & M^{-1}x_i \end{array} \right. \tag{6.5.4}$$

with

$$x_i, y_i \in \mathbb{R}^n, \qquad v_i, w_i \in \mathbb{R}^p$$

We have

$$\left. \begin{array}{rcl} B^{\mathrm{T}}B & = & QA^{\mathrm{T}}NAQ \\ B^{\mathrm{T}}Bw_i & = & \lambda_i w_i \\ w_i & = & Qv_i \end{array} \right\} \implies QA^{\mathrm{T}}NAPv_i = \lambda_i Qv_i \tag{6.5.5}$$

and, as Q is invertible

$$A^{\mathrm{T}}NAPv_i = \lambda_i v_i \tag{6.5.6}$$

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 $B^{\mathrm{T}}B$ is symmetric, whereas matrix $A^{\mathrm{T}}NAP$ 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 $B^{\mathrm{T}}Bw_i = \lambda_i w_i$ 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

$$\left(\sum_i w_i \right) g = \sum_i w_i a_i \tag{6.5.7}$$

or

$$g = \frac{1}{w} \sum_i w_i a_i, \qquad w = \sum_i w_i \tag{6.5.8}$$

The centered cloud is the cloud with points

$$\overline{a}_i = a_i - g \tag{6.5.9}$$

One checks that

$$\begin{aligned}\sum_i w_i \overline{a}_i &= \sum_i w_i a_i - \left(\sum_i w_i\right) g \\ &= wg - wg \\ &= \mathbf{0}\end{aligned}$$

• **Geometric approach: attached point cloud:** Let $\mathcal{A}$ 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 $\mathcal{B}$ 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

$$MAQ = [\sqrt{\nu_i \pi_j}\, \alpha_{ij}]_{i,j}$$

and, in $\mathbb{R}^p$

$$\begin{aligned} d^2(b_i, b_k) &= \sum_j (\sqrt{\nu_i \pi_j}\, \alpha_{ij} - \sqrt{\nu_k \pi_j}\, \alpha_{kj})^2 \\ &= \sum_j \pi_j \, (\sqrt{\nu_i}\, \alpha_{ij} - \sqrt{\nu_k}\, \alpha_{kj})^2 \end{aligned} \tag{6.5.10}$$

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 $\mathcal{A}_{\mathrm{M}}$ 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.

$$\sum_i a_i = 0 \tag{6.5.11}$$

The variances (or norms) of columns of A can vary significantly. In such a case, the variance/covariance matrix $\Sigma = A^{\mathrm{T}} A$ 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

$$a_{\bullet j} \longrightarrow a'_{\bullet j} = \frac{a_{\bullet j}}{\|a_{\bullet j}\|} \tag{6.5.12}$$

Hence

$$\|a'_{\bullet j}\| = 1 \tag{6.5.13}$$

This can be read as a `PCA` with inner product $N = \mathbb{I}_n$ in $\mathbb{R}^n$ and $P = \mathrm{diag}\left(\frac{1}{\|a_{\bullet j}\|}\right)$ 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].

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 $\mathbb{R}^n$:

$$\begin{cases} y_j \in E \subset \mathbb{R}^n \\ v_j \in F \subset \mathbb{R}^p \end{cases} \tag{6.6.1}$$

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

$$\mathbb{R}^n \xrightarrow{R=UU^{\mathrm{T}}} E, \qquad \mathbb{R}^p \xrightarrow{S=VV^{\mathrm{T}}} F \tag{6.6.2}$$

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 $\|v\|_{\mathrm{P}} = 1$. The projection x' of x on $F = \mathbb{R}v$ is given by

$$\begin{aligned} x' &= \langle x, v\rangle_{\mathrm{P}} v \\ &= \langle x, Pv\rangle v \\ &= \langle Pv, x\rangle v \\ &= (v \otimes Pv).x \end{aligned} \tag{6.6.3}$$

Hence, the projector on $\mathbb{R}v$ with the inner product defined by P is $v \otimes Pv$. If $F = \mathrm{span}(v_1, \ldots, v_r)$, we have

$$\begin{aligned} x' &= \sum_a \langle x, Pv_a\rangle v_a \\ &= \sum_a (v_a \otimes Pv_a)x \end{aligned} \tag{6.6.4}$$

and the projector is

$$\begin{aligned} S &= \sum_a v_a \otimes Pv_a \\ &= V(PV)^{\mathrm{T}} \\ &= VV^{\mathrm{T}}P \\ &= PVV^{\mathrm{T}} \end{aligned} \tag{6.6.5}$$

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^{T} are symmetric, hence $(VV^{\mathrm{T}})P$ is symmetric and $VV^{\mathrm{T}}P = (VV^{\mathrm{T}}P)^{\mathrm{T}} = PVV^{\mathrm{T}}$.

- Similarly, we have

$$R = UU^{\mathrm{T}}N \tag{6.6.6}$$

and

$$\begin{array}{rcl} A' & = & RAS \\ & = & UU^{\mathrm{T}}NAPVV^{\mathrm{T}} \end{array} \tag{6.6.7}$$

- 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

$$A'' = MA'Q, \qquad \text{with} \quad N = M^2, \quad P = Q^2$$

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 χ^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 χ^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

$$A = \frac{T}{T_{++}}, \qquad T_{++} = \sum_{i,j} T_{ij} \tag{7.1.1}$$

The general term in A is denoted α_{ij} and we have

$$A \in \mathbb{R}^{I\times J} \simeq \mathbb{R}^I \otimes \mathbb{R}^J \tag{7.1.2}$$

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

$$\begin{cases} r_i = \displaystyle\sum_j \alpha_{ij} \\ c_j = \displaystyle\sum_i \alpha_{ij} \end{cases} \tag{7.1.3}$$

Let us denote by D_r and D_c the square diagonal matrices with diagonal respectively r and c:

$$D_r = \operatorname{diag} r, \qquad D_c = \operatorname{diag} c \tag{7.1.4}$$

- In case of independence between both variables, the expectation for A is

$$\widetilde{A} = r \otimes c \tag{7.1.5}$$

Indeed, we have, ignoring the value of the other variable

$$P(X_r = i) = r_i, \qquad P(X_c = j) = c_j$$

Then, in case of independence

$$P(X_r = i\,;\, X_c = j) = r_i c_j$$

Then, a first observation is that

$$\chi^2(A) = \|A - \widetilde{A}\|_{D_r^{-1} \otimes D_c^{-1}} \tag{7.1.6}$$

Proof. Indeed, we have

$$\begin{aligned}
\chi^2(A) &= \sum_{i,j} \frac{(\alpha_{ij} - r_i c_j)^2}{r_i c_j} \\
&= \sum_{i,j} \left(\frac{\alpha_{ij} - r_i c_j}{\sqrt{r_i c_j}} \right)^2 \\
&= \sum_{i,j} \left(\frac{1}{\sqrt{r_i}} (\alpha_{ij} - r_i c_j) \frac{1}{\sqrt{c_j}} \right)^2 \\
&= \left\| D_r^{-1/2} (A - r \otimes c) D_c^{-1/2} \right\|^2 \\
&= \|A - r \otimes c\|^2_{D_r^{-1} \otimes D_c^{-1}}
\end{aligned} \tag{7.1.7}$$

□

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.

given	T (contingency table)
compute	$T_{++} = \sum_{i,j} T_{ij}$ $A = \frac{T}{T_{++}}$ $r_i = \sum_j \alpha_{ij}$ $c_j = \sum_i \alpha_{ij}$
run	`PCAmet`
on	$A - r \otimes c$
with diagonal metrics	$1/r_i$ on row i $1/c_j$ on column j

We have

$$\begin{cases} M = \operatorname{diag} 1/\sqrt{r_i} \\ Q = \operatorname{diag} 1/\sqrt{c_j} \end{cases}$$

Then

$$\begin{aligned} A_{\text{M,Q}} &= M(A - r \otimes c)Q \\ &= \left[\frac{\alpha_{ij} - r_i c_j}{\sqrt{r_i c_j}}\right]_{i,j} \end{aligned} \tag{7.2.1}$$

This yields the following algorithm:

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

input $T \in \mathbb{R}^{I \times J}$, a contingency table
compute $A = T/T_{++}$, with $T_{++} = \sum_{i,j} T_{ij}$
compute $r_i = \sum_j \alpha_{ij}$, $D_r = \operatorname{diag} r$
compute $c_j = \sum_i \alpha_{ij}$, $D_c = \operatorname{diag} c$
compute $M = \operatorname{diag}(1/\sqrt{r_i})$
compute $Q = \operatorname{diag}(1/\sqrt{c_j})$
compute $A_{\text{M,Q}} = M(A - r \otimes c)Q = \left[\frac{\alpha_{ij} - r_i c_j}{\sqrt{r_i c_j}}\right]_{i,j}$
compute $Z, X, \Lambda = \text{PCA\_CORE}(A_{\text{M,Q}})$
compute $Y_r = M^{-1} Z$
compute $Y_c = Q^{-1} X$
return Y_r, Y_c, Λ

7.3 CoA and geometry of point clouds

Let us consider the point cloud $\mathcal{X}$ of I points in $\mathbb{R}^{\text{J}}$ of coordinates

$$X_{ij} = \frac{\alpha_{ij}}{\sqrt{r_i c_j}} \tag{7.3.1}$$

in a Euclidean space with standard inner product (i.e. $\mathbb{I}_{\text{J}}$). Then, if $x_i, x_{i'}$ are two points in $\mathcal{X}$, we have

$$\begin{aligned} d(x_i, x_{i'})^2 &= \sum_j \left(\frac{\alpha_{ij}}{\sqrt{r_i c_j}} - \frac{\alpha_{i'j}}{\sqrt{r_{i'} c_j}}\right)^2 \\ &= \sum_j \frac{1}{c_j}\left(\frac{\alpha_{ij}}{\sqrt{r_i}} - \frac{\alpha_{i'j}}{\sqrt{r_{i'}}}\right)^2 \end{aligned} \tag{7.3.2}$$

It is standard to say that point cloud $\mathcal{X} = (x_i)_i$ is embedded with weights $1/c_j$ for column (= variable) j and metrics defined by diagonal matrix of term $1/r_i$ in $\mathbb{R}^I$.

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 slighltly 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 $\mathbb{R}^{\mathrm{J}}$, with point x_i of coordinates

$$x_i = \left[\frac{\alpha_{ij}}{r_i}\right]_j, \qquad r_i = \sum_j \alpha_{ij} \tag{7.4.1}$$

  $\rightarrow$ a cloud of column profiles, as J points x_j^{T} in $\mathbb{R}^{\mathrm{I}}$, with point x_j^{T} of coordinates

$$x_j^{\mathrm{T}} = \left[\frac{\alpha_{ij}}{c_j}\right]_i, \qquad c_i = \sum_i \alpha_{ij} \tag{7.4.2}$$

- Then, metrics with diagonal matrices respectively D_c^{-1} for R^{J} and D_r^{-1} for $\mathbb{R}^{\mathrm{I}}$ are selected. Hence, distances between points x_i and x_i' in $\mathbb{R}^{\mathrm{J}}$ are computed as

$$d^2(x_i, x_i') = \sum_j \frac{1}{c_j} \left(\frac{\alpha_{ij}}{r_i} - \frac{\alpha_{i'j}}{r_{i'}}\right)^2 \tag{7.4.3}$$

and between points x_j^{T} and $x_{j'}^{\mathrm{T}}$ in $\mathbb{R}^{\mathrm{I}}$ as

$$d^2(x_j^{\mathrm{T}}, x_{j'}^{\mathrm{T}}) = \sum_i \frac{1}{r_i} \left(\frac{\alpha_{ij}}{c_j} - \frac{\alpha_{ij'}}{c_{j'}}\right)^2 < \tag{7.4.4}$$

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 $\forall \mathrm{J},\ \alpha_{ij}/r_i = \alpha_{i'j}/r_{i'}$), 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])

	R: row profiles in $\mathbb{R}^J$	C: column profiles in $\mathbb{R}^I$
point cloud	$D_r^{-1}A$	$D_c^{-1}A^{\mathrm{T}}$
metric	D_c^{-1}	D_r^{-1}
weights	D_r	D_c

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

$$\begin{aligned}
I_{\text{R}}^2 &= \sum_i r_i \|(D_r^{-1}A)_i - c\|_P^2 \\
&= \sum_i r_i \left\| \frac{a_i}{r_i} - c \right\|_P^2, && (D_r^{-1}A)_i = \tfrac{a_i}{r_i} \\
&= \sum_i r_i \left(\frac{1}{c_j} \sum_j \left(\frac{a_{ij}}{r_i} - c_j \right)^2 \right), && P = \text{diag}\left(\tfrac{1}{c_j}\right) \\
&= \sum_i r_i \left(\sum_j \left(\frac{a_{ij}}{r_i\sqrt{c_j}} - \sqrt{c_j} \right)^2 \right) && Q = \text{diag}\left(\tfrac{1}{\sqrt{c_j}}\right) \\
&= \sum_{i,j} \left(\sqrt{r_i}\frac{a_{ij}}{r_i\sqrt{c_j}} - \sqrt{r_i}\sqrt{c_j} \right)^2 \\
&= \sum_{i,j} \left(\frac{a_{ij} - r_i c_j}{\sqrt{r_i c_j}} \right)^2
\end{aligned} \tag{7.4.5}$$

• 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 $I = \|X\|_P = \|XQ\|$. 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 $I = \|X\|_N = \|MX\|$. If we have weights on rows and metric in $\mathbb{R}^p$, we have $I = \|X\|_{P\otimes N} = \|MXQ\|$. Let $X = D_r^{-1}A - c \otimes \mathbf{1}_n$, $P = D_c^{-1}$ hence $Q = D_c^{-1/2}$, and $w_r = r$ hence $M = D_r^{1/2}$. Then, knowing that $A(x \otimes y)B = Bx \otimes Ay$

$$\begin{aligned}
I_{\text{R}} &= \|MXQ\| \\
&= \|D_r^{1/2}(D_r^{-1}A - c \otimes \mathbf{1}_n)D_c^{-1/2}\| \\
&= \|D_r^{-1/2}AD_c^{-1/2} - D_r^{1/2}(c \otimes \mathbf{1}_n)D_c^{-1/2}\| \\
&= \|D_r^{-1/2}AD_c^{-1/2} - D_c^{-1/2}c \otimes D_r^{1/2}\mathbf{1}_n\| \\
&= \|D_r^{-1/2}AD_c^{-1/2} - D_c^{-1/2}c \otimes D_r^{-1/2}r\| \\
&= \|D_r^{-1/2}(A - c \otimes r)D_c^{-1/2}\|
\end{aligned}$$

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

$$I_{\text{C}}^2 = \sum_{i,j} \left(\frac{a_{ij} - r_i c_j}{\sqrt{r_i c_j}} \right)^2 = I_{\text{R}}^2 \tag{7.4.6}$$

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 [NG07]. 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 $y_{\text{A}} \in \operatorname{span} A$ (resp. $y_{\text{B}} \in \operatorname{span} B$) can be written as Au_{A} (resp. Bu_{B}). The correlation between y_{A} and y_{B} is

$$\operatorname{corr}(y_{\text{A}}, y_{\text{B}}) = \frac{\langle y_{\text{A}}, y_{\text{B}}\rangle}{\|y_{\text{A}}\|\|y_{\text{B}}\|}$$

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	$u_{\text{A}} \in \mathbb{R}^p$, $u_{\text{B}} \in \mathbb{R}^q$
such that	$\dfrac{\langle Au_{\text{A}}, Bu_{\text{B}}\rangle}{\|Au_{\text{A}}\|\|Bu_{\text{B}}\|}$ maximal

As such the problem is difficult to solve. One reason is that $\|u_{\text{A}}\|$ and $\|u_{\text{B}}\|$ can take any non zero value (the correlation remains unchanged by a rescaling of $\|u_{\text{A}}\|$ or $\|u_{\text{B}}\|$). One could add a constraint like $\|u_{\text{A}}\| = \|u_{\text{B}}\| = 1$, 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 $y_{\text{A}} = Au_{\text{A}}$

(resp. $y_{\text{B}} = Bu_{\text{B}}$):

Given	$A \in \mathbb{R}^{n\times p}$, $B \in \mathbb{R}^{n\times q}$
Find	$u_{\text{A}} \in \mathbb{R}^p$, $u_{\text{B}} \in \mathbb{R}^q$
with	$\|Au_{\text{A}}\| = 1$, $\|Bu_{\text{B}}\| = 1$
such that	$\langle Au_{\text{A}}, Bu_{\text{B}}\rangle$ 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

$$\nabla f - \lambda \nabla g = \mathbf{0}$$

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 $(u_{\text{A}}, u_{\text{B}})$, the function f is $f(u_{\text{A}}, u_{\text{B}}) = \langle Au_a, Bu_b\rangle$ and g is $\|Au_{\text{A}}\| = \|Bu_{\text{B}}\| = 1$. One computes seperately the partial derivatives with resect to u_{A} and u_{B}, denoting them $\nabla_{u_{\text{A}}}$ and $\nabla_{u_{\text{B}}}$. One has

$$\left\{\begin{array}{lcllcl} \nabla_{u_{\text{A}}}\langle Au_{\text{A}}, Bu_{\text{B}}\rangle &=& A^{\text{T}}Bu_{\text{B}} & \nabla_{u_{\text{B}}}\langle Au_{\text{A}}, Bu_{\text{B}}\rangle &=& B^{\text{T}}Au_{\text{A}} \\ \nabla_{u_{\text{A}}}\|Au_{\text{A}}\|^2 &=& 2A^{\text{T}}Au_{\text{A}} & \nabla_{u_{\text{B}}}\|Bu_{\text{B}}\|^2 &=& 2B^{\text{T}}Bu_{\text{B}} \end{array}\right. \tag{8.2.1}$$

Then, the solution satisfies to

$$\left\{\begin{array}{lcl} \nabla_{u_{\text{A}}}: & A^{\text{T}}Bu_{\text{B}} &= \lambda A^{\text{T}}Au_{\text{A}} \\ \nabla_{u_{\text{B}}}: & B^{\text{T}}Au_{\text{A}} &= \mu B^{\text{T}}Bu_{\text{B}} \end{array}\right. \tag{8.2.2}$$

• We first show that $\lambda = \mu$.

Proof. Therefore, we observe that

$$\left\{\begin{array}{lcl} \langle A^{\text{T}}Bu_{\text{B}}, u_{\text{A}}\rangle &=& \lambda\langle A^{\text{T}}Au_{\text{A}}, u_{\text{A}}\rangle \\ \langle B^{\text{T}}Au_{\text{A}}, u_{\text{B}}\rangle &=& \mu\langle B^{\text{T}}B, u_{\text{B}}\rangle \end{array}\right.$$

and that

$$\begin{array}{lcllcl} \langle A^{\text{T}}Au_{\text{A}}, u_{\text{A}}\rangle &=& \langle Au_{\text{A}}, Au_{\text{A}}\rangle &=& 1 \\ \langle B^{\text{T}}Bu_{\text{B}}, u_{\text{B}}\rangle &=& \langle Bu_{\text{B}}, Bu_{\text{B}}\rangle &=& 1 \end{array}$$

Then

$$\lambda = \langle A^{\text{T}}Bu_{\text{B}}, u_{\text{A}}\rangle = \langle B^{\text{T}}Au_{\text{A}}, u_{\text{B}}\rangle = \mu$$

□

- Thus, equation (8.2.2) reads

$$\left\{\begin{array}{rcl} A^{\mathrm{T}}Bu_{\mathrm{B}} & = & \lambda A^{\mathrm{T}}Au_{\mathrm{A}} \\ B^{\mathrm{T}}Au_{\mathrm{A}} & = & \lambda B^{\mathrm{T}}Bu_{\mathrm{B}} \end{array}\right. \tag{8.2.3}$$

Multiplying leftwise the first equation by $(A^{\mathrm{T}}A)^{-1}$ and the second by $(B^{\mathrm{T}}B)^{-1}$ yields

$$\left\{\begin{array}{rcl} (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}Bu_{\mathrm{B}} & = & \lambda u_{\mathrm{A}} \\ (B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}Au_{\mathrm{A}} & = & \lambda u_{\mathrm{B}} \end{array}\right. \tag{8.2.4}$$

and, having in mind that $Au_{\mathrm{A}} = y_{\mathrm{A}}$ and $Bu_{\mathrm{B}} = y_{\mathrm{B}}$

$$\left\{\begin{array}{rcl} A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}y_{\mathrm{B}} & = & \lambda y_{\mathrm{A}} \\ B(B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}y_{\mathrm{A}} & = & \lambda y_{\mathrm{B}} \end{array}\right. \tag{8.2.5}$$

- One recognizes in the l.h.s. of (8.2.5)

$$\left\{\begin{array}{l} A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}} = \mathcal{P}_{\mathrm{A}} \\ B(B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}} = \mathcal{P}_{\mathrm{B}} \end{array}\right.$$

i.e. the projectors on the spaces spanned by the columns of A and of B respectively. This leads to

$$\left\{\begin{array}{rcl} \mathcal{P}_{\mathrm{A}}y_{\mathrm{B}} & = & \lambda y_{\mathrm{A}} \\ \mathcal{P}_{\mathrm{B}}y_{\mathrm{A}} & = & \lambda y_{\mathrm{B}} \end{array}\right.$$

or

$$\left\{\begin{array}{rcl} \mathcal{P}_{\mathrm{A}}\mathcal{P}_{\mathrm{B}}y_{\mathrm{A}} & = & \lambda^2 y_{\mathrm{A}} \\ \mathcal{P}_{\mathrm{B}}\mathcal{P}_{\mathrm{A}}y_{\mathrm{B}} & = & \lambda^2 y_{\mathrm{B}} \end{array}\right. \tag{8.2.6}$$

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

$$\begin{array}{ccccc} \text{span } A & \xrightarrow{\mathcal{P}_{\mathrm{B}}} & \text{span } B & \xrightarrow{\mathcal{P}_{\mathrm{A}}} & \text{span } A \\ y_{\mathrm{A}} & \longrightarrow & y'_{\mathrm{B}} & \longrightarrow & y''_{\mathrm{A}} \end{array}$$

The same can be written for y_{B}:

$$\begin{array}{ccccc} \text{span } B & \xrightarrow{\mathcal{P}_{\mathrm{A}}} & \text{span } A & \xrightarrow{\mathcal{P}_{\mathrm{B}}} & \text{span } B \\ y_{\mathrm{B}} & \longrightarrow & y'_{\mathrm{A}} & \longrightarrow & y''_{\mathrm{B}} \end{array}$$

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 $\mathcal{P}_{\mathrm{A}}\mathcal{P}_{\mathrm{B}}$ (resp. $\mathcal{P}_{\mathrm{B}}\mathcal{P}_{\mathrm{A}}$).

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

$$\left\{ \begin{array}{rcl} \mathcal{P}_{\mathrm{A}}\mathcal{P}_{\mathrm{B}}y_{\mathrm{A}} & = & \lambda^2 y_{\mathrm{A}} \\ \mathcal{P}_{\mathrm{B}}\mathcal{P}_{\mathrm{A}}y_{\mathrm{B}} & = & \lambda^2 y_{\mathrm{B}} \end{array} \right.$$

Let us recall that

$$\left\{ \begin{array}{rcl} \mathcal{P}_{\mathrm{A}} & = & A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}} \\ \mathcal{P}_{\mathrm{B}} & = & B(B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}} \end{array} \right.$$

Then

$$\left\{ \begin{array}{rcl} A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}B(B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}y_{\mathrm{A}} & = & \lambda^2 y_{\mathrm{A}} \\ B(B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}y_{\mathrm{B}} & = & \lambda^2 y_{\mathrm{B}} \end{array} \right. \tag{8.3.1}$$

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

$$\left\{ \begin{array}{ll} T & = A^{\mathrm{T}}B \\ N & = (A^{\mathrm{T}}A)^{-1} \\ P & = (B^{\mathrm{T}}B)^{-1} \end{array} \right. \tag{8.3.2}$$

Then

$$\left\{ \begin{array}{rcl} ANTPB^{\mathrm{T}}y_{\mathrm{A}} & = & \lambda^2 y_{\mathrm{A}} \\ BPT^{\mathrm{T}}NA^{\mathrm{T}}y_{\mathrm{B}} & = & \lambda^2 y_{\mathrm{B}} \end{array} \right.$$

Let us recall that

$$y_{\mathrm{A}} = Au_{\mathrm{A}}, \qquad y_{\mathrm{B}} = Bu_{\mathrm{B}}$$

Then

$$\left\{ \begin{array}{rcl} ANTPT^{\mathrm{T}}u_{\mathrm{A}} & = & \lambda^2 Au_{\mathrm{A}} \\ BPT^{\mathrm{T}}NTu_{\mathrm{B}} & = & \lambda^2 Bu_{\mathrm{B}} \end{array} \right.$$

We can "simplify" by A and B by left multiplication by $(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ and $(B^{\mathrm{T}}B)^{-1}B^{\mathrm{T}}$. We have

$$\left\{ \begin{array}{rcl} NTPT^{\mathrm{T}}u_{\mathrm{A}} & = & \lambda^2 u_{\mathrm{A}} \\ PT^{\mathrm{T}}NTu_{\mathrm{B}} & = & \lambda^2 u_{\mathrm{B}} \end{array} \right. \tag{8.3.3}$$

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,

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

Then, equation (8.3.3) reads

$$\left\{ \begin{array}{rcl} MRR^{\mathrm{T}}M^{-1}u_{\mathrm{A}} & = & \lambda^2 u_{\mathrm{A}} \\ QR^{\mathrm{T}}RQ^{-1}u_{\mathrm{B}} & = & \lambda^2 u_{\mathrm{B}} \end{array} \right. \tag{8.3.4}$$

Let us denote

$$w_{\mathrm{A}} = M^{-1}u_{\mathrm{A}}, \qquad w_{\mathrm{B}} = Q^{-1}u_{\mathrm{B}}$$

Then, by left multiplication by M^{-1} of the first equation and by Q^{-1} of the second, we have

$$\left\{ \begin{array}{rcl} RR^{\mathrm{T}} w_{\mathrm{A}} & = & \lambda^2 w_{\mathrm{A}} \\ R^{\mathrm{T}} R w_{\mathrm{B}} & = & \lambda^2 w_{\mathrm{B}} \end{array} \right. \tag{8.3.5}$$

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_{A} is a principal axis of the `PCA` of R^{T}, and w_{B} 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 the w_{B} as a principal axis of R. Then, w_{A} as a principal axis of R^{T} is a principal component of R and related to w_{B} by

$$w_{\mathrm{A}} = R w_{\mathrm{B}}, \qquad \text{with} \quad \begin{cases} R & \in \mathbb{R}^{p \times q} \\ w_{\mathrm{A}} & \in \mathbb{R}^p \\ w_{\mathrm{B}} & \in \mathbb{R}^q \end{cases} \tag{8.3.6}$$

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 $T = A^{\mathrm{T}} B$ with an inner product defined by N on rows and by P on columns where $N = (A^{\mathrm{T}} A)^{-1}$ and $P = (B^{\mathrm{T}} B)^{-1}$. We then have

$$u_{\mathrm{A}} = M w_{\mathrm{A}}, \qquad u_{\mathrm{B}} = Q w_{\mathrm{B}}$$

and

$$y_{\mathrm{A}} = A u_{\mathrm{A}}, \qquad y_{\mathrm{B}} = B u_{\mathrm{B}}$$

• **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: $\mathrm{CCA}(A, B)$

input $A \in \mathbb{R}^{n \times p}$, $B \in \mathbb{R}^{n \times q}$ with $p \geq q$
compute $T = A^{\mathrm{T}} B$
compute $N = (A^{\mathrm{T}} A)^{-1}$
compute $P = (B^{\mathrm{T}} B)^{-1}$
compute $M = N^{1/2}$
compute $Q = P^{1/2}$
compute $R = MTQ$
compute $W_{\mathrm{A}}, W_{\mathrm{B}}, \Psi = \mathrm{PCA\_CORE}(R)$
compute $\Lambda = \sqrt{\Psi}$
compute $U_{\mathrm{A}} = M W_{\mathrm{A}}$
compute $U_{\mathrm{B}} = Q W_{\mathrm{B}}$
compute $Y_{\mathrm{A}} = A U_{\mathrm{A}}$
compute $Y_{\mathrm{B}} = B U_{\mathrm{B}}$
return $Y_a, Y_{\mathrm{B}}, U_{\mathrm{A}}, U_{\mathrm{B}}, \Lambda$

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 $y_A = Au_A$, $y_B = Bu_B$ 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 from $\|y_A \wedge y_B \wedge y_C\|$ (vectors are independent when their wedge product is maximal), 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, called as well completely disjunctive table. 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

$$w = M^T M w, \qquad v = D_B^{1/2} w$$

with

$$M = D_A^{1/2} T D_B^{1/2}, \qquad T = A^T B, \qquad D_A = (A^T A)^{-1}, \qquad D_B = (B^T B)^{-1}$$

Key observations: Let us note three things:

→ $(A^T A)^{-1} \in \mathbb{R}^{p \times p}$ (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 $D_A^{1/2}$ in $\mathbb{R}^p$ and by $D_B^{1/2}$ 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 $p+q \leq n$ be two data sets, each a set of quantitative variables (the columns of A and B) on a same set of items (the rows of A and B). Let us have in mind the Canonical Analysis of (A, B), which will be developed here along another calculation.

Therefore, 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

$$\left|\begin{array}{lcll} D_{\mathrm{A}} & = & (A^{\mathrm{T}}A)^{-1} & \in \mathbb{R}^{p\times p} \\ D_{\mathrm{B}} & = & (B^{\mathrm{T}}B)^{-1} & \in \mathbb{R}^{q\times q} \end{array}\right.$$

and

$$D = \begin{pmatrix} D_{\mathrm{A}} & 0 \\ 0 & D_{\mathrm{B}} \end{pmatrix} \in \mathbb{R}^{(p+q)\times(p+q)}$$

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

$$T = A^{\mathrm{T}}B, \qquad M = D_{\mathrm{A}}^{1/2} T D_{\mathrm{B}}^{1/2}$$

Then, performing the `PCA` of X with row distances given by D is performing the `PCA` of M. If A and B are indicator arrays, this is performing the `CoA` of T, which is the contingency table of A and B.

Proof. The solution is performing a `PCA` of

$$R = XD^{1/2}, \qquad R \in \mathbb{R}^{n\times(p+q)}$$

i.e. performing an EVD of $R^{\mathrm{T}}R$. We will denote the matrix dimensions under each block. We have

$$\underset{n\times(p+q)}{R} = \left[\underset{n\times p}{AD_{\mathrm{A}}^{1/2}}\ ,\ \underset{n\times q}{BD_{\mathrm{B}}^{1/2}}\right], \qquad \underset{(p+q)\times n}{R^{\mathrm{T}}} = \left[\begin{array}{c} \underset{p\times n}{D_{\mathrm{A}}^{1/2}A^{\mathrm{T}}} \\ \\ \underset{q\times n}{D_{\mathrm{B}}^{1/2}B^{\mathrm{T}}} \end{array}\right]$$

So

$$R^{\mathrm{T}}R = \left[\begin{array}{cc} \underset{p\times p}{D_{\mathrm{A}}^{1/2}A^{\mathrm{T}}AD_{\mathrm{A}}^{1/2}} & \underset{p\times q}{D_{\mathrm{A}}^{1/2}A^{\mathrm{T}}BD_{\mathrm{B}}^{1/2}} \\ \\ \underset{q\times p}{D_{\mathrm{B}}^{1/2}B^{\mathrm{T}}AD_{\mathrm{A}}^{1/2}} & \underset{q\times q}{D_{\mathrm{B}}^{1/2}B^{\mathrm{T}}BD_{\mathrm{B}}^{1/2}} \end{array}\right]$$

Let us observe that

$$A^{\mathrm{T}}A = D_{\mathrm{A}}^{-1}, \qquad B^{\mathrm{T}}B = D_{\mathrm{B}}^{-1}$$

Hence

$$R^{\mathrm{T}}R = \left[\begin{array}{cc} \mathbb{I}_p & \underset{p\times q}{D_{\mathrm{A}}^{1/2}A^{\mathrm{T}}BD_{\mathrm{B}}^{1/2}} \\ \\ \underset{q\times p}{D_{\mathrm{B}}^{1/2}B^{\mathrm{T}}AD_{\mathrm{A}}^{1/2}} & \mathbb{I}_q \end{array}\right]$$

Let

$$x = \begin{bmatrix} u \\ v \end{bmatrix}$$

be such that

$$R^{\mathrm{T}} R x = \lambda x$$

This yields

$$\begin{cases} D_{\mathrm{A}}^{1/2} A^{\mathrm{T}} B D_{\mathrm{B}}^{1/2} v &= (\lambda - 1) u \\ D_{\mathrm{B}}^{1/2} B^{\mathrm{T}} A D_{\mathrm{A}}^{1/2} u &= (\lambda - 1) v \end{cases}$$

Let us denote

$$T = A^{\mathrm{T}} B, \qquad M = D_{\mathrm{A}}^{1/2} T D_{\mathrm{B}}^{1/2}$$

Then

$$\begin{cases} Mv &= (\lambda - 1) u \\ M^{\mathrm{T}} u &= (\lambda - 1) v \end{cases}$$

or

$$\begin{cases} M^{\mathrm{T}} M v &= (\lambda - 1)^2 v \\ M M^{\mathrm{T}} u &= (\lambda - 1)^2 u \end{cases}$$

where we recognize the `PCA` of M, hence the `PCA` of $T = A^{\mathrm{T}} B$ with metrics defined by $(A^{\mathrm{T}} A)^{-1}$ on columns and $(B^{\mathrm{T}} B)^{-1}$ on rows, hence the `CoA` of T as a contingency table. □

9.3 Multiple Canonical Analysis

Knowing that, the extension if Canonical Analysis to more than two quantitative variables is straightforward.

• Let us have m qualitative variables on n items, each with indicator matrix A_ℓ ($\ell \in [\![1, m]\!]$) with

$$A_\ell \in \mathbb{R}^{n \times p_\ell}, \qquad \sum_\ell p_\ell \leq n$$

Let us define

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

Let us build

$$X = [A_1 | \dots | A_m]$$

and

$$D = \begin{bmatrix} D_1 & 0 & \dots & \dots & 0 \\ 0 & D_2 & 0 & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & D_m \end{bmatrix}$$

Then, the `MCoA` of $(A_1, \dots, A_m)$ is the `PCA` of X with distances on rows given by D. It is the `PCA` of

$$R = [A_1 D_1 | \dots | A_m D_m] \tag{9.3.1}$$

We have

$$R^{\mathrm{T}} = \begin{bmatrix} D_1 A_1^{\mathrm{T}} \\ \vdots \\ D_m A_m^{\mathrm{T}} \end{bmatrix}$$

and $M = R^{\mathrm{T}} R$ can be written blockwise as

$$\begin{cases} M_{\ell\ell} = \mathbb{I}_{p_\ell} & \text{if } \ell = \ell' \\ M_{\ell\ell'} = D_\ell A_\ell^{\mathrm{T}} A_{\ell'} D_{\ell'} & \text{if } \ell \neq \ell' \end{cases}$$

9.4 Summary of relationships between some methods

Let us recall that, in this document, we denote:

`CoA`	:	Correspondence Analysis
`Can`	:	Canonical Analysis
`PCAmet`	:	Principal Component Analysis with metrics

Let us recall that:

→ `CoA` is the PCA of a contingency table T with metrics on rows and columns as the inverse of the row and column marginals,

→ `Can` is the analyss of two quanttative arrays (A, B) in order to find the most common components,

→ `PCAmet` the PCA of a quantitative array A with metrics defined by N on columns and P on rows.

◇ Let (A, B) be two indicator arrays of one categorical variable each on the same items. Then, The Canonical Analysis of (A, B) is equivalent to the Correspondence Analysis of $X = A^{\mathrm{T}} B$

$$\texttt{Can}(A, B) = \texttt{CoA}(A^{\mathrm{T}} B) \tag{9.4.1}$$

◇ Let (A, B) be two quantitative arrays. Let us consider their concatenation $X = [A|B]$, and the `PCAmet` of X with metrics defined by D on rows, with

$$D = \begin{pmatrix} D_{\mathrm{A}} & 0 \\ 0 & D_{\mathrm{B}} \end{pmatrix} \qquad \text{with} \quad \begin{cases} D_{\mathrm{A}} &= (A^{\mathrm{T}} A)^{-1} & \in \mathbb{R}^{p \times p} \\ D_{\mathrm{B}} &= (B^{\mathrm{T}} B)^{-1} & \in \mathbb{R}^{q \times q} \end{cases}$$

Then, the `PCA` of X with metrics on rows defined by D is equivalent to performing the Canonical Analysis of (A, B)

$$\texttt{Can}(A, B) = \texttt{PCAmet}(X, D) \tag{9.4.2}$$

◇ Let us now assume that A and B are each the indicator array of a categorical variable. Then, $\texttt{Can}(A, B)$ is equivalent to the Correspondence Analysis of $T = A^{\mathrm{T}} B$. Then, $\texttt{PCAmet}(X, D)$,

equivalent to `Can`(A, B), is equivalent to the Correspondence Analysis of $T = A^{\mathrm{T}}B$ as well by transitivity.

$\diamond$ The Canonical Analysis of two tables (A, B) as the `PCAmet` of $X = [A|B]$ with metric defined by D on the rows of X has been extended to the analysis of m arrays A_ℓ as the `PCAmet` of table

$$X = [A_1|\dots|A_m]$$

with metric defined by matrix D blockwise diagonal defined as

$$D = \mathrm{Diag}(D_\ell) \qquad \text{with} \quad D_\ell = (A_\ell^{\mathrm{T}} A_\ell)^{-1}$$

This leads to Multiple Correspondence Analysis (`MCA`) as follows.

9.5 Multiple Correspondence Analysis

Let us have m categorical variables observed each on n items. Let us denote by A_ℓ with $\ell \in [\![1, m]\!]$ the indicator array of variable ℓ, i.e. is a $n \times p_\ell$ binary array with

$$\alpha_{ij_\ell} = \begin{cases} 1 & \text{if modality } j_\ell \text{ has been observed for item } i \\ 0 & \text{otherwise} \end{cases}$$

Let us define

$$X = [A_1|\dots|A_m]$$

and the metric on rows of X defined by the matrix D blockwise diagonal defined as

$$D = \mathrm{Diag}(D_\ell) \qquad \text{with} \quad D_\ell = (A_\ell^{\mathrm{T}} A_\ell)^{-1}$$

Then, the Multiple Correspondence Analysis of $(A_1, \dots, A_m)$ is equivalent to the `PCAmet` of X with metric defined by D on rows.

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]. Since then, the diversity of related methods have flourished, and a more recent panorama is given in [GB06]. According to the introduction of [Gre84], L. Guttman should be credited for all the basic ideas at the root of Multiple Correspondence Analysis, in Guttman, L. 1941. *The quantification of a class of attributes: A theory and method of scale construction*, Chapter in *The Prediction of Personal Adjustment*, P. Horst, eds. New York : Social Science Research Council. The article [TY85] is a thorough survey of different methods associated with Multiple Correspondence Analysis, with both a historical background and, much prior to the present notes, an organization of methods to show that they all lead to the same equation to analyze the data. The unifying formalism selected in [TY85] is the duality diagram.

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

$$\mathcal{X} = (x_i)_{1 \leq i \leq n}, \qquad x_i \in \mathbb{R}^r$$

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 ℓ^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 Σ
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, Σ_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 $X = (x_i)_i$ 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

$$\left| \begin{array}{rcl} g_{ij} & = & \langle x_i, x_j \rangle \\ d_{ij}^2 & = & \|x_i - x_j\|^2 \end{array} \right.$$

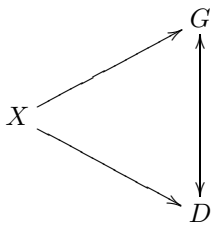
Then

$$\left\{ \begin{array}{rcl} d_{ij}^2 & = & g_{ii} + g_{jj} - 2g_{ij} \\ g_{ij} & = & -\frac{1}{2}\left(d_{ij}^2 - d_{i\bullet}^2 - d_{j\bullet}^2 + d_{\bullet\bullet}^2\right) \end{array} \right. \tag{10.1.1}$$

with

$$\left\{ \begin{array}{rcl} d_{i\bullet}^2 & = & \frac{1}{n}\sum_j d_{ij}^2 \\ \\ d_{\bullet\bullet}^2 & = & \frac{1}{n^2}\sum_{i,j} d_{ij}^2 = \frac{1}{n}\sum_i d_{i\bullet}^2 \end{array} \right. \tag{10.1.2}$$

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



• A key question is to know whether one-way arrows $X \longrightarrow G$ and $X \longrightarrow D$ 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

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

then

$$G_{ij} = \langle x_i, x_j \rangle \tag{10.2.1}$$

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

$$G = XX^{\mathrm{T}} \tag{10.2.2}$$

Next step is about computing X knowing XX^{T}.

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

$$Gu_\alpha = \sigma_\alpha u_\alpha \tag{10.2.3}$$

with

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$$

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

$$GU = U\Sigma \tag{10.2.4}$$

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

$$G = U\Sigma U^{\mathrm{T}} \tag{10.2.5}$$

We recognize here the SVD of G (hence the choice of Σ 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

$$G = (U\Lambda)(U\Lambda)^{\mathrm{T}} \tag{10.2.6}$$

and we can select

$$X = U\Lambda, \qquad \Lambda = \Sigma^{1/2} \tag{10.2.7}$$

It is not the only solution, because any matrix $X' = X\Omega$ where Ω 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 $G = U\Sigma U^{\mathrm{T}}$
4: **compute** $\Lambda = \Sigma^{1/2}$
5: **compute** $X = U\Lambda$
6: **return** X, Λ

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 $\mathcal{X}$ is done by the `PCA` of X.

- Let us recall that

$$X = U\Sigma^{1/2} \qquad \text{with} \quad U^{\mathrm{T}}U = \mathbb{I}_n \tag{10.3.1}$$

We recognize here the SVD of X as $X = U\Lambda V^{\mathrm{T}}$, with $\Lambda = \Sigma^{1/2}$ and $V = \mathbb{I}_n$. 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 σ_α) 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: $X, \Sigma = \mathrm{D}, r$

1: **input:** a distance matrix $D \in \mathbb{R}^{n \times n}$; a dimension $r < n$
2: **compute** the Gram matrix of D: $G = \textsc{gram}(D)$
3: **compute** the eigenpairs $(u_\alpha, \lambda_\alpha)$ such that $Gu_\alpha = \sigma_\alpha u_\alpha$
4: or **compute** the SVD of G as $G = U\Sigma U^{\mathrm{T}}$
5: keep in U the columns associated to non-negative eigenvalues of G; clip them off in Σ
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, Λ

- **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 Σ). If the SVD is selected, we have $G = U\Sigma U^{\mathrm{T}}$. We have the choice of the sign in columns of U. If the choice is the same in U and $V^{\mathrm{T}} = U^{\mathrm{T}}$, then the sign of σ_α 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`	$(Y, V, \Lambda) = \text{PCA\_CORE}(A)$	$C = A^{\text{T}} A$ $(\Lambda, V) = \text{EIG}(C)$ $Y = AV$ or $(U, \Sigma, V) = \text{SVD}(A)$ $\Lambda = \Sigma^2$ $Y = U\Sigma$
`PCAiv`	$(Y, V, \Lambda) = \text{PCA-IV}(A, U, V)$	$\mathcal{P}_E = U(U^{\text{T}}U)^{-1}U^{\text{T}}$ $\mathcal{P}_F = V(V^{\text{T}}V)^{-1}V^{\text{T}}$ $A_{\text{U,V}} = \mathcal{P}_{\text{E}} A \mathcal{P}_{\text{F}}$ $(Y, \Lambda, V) = \text{PCA\_CORE}(A_{\text{U,V}})$
`PCAmet`	$(Y, V, \Lambda) = \text{PCA-MET}(A, M, Q)$	$A_{\text{M,Q}} = MAQ$ $(Z, \Lambda, X) = \text{PCA\_CORE}(A_{\text{M,Q}})$ $Y = M^{-1}Z$ $V = Q^{-1}X$
`CoA`	$(Y_r, Y_c, \Lambda) =$ `CoA` (T)	$A = T/T_{++}, \quad T_{++} = \sum_{i,j} T_{ij}$ $r_i = \sum_j \alpha_{ij}, \quad c_j = \sum_i \alpha_{ij}$ $M = \text{diag}\,(1/\sqrt{r_i}), \quad Q = \text{diag}\,(1/\sqrt{c_j})$ $A_{\text{M,Q}} = M(A - r \otimes c)Q = \frac{\alpha_{ij} - r_i c_j}{\sqrt{r_i c_j}}$ $Z, \Lambda, X = \text{PCA\_CORE}(A_{\text{M,Q}})$ $Y_r = M^{-1}Z, \quad Y_c = Q^{-1}X$
`CCA`	$(Y_{\text{A}}, Y_{\text{B}}, U_{\text{A}}, U_{\text{B}}, \Lambda) = \text{CCA}(A, B)$	$T = A^{\text{T}}B$ $M = (A^{\text{T}}A)^{-1/2}, \quad Q = (B^{\text{T}}B)^{-1/2}$ $R = MTQ$ $W, \Psi, W = \text{PCA\_CORE}(R)$

12 References in textbooks

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

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]	secion 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

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

Abbreviations

CCA	Canonical Correlation Analysis
CoA	Correspondance Analysis
IV	Instrumental Variables
MDS	Mulidimensional Sca;ling
PCA	Princial Component Analysis
SDP	Symmetric Definite Positive
SGF	Symmetric Gauge Function
SVD	Singular Value Decomposition
UIN	Unitarily Invariant Norm

References

[ACD+22] E. Agullo, O. Coulaud, A. Denis, M. Faverge, A. Franc, J.-M. Frigerio, N. Furmento, A. Guilbaud, E. Jeannot, R. Peressoni, 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.

[GB06] M. Greenacre and J. Blasius, editors. *Multiple Correspondence Analysis and Related Methods*. Chapman & al., 2006.

[Git85] R. Gittins. *Canonical Analysis: A Review with Applications in Ecology*, volume 12 of *Biomathematics*. Springer, 1985.

[Gre84] M. Greenacre. *Theory and Applications of Correspondence Analysis*. Acadelic Press, 1984.

[Hil74] M. O. Hill. Correspondence analysis: A neglected multivariate method. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 23(3):340–354, 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 sstatistical 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 dat. *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.

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399