Linear Dimensionality Reduction

Alain Franc

Linear Dimensionality Reduction

Alain Franc

RESEARCH REPORT
N° 9488
May 2023 (2nd revision)
Project-Team Pleiade

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

Linear Dimensionality Reduction

Alain Franc*†

Project-Team Pleiade

Research Report n° 9488 — May 2023 (2nd revision) — 99 pages

Abstract: These notes are an overview of some classical linear methods in Multivariate Data Analysis. This is a 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 like distributed memory and task-based programming or to improve the efficiency of the algorithms like Randomised SVD can focus on this shared core method, and benefit to all methods.

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 peut se ramener 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

Motivation

The writing of this document is 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 developed by a French school, in parallel with more statistical approaches in Anglo-Saxon countries.

◇ The main methods are PCA (Principal Component Analysis) associating items and features, CoA (Correspondence Analysis) of contingency tables, Canonical Analysis for analysis of two tables, Multiple Correspondence Analysis for analysis of two tables or more, MDS (Multidimensional Scaling) for building a point cloud from a table of distances, etc... 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), mainly because statistical learning methods are sometimes inoperative in spaces of very large dimension. A pretreatment as a projection on a space of lower dimension can make them operational, provided the projection respects the structure of the dataset.

◇ 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 a Singular Value Decomposition (SVD), 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.

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 [BM01, HMT11]),
- a distributed memory implementation of the basic matrix operations,
- the use of a task-based programming paradigm for the assembly of these steps.

Random projection if often presented itself as a dimensionality reduction method, by projecting a point cloud on a space of lower dimension while respecting the pairwise distances (see e.g. [BM01]). In these notes, it is used as a tool for lowering the complexity of the calculation of the SVD of a large matrix, as in [HMT11]. The integration of the rSVD in the MDS algorithm has been made in [Par18, BCF+18]. Numerical implementation for very large matrices (namely $10^6 \times 10^6$) with distributed-memory and task-based programming has been made in [ACD+22]. A motivation for these notes is to show how such an approach developed for MDS can be operational for any linear dimensionality reduction method that relies on a SVD.

Acknowledgements: These notes have been written as a methodological companion to two ADT (Action de Développement Technologique) built on collaborations between INRIA and Inrae at Bordeaux: Gordon (2019-2020) and Diodon (2021-2022), the aim of which was to provide librairies for running linear dimension reduction on massive data sets [ACD+22] with rSVD, possibly (in C++) distributed memory and task based programming. Most of the pseudocodes presented in these notes, including the integration of Randomized SVD, have been implemented in

- a C++ library written by INRIA, called `cppdiodon`, publicly available at `https://diodon.gitlabpages.inria.fr/cppdiodon/index.html`,
- a python library written by Inrae, called `pydiodon`, publicly available at `https://gitlab.inria.fr/diodon/pydiodon`.

I am particularly grateful to Emmanuel Agullo, Pierre Blanchard, Olivier Coulaud, Jean-Marc Frigerio, Romain Péressoni, Florent Pruvost for many discussions throughout these projects and before, especially on linear algebra, and their encouragements to make explicit this methodological companion which enabled the continuation of the Gordon ADT through the Diodon ADT. I am particularly indebted to Francis Cailliez, Daniel Chessel, Jean-Baptiste Denis, Yves Escoufier, Jean-Dominique Lebreton and Robert Sabatier, who taught me multivariate data analysis many decades ago.

Contents

1 Introduction

Let us start with an example: supervised learning with Support Vector Machine (SVM). 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 (J.-L.) telling that there exists a space H of smaller dimension $d \ll p$ (of $\mathcal{O}(\mathrm{Log}\, p)$) 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, linear dimension reduction methods in multivariate data analysis. This establishes a tight link between machine learning and multivariate data analysis.

Multivariate Data Analysis (MDA) 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. MDA 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:

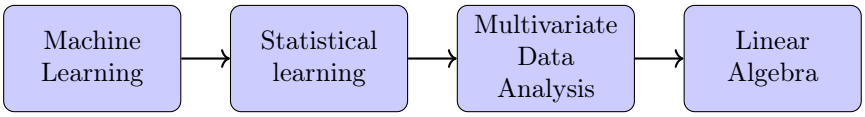
Machine learning	**Statistics**
weights	parameters
learning	fitting
supervised learning	regression/classification
unsupervised learning	density estimation, clustering

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. [Mur12, SSBD14]). A grail is to mimic the way a brain learns from experience. This is beyond the scope

[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

of those notes.

MDA 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 MDA in Machine Learning. Murphy recognized two approaches in Machine Learning [Mur12, sec. 1.1.2], which are consistent with 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. MDA plays a key role in such a framework to produce a low dimension representation of an array connecting items and features, as close as possible to the original array. This is called Dimensionality Reduction (see e.g. [LV07]), and one of the key (linear) technique therefore is Principal Component Analysis (PCA, 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

MDA is at the crossroads 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 crossroads between algebra, geometry and statistics.

As arrays of data are matrices, `MDA` heavily relies 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 is to describe how items and features are related, which is usually addressed by a 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 be modeled as the realization of a set of random variables $(X_1, \ldots, X_p)$. Questions of interest are the study of the dependence structure of the X_j, given by the variance-covariance matrix of observed data, or exhibiting low dimension latent variables z, such that observations x can be modeled as $\mathbb{P}(x \mid z)$.

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, PCA can be built as a consequence of the Singular Value Decomposition of A: $A = U\Sigma V^{\mathrm{T}}$. These methods are constantly co-evolving with numerical linear algebra, and have benefited from[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 progress 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-memory parallelization), and progress in technology of computing infrastructures (like Graphic Processing Units).

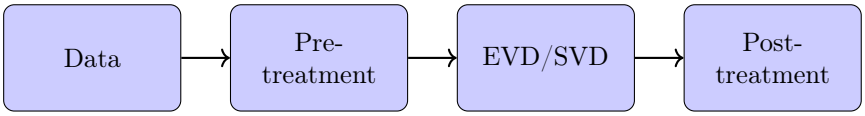
The high diversity of existing method can be organized into a small set of iconic methods, knowing that such a classification is far from being either unique or universally adopted. However, most textbooks presenting these methods progressively reach an agreement on the poles around which to organize 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. 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] and a current research with "Principal Component Analysis" on Google Scholar provided 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 MDA or dimensionality reduction, like [And58, MKB79, CC80, LV07, Ize08, Wan12]. Each has a special flavour: [And58] is the seminal book on MDA (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 framework (for presentation of the duality diagram, see as well [DD07, DlCH11]). See as well [LMT77, LMF82, LMP00, 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)
$\|.\|_{\text{sp}}$	Spectral norm
$[\![a, b]\!]$	the set of integers i with $a \leq i \leq b$
a_{ij}	coefficient in row i and column 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_r	a matrix of rank r
$A \geq 0$	a non-negative matrix
$\mathcal{A}$	a point cloud of n points in $\mathbb{R}^p$
E_r	a subspace of $\mathbb{R}^p$ of dimension r
$\mathbb{I}_n$	the identity matrix in $\mathbb{R}^{n \times n}$
$\mathcal{L}(E, F)$	the space of linear functions from E to F
r	prescribed rank for best low rank approximation
$\mathbb{R}^{n \times p}$	the space of matrices with n rows and p columns

Notations more specific to a section are given at the beginning of the corresponding section. We have tried to be as consistent as possible between sections, but the diversity of situations and matrices involved makes such a challenge sometimes ... challenging.

3 Principal Component Analysis (PCA)

Notations

Symbol	space	meaning
α	$\mathbb{N}$	index of eigenvalues of C
a_{i*}	$\mathbb{R}^p$	row i of A, point i in $\mathcal{A}$
$a_{i,:}$	$\mathbb{R}^p$	row i of A, point i in $\mathcal{A}$
$\widetilde{a_{i*}}$	$\mathbb{R}^p$	projection of a_{i*} on E_r
A	$\mathbb{R}^{n\times p}$	matrix to analyse
$\mathcal{A}$		point cloud associated with A
$\mathcal{A}_r$		projection of $\mathcal{A}$ on E
A_r	$\mathbb{R}^{n\times p}$	best rank r matrix for approximating A
C	$\mathbb{R}^{p\times p}$	variance-covariance matrix
E_r	$\subset \mathbb{R}^p$	best $r-$dimensional subspace
λ	$\mathbb{R}$	eigenvalue of C
Λ	$\mathbb{R}^{p\times p}$	diagonal matrix of eigenvalues of C
r	$\mathbb{N}$	rank
σ	$\mathbb{R}$	singular value of A
Σ	$\mathbb{R}^{p\times p}$	diagonal matrix of singular values of A
U	$\mathbb{R}^{n\times p}$	left singular vectors of U, colulmnwise
v	$\mathbb{R}^p$	eigenvector of C (principal axis)
V	$\mathbb{R}^{p\times p}$	matrix of eigenvalues of C, columnwise right singular vectors of A, columnwise
Y	$\mathbb{R}^{n\times p}$	matrix of principal components

Matrices involved (PCA in a nutshell)

Matrix	dimensions	what it is
A	$n \times p$	matrix to be analyzed
C	$p \times p$	variance - covariance matrix of A
Λ	$p \times p$	diagonal matrix of eigenvalues of C
Σ	$p \times p$	diagonal matrix of singular values of A
U	$n \times p$	left singular vectors of A
V	$p \times p$	principal axis (columnwise) right singular vectors of A
Y	$n \times p$	principal components

with

$$\begin{array}{rcl} C & = & A^{\mathrm{T}}A \\ CV & = & V\Lambda \\ A & = & U\Sigma V^{\mathrm{T}} \\ \Lambda & = & \Sigma^2 \\ Y & = & AV \\ Y & = & U\Sigma \\ U^{\mathrm{T}}U & = & \mathbb{I}_p \\ V^{\mathrm{T}}V & = & \mathbb{I}_p \end{array}$$

PCA 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 $\mathcal{A}_{\mathrm{E}}$ 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 Frobenius (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 PCA is well known and well documented. According to [Bas94, chap. 3], its origin can be traced in the work of Bravais in 1846 (where the notion of principal axis emerged, in "in the form of rotating an ellipse to 'axes principaux' in order to achieve independence in a multivariate normal distribution"). Classically, its origin is attributed to Pearson in [Pea01] under the guise of both a statistical and a geometrical derivation, as an extension of the linear regression. The aim is statistical, but the idea behind is clearly geometric. Pearson's aim was to escape from the non symmetry of dependent and independent variables in linear regression (the regression line changes if the status dependent - independent are reversed), by giving equal status for variations on both types of variables. He was led to find the best fit of, say, a system of points in the plane by a line. One guise of his approach is statistical, in the sense that it uses the notions of mean, variance, standard deviation of a sample, but the notion of statistical model as it is understood nowadays was not available at this time. So, Pearson's approach can be qualified as geometrical. 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 SVD

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]. The triple nature of PCA (algebraic, geometrical, statistical) can be found in [VMS16], where chapter 2 is a thorough presentation of PCA. PCA is developed in every textbook in MDA, 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 under various names according to the field it is applied to, as mentioned in `https://en.wikipedia.org/wiki/Principal_compo` Principal Component Analysis, Karhunen–Loève transform (KLT) in signal processing, proper orthogonal decomposition (POD) in mechanical engineering, empirical orthogonal functions (EOF) in meteorological science, among others.

If the theory can be derived for any pair (p, r) of dimensions with $1 \leq r < p$, it is most interesting when p is large and $r \ll p$.

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	$\operatorname{rank} A_r = r$	
such that	$\|A - A_r\|$	minimal

- **Geometric approach:**

In this approach, a cloud denoted $\mathcal{A}$ of n points in $\mathbb{R}^p$, labeled $(a_i)_i$ with $1 \leq i \leq n$, is associated to a matrix $A \in \mathbb{R}^{n \times p}$, where point a_i is row i of A: $a_i = a_{i,:}$. The point cloud associated to A_r is denoted $\mathcal{A}_r$

Given a point cloud	$\mathcal{A} = (a_1, \ldots, a_n)$ $a_i \in \mathbb{R}^p$ $0 < r < p$
Find with	a subspace $E_r \subset \mathbb{R}^p$ $\dim E_r = r$
such that	$d(\mathcal{A}, \mathcal{A}_r)$ minimal
where and	$d(\mathcal{A}, \mathcal{A}_r) = \sum_i \|a_i - \widetilde{a_i}\|^2$ $\widetilde{a_i}$ 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 $(v_1, \ldots, v_r)$ of E_r, and let us complete it to have an orthonormal basis of $\mathbb{R}^p$ as $(v_1, \ldots, v_r, v_{r+1}, \ldots, v_p)$. If the projection is exact, i.e. if $\forall\, i$, $a_i = \widetilde{a_i}$, 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	$\mathcal{A} = (a_1, \ldots, a_n)$ $a_i \in \mathbb{R}^p$ $0 < r < p$	
Find with	a subspace $E_r \subset \mathbb{R}^p$ $\dim E_r = r$	(3.2.1)
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 a first algorithm, for a function called PCA_EVD():

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

About the choice between SVD and EVD: Adopting the SVD viewpoint has some numerical advantages:

→ there exists efficient and stable algorithms for computing an 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 9),

→ 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.

However, matrix $C = A^{\mathrm{T}}A$ is the variance covariance matrix of the distribution of the variables in the statistical approach, and must not be ignored.

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 $\mathcal{O}(n^2p)$. 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 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, see appendix C), 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 $x, y \in X$

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

A demonstration of the existence relies on 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 built 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** $B = Q^{\mathrm{T}}A$
6: **run** the SVD of B: $B = U_{\mathrm{B}}\Sigma V^{\mathrm{T}}$, or $(U_{\mathrm{B}}, \Sigma, V) = \text{SVD}(B)$
7: **compute** $U = QU_{\mathrm{B}}$
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	$k \times p$	$B = Q^{\mathrm{T}}A$
U_{B}	$k \times k$	$B = U_{\mathrm{B}}\Sigma V^{\mathrm{T}}$
Σ	$k \times k$	idem
V	$k \times p$	idem
U	$n \times k$	$U = QU_{\mathrm{B}}$

- In section 10, we will see that MDS relies on the SVD or EVD of the Gram matrix of a point cloud (if $\mathcal{X} = (x_i)_i$ is a point cloud with $x_i \in \mathbb{R}^p$ and $1 \leq i \leq n$, the Gram matrix G of $\mathcal{X}$ is the symmetric definite positive matrix of elements $g_{ij} = \langle x_i, x_j \rangle$). Then, $G = XX^{\mathrm{T}}$. MDS is solving the reverse problem: finding X knowing G. If $G = U\Sigma U^{\mathrm{T}}$ is the SVD of G, then a solution to MDS is $X = U\Sigma^{1/2}$. The SVD can be done by Random Projection by selecting Ω, computing $Y = G\Omega$, Q such that $Y = QR$, defining $G' = QQ^{\mathrm{T}}G \approx G$ and doing the SVD of G'. If $B = Q^{\mathrm{T}}G$, the SVD of B is $B = U_{\mathrm{B}}\Sigma V^{\mathrm{T}}$ and $G' = QB = U\Sigma V^{\mathrm{T}}$ with $U = QU_{\mathrm{B}}$. However, one cannot write $G' = X'X'^{\mathrm{T}}$ because G' is not symmetric, and $V \neq U$. To solve this, one can do a double projection on G, rowwise and columnwise, and define $G'' = (QQ^{\mathrm{T}})G(QQ^{\mathrm{T}})$, with still $G'' \approx G$. One defines $C = Q^{\mathrm{T}}GQ$, which is symmetric, do its SVD by $C = U_{\mathrm{C}}\Sigma U_{\mathrm{C}}^{\mathrm{T}}$, hence $G'' = U\Sigma U^{\mathrm{T}}$ with $U = QU_{\mathrm{C}}$. One then can write $G \approx X''X''^{\mathrm{T}}$ with $X'' = U\Sigma^{1/2}$. This can be derived as follows:

Algorithm 4 SVD of a Gram matrix with Gaussian Random Projection SVD_GRP_GRAM(G, k))

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

- Here are the dimensions of the involved matrices:

Matrix	dimensions	computation
G	$n \times n$	Gram matrix
Ω	$n \times k$	Gaussian random matrix
Y	$n \times k$	$Y = G\Omega$
Q	$n \times k$	$Y = QR$
C	$k \times k$	$B = Q^{\mathrm{T}}GQ$
U_{C}	$k \times k$	$C = U_{\mathrm{C}}\Sigma U_{\mathrm{C}}^{\mathrm{T}}$
Σ	$k \times k$	idem
U	$n \times k$	$U = QU_{\mathrm{C}}$

One observes that the complexity of the calculation of the SVD of C is in $\mathcal{O}(k^3)$, whereas the calculation of $Y = G\Omega$ is in $\mathcal{O}(n^2k)$, and more "expensive".

Notes and references: The property behind Random Projection (RP) is counter-intuitive: "RP, while reducing dimensionality, approximatively preserves pairwise distances with high probability" ([Vem04, p. 2]). This is a consequence of Johnson-Lindenstrauss lemma ([JL84], see appendix C for details). [Vem04] presents a variety of domains of application of random projection. This is highly counter-intuitive, because it can undermine the very notion of PCA. Indeed, a point cloud and a rank being given, PCA is about finding the best subspace of dimension k as far as quality of projection is concerned. RP answer is that any randomly selected subspace of dimension k will make the job! Actually, this is not true for very small k, like first two or three axis useful for visualizing the shape of the point cloud. However, when k is significantly larger, this is true. It is possible to show (see appendix C) that the best choice is that k is in $\mathcal{O}(\frac{\mathrm{Log}\, n}{\epsilon^2})$ where ϵ is the relative accuracy in distances preservation. The presence of ϵ^2 at the denominator forces n to be very large for $\mathrm{Log}\, n < n\epsilon^2$. This is however true for any point cloud, including random ones. In applications, point clouds have a pattern or a structure, and RP works for much smaller values of n. See for example [BM01] for a comparison in accuracy and computing time with some other dimensionality reduction tools like PCA and an overview of various methods for selecting the matrix Ω (Gaussian matrix is not the only possible choice). The use of randomized algorithms to compute efficiently the SVD of a very large matrix is fully developed in [HMT11]. This section explains how SVD with Gaussian Random Projection works. To understand why it works, see appendix C.

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 5 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 pre-treatment. The name PCA designates classically 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, Λ, V) 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 an 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. $(y \otimes x)_{ij} = y_i v_j$). Let us denote by y_j the column j of Y, and v_j the column j of V. 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

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

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:

  $\rightarrow$ select a rank r, and deduce the quality ρ of the approximation,

  $\rightarrow$ 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 6 PCA-SC(A)

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

Even if there exists a rigorous definition in relation with the theory of the measure, let us assume that a random variable is a probability distribution on the outcome of a random experiment. Typical examples are dice rolling or coin flipping. In dice rolling, X is the observed value on the upwards face of the dice when it stops rolling. The possible outcomes are $\{1, 2, 3, 4, 5, 6\}$ and the random variable X is defined by

$$p_1 = \mathbb{P}(X = 1), \quad p_2 = \mathbb{P}(X = 2), \quad \ldots \tag{4.1.1}$$

If the dice rolls n times, the outcome is a tuple in $\{1, \ldots, 6\}^n$, referred to as n realisations if X. They are called i.i.d. for independent identically distributed.

- Let X be a random variable with values in $\mathbb{R}$. Let $x = (x_i)_i$ with $1 \leq i \leq n$ be a vector of n realizations of X (it is called a sample). Then, the sample mean is

$$\mathbb{E}(x) = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4.1.2}$$

It must not be confounded with the population mean $\mathbb{E}(X)$. For sake of clarity, we denote

$\mathbb{E}(X)$	the population mean
$\mathbb{E}(x)$ or $\overline{x}$	the sample mean

A sample is a subset of a possibly infinite population. Statistics are about inferring properties for the population knowing a sample. For example, if an infinite population follows a Gaussian law of mean μ and variance σ^2, the population mean is μ whereas the sample mean is $\overline{x} = (1/n)\sum_i x_i$. The sample mean is an unbiased estimator of the mean of the population.

The population variance is defined as

$$\operatorname{var} X = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) \tag{4.1.3}$$

The sample variance is defined as

$$\operatorname{var} x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{4.1.4}$$

It is an unbiased estimator of the population variance.

The standard deviation is the square root of the variance

$$\operatorname{std} x = \sqrt{\operatorname{var} x} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{4.1.5}$$

If $\mathbb{E}(x) = 0$, $\operatorname{var} x = \frac{1}{n-1}\sum_i x_i^2 = \|x\|^2/(n-1)$, which establishes a direct link between the geometric approach relying on Frobenius norm and statistical approach relying on variance. Let E_r be a $r-$dimensional subspace in $\mathbb{R}^p$. Each point $x_i \in \mathbb{R}^p$ has a projection $\widetilde{x_i}$ on E_r. Pythagore theorem

$$\|x_i\|^2 = \|x_i - \widetilde{x_i}\|^2 + \|\widetilde{x_i}\|^2$$

can be read as a decomposition of the variance of $X = (x_1, \ldots, x_n)$. The projection is optimal when the discrepancy $\|x_i - \widetilde{x_i}\|^2$ is minimal, or $\|\widetilde{x_i}\|^2$ is maximal, i.e. $\operatorname{var}\widetilde{x_i}$ maximal. This leads to statistical approach of PCA.

The covariance of two random variables X and Y is

$$\operatorname{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \tag{4.1.6}$$

and

$$\operatorname{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{4.1.7}$$

It is a symmetric, semi-definite, positive bilinear form. It is definite, hence an inner product, for those samples with $\overline{x} = 0$: the map $(x, y) \longrightarrow \operatorname{cov}(x, y)$ is an inner product, and $x \longrightarrow \operatorname{var} x$ is a norm, proportional to the Frobenius norm. Two random variables are independent if $\operatorname{cov}(X, Y) = 0$. In such a case

$$\operatorname{var}(X + Y) = \operatorname{var} X + \operatorname{var} Y \tag{4.1.8}$$

This is Pythagore theorem.

4.2 Statistical approach

The statistical approach to PCA is a complex topic with many guises. Some rely on empirical distributions, without inference, and some rely on inference, with or without latent variables. Some are presented here.

• Statistical approach has been set first with data sets being realisation of random variables. Let us have a $p-$variate random variable (rv) $X = (X_1, \ldots, X_p)$ with zero mean: $\mathbb{E}(X) = (0, \ldots, 0)$, observed on n independent items . A typical example is a $p-$variate Gaussian distribution with zero mean and Σ as variance-covariance matrix. The n observations for variable X_j are the column j of a matrix X. Each variable X_j is centered: $\mathbb{E}(X_j) = 0$. `PCA` at rank r is about finding r independent (non correlated) linear combinations $\widetilde{X_k} = Xu_k = \sum_{j=1}^{p} u_{kj} X_j$, $u_k \in \mathbb{R}^p$, $1 \leq k \leq r$, of the X_j which have maximum variance under the constrain $\|u_k\| = 1$. Here, means, variances, covariances are sample mean, sample variances and sample covariances. Statistical PCA relies on the observation that the sample variance of a realisation $x \in \mathbb{R}^p$ of a centered $p-$variate rv is proportional to the square of its Frobenius norm: if $\mathbb{E}(x) = 0$, $\operatorname{var} x \propto \|x\|^2 = \langle x, x \rangle$.

• 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 variable X to be Gaussian, although some more can be said if it is 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

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

Maximizing $\operatorname{var} Xu$ with the constraint $\|u\| = 1$ yields that u is the eigenvector of Σ associated to its largest eigenvalue, denoted λ. Then, as $\Sigma u = \lambda u$, $\operatorname{var} Xu = \langle u, \Sigma u \rangle = \lambda \langle u, u \rangle = \lambda$.

• 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 a rotation $U \in \mathbb{O}(\mathbb{R}^p)$ defining

$$V = XU$$

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 $V_1, \ldots, V_{k-1}$.

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].

4.3 Factor Analysis and Probabilistic PCA

Let us have a $p-$variate Gaussian variable, with n iid realisations, considered as observations. Correlations between observed variables are expressed by the variance-covariance matrix Σ. Factor analysis (FA) models a situation where there exists r independent unobserved variables, with $r < p$, the response to which explains the correlation between the observed variables. It is a

[4] Σ is the standard notation for the diagonal matrix of singular values of a given matrix as well. Confusion can easily be avoided from context.

latent variable statistical model. Unobserved variables are called *hidden variables* or *latent variables*. Let z denote the latent variable, and x the observed variable (the data). Let us assume that random variable Z follows a certain law, denoted

$$p(z_i \mid \theta)$$

(it will be specified later). Then, the model for x_i in FA can be written

$$p(x_i \mid z_i, \theta')$$

and solving FA is about inferring parameters in θ, θ' and recovering the latent variables z_i. In Bayesian perspective, $p(z \mid \theta)$ is called the *prior*.

- This makes FA rather complex, and this complexity has nourished decades of debates about Factor Analysis and PCA. Therefore, many authors (including [Jol02, chap. 7]) have insisted on the difference between PCA and FA. In fact, FA is a statistical approach of PCA where principal axis (and their realizations as components) are latent variables.

- Let us select as prior for the latent variables a multivariate Gaussian probability

$$p(z_i) = \mathcal{N}(z_i \mid \mu_0, \Sigma_0) \tag{4.3.1}$$

(here, $\theta = (\mu_0, \Sigma_0)$). Then, we select for $p(x \mid z)$ a Gaussian law as well, denoted

$$p(x_i \mid z_i, \theta) = \mathcal{N}(Wz_i + \mu, \Psi) \tag{4.3.2}$$

($\theta' = (W, \mu, \Psi)$) where the mean has been selected as a linear function of the hidden input ($x_i \in \mathbb{R}^p$, $z_i \in \mathbb{R}^r$ and $W \in \mathbb{R}^{p \times r}$), and Ψ to be diagonal. The special case where $\Psi = \sigma^2 \mathbb{I}_p$ is known as *probabilistic PCA*. If $\sigma \to 0$, probabilistic PCA converges to PCA which is non-probabilistic. We then have the following reductions

$$\text{FA: } \Psi \xrightarrow{\Psi=\sigma^2\mathbb{I}} \text{probabilistic PCA} \xrightarrow{\sigma=0} \text{PCA}$$

- Marginalizing over the latent variable leads to

$$\begin{aligned} p(x_i \mid \theta) &= \int_{z_i} p(x_i \mid z_i)\, p(z_i \mid \theta)\, dz_i \\ &= \mathcal{N}(x_i \mid W\mu_0 + \mu\,,\, \Psi + W\Sigma_0 W^{\mathrm{T}}) \end{aligned} \tag{4.3.3}$$

This shows that t is possible to chose $\mu_0 = 0$ and $\Sigma_0 = \mathbb{I}_r$ without loss of generality:

$$p(z_i \mid \theta) = \mathcal{N}(0, \mathbb{I}_r) \tag{4.3.4}$$

which leads to

$$p(x_i \mid z_i, \theta') = \mathcal{N}(Wz_i + \mu\,,\, WW^{\mathrm{T}} + \Psi) \tag{4.3.5}$$

and

$$p(x_i \mid \theta) = \mathcal{N}(x_i \mid \mu, WW^{\mathrm{T}} + \Psi). \tag{4.3.6}$$

The link with `PCA` can be read in the observation that the variance-covariance matrix of the data x_i is given by a low-rank matrix $WW^{\mathrm{T}} + \Psi$, as $W \in \mathbb{R}^{p \times r}$ and $\operatorname{rank} WW^{\mathrm{T}} = r$. 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.

- Next step is to estimate the parameters (W, μ, σ) or (W, μ, Ψ) of the model knowing the observations, not presented here. This is not so easy, and is presented in [Bis06, sect. 12.2.1]. It can be done with EM procedure (see Mur12, sec. 12.1.5.

Notes and references: Factor analysis has been developed all over the 20th century in parallel with PCA, sometimes with some controversies. One probable reason for those controversies is that sometimes it is said to inherit from the work of Pearson in 1901, and sometimes of Spearman in 1904, who was interested in finding one factor explaining a diversity of correlated features. His work has been extended to multivariate factors by Thurstone in 1935. The coexistence of PCA and FA with the same tools has probably contributed to some controversies. An history of FA can be found in the introduction of [Bas94]. A sound utilisation of notions in statistical modeling has clarified the situation. A classical presentation with this approach is [And58, sect. 4.7], or [Bas94]. It is currently developed as a statistical model with latent variables, where the correlated features are the observed variables, and the independent factors the latent variables. Classical setting is with Gaussian models. Probabilistic PCA (PPCA) has been proposed in 1999 in [TB99], who have presented as well the links and differences between PCA, PPCA and FA. See as well [Bis06, sect. 12.2] for a clear presentation of statistical approach of `PCA`, probabilistic `PCA`, and factor analysis. We have followed [Bis06] and [Mur12, chap. 12] here.

4.4 Distribution of eigenvalues of random matrices

A random matrix is the realization of a random variable over a space of matrices. Different spaces have been studied, and classically referred to for historical reasons as "ensembles". For example, the Ginibre ensemble is the set of $n \times n$ complex matrices whose entries are i.i.d. realizatons of the complex Gaussian normal law, i.e. $a_{k\ell} = x_{k\ell} + i\, y_{k\ell}$ with $x_{k\ell}, y_{k\ell} \sim \mathcal{N}(0,1)$. There are several types of results about the distribution of the eigenvalues of random matrices, like :

→ what is the distribution of the eigenvalues of a matrix in a given ensemble for size n ?

→ when $n \to \infty$, does the distribution of the eigenvalues converge to a given "universal" distribution ?

Wishart matrix: Let $A \in \mathbb{R}^{n\times p}$ be a random matrix, with rows being i.i.d. realisations of a $p-$variate Gaussian distribution X of mean $\mu = 0$ and variance-covariance matrix Σ. Let us denote $C = A^{\mathrm{T}}A \in \mathbb{R}^{p\times p}$. Then, elements in C follows a Wishart distribution, denoted $W(n,p,\Sigma)$, which has been analytically computed by Wishart in 1928. For $\Sigma = \mathbb{I}_p$, it is given by

$$p(C) = w(n,p)\,(\det C)^{(n-p-1)/2}\,\exp -\frac{1}{2}\mathrm{Tr}\,C, \tag{4.4.1}$$

where $w(n,p)$ is the normalizing constant

$$\frac{1}{w(n,p)} = \pi^{p(p-1)/4}\,2^{np/2}\prod_{j=1}^{p}\Gamma\left(\frac{n-j+1}{2}\right). \tag{4.4.2}$$

The formula for a general Σ is

$$p(C) = w(n,p,\Sigma)\,(\det C)^{(n-p-1)/2}\,\exp -\frac{1}{2}\mathrm{Tr}\,\Sigma^{-1}C, \tag{4.4.3}$$

with

$$\frac{1}{w(n,p,\Sigma)} = \pi^{p(p-1)/4}\,2^{np/2}\,(\det\Sigma)^{n/2}\prod_{j=1}^{p}\Gamma\left(\frac{n-j+1}{2}\right). \tag{4.4.4}$$

Marčenko-Pastur semi-circular law: Let $A \in \mathbb{R}^{n\times p}$ be a random matrix as above with its rows being i.i.d. realizations of a $p-$variate Gaussian law of mean 0 and variance-covariance Σ. Let us denote by

$$\lambda_1 > \dots \lambda_p > 0$$

the p eigenvalues, assumed to be separated. Let the size $n \times p$ of the matrix A grow with

$$\lim_{n\to\infty} \frac{p(n)}{n} = \alpha.$$

Let

$$C_p = \frac{1}{n} A_p^{\mathrm{T}} A_p$$

be an estimator of the variance covariance matrix. Let us assume that $0 < \alpha \leq 1$. Let $\Sigma = \mathbb{I}_p$, and define

$$\begin{cases} a &= (1-\sqrt{\alpha})^2 \\ b &= (1+\sqrt{\alpha})^2. \end{cases}$$

Let μ be the measure such that

$$\mu(\Omega) = \frac{1}{p} \#\{i \mid \lambda_i \in \Omega\} \tag{4.4.5}$$

for $\Omega \subset \mathbb{R}$. Then,

$$\lambda \in [a, b], \tag{4.4.6}$$

and, if $x \in [a, b]$

$$d\mu(x) = \frac{1}{2\pi\alpha} \frac{\sqrt{(b-x)(x-a)}}{x}. \tag{4.4.7}$$

So, if $a \leq z < z' \leq b$

$$\mathbb{P}(z \leq \lambda \leq z') = \frac{1}{2\pi\alpha} \int_z^{z'} \frac{\sqrt{(b-x)(x-a)}}{x}\, dx. \tag{4.4.8}$$

Notes and references: The Wishart matrix has been thoroughly studied, and much can be said about it. See e.g. [And58, chapter 7] devoted to it. The expressions of the joint law of the coefficients of a Wishart matrix in the case where $\Sigma = \mathbb{I}_p$ and for any Σ have been borrowed from [And58, section 7.2, formula (1)] and [Mec19, section 2.2]. For the asymptotic distribution of the eigenvalues of a Wishart matrix through Marčenko-Pastur theorem, we have followed [Mec20], which is a gem.

4.5 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

$$\forall \begin{cases} U & \in \mathbb{U}(\mathbb{C}^n) \\ V & \in \mathbb{U}(\mathbb{C}^p) \\ A & \in \mathbb{C}^{n\times p} \end{cases}, \qquad \|UAV^*\| = \|A\| \tag{4.5.1}$$

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

$$\forall \begin{cases} U & \in \mathbb{O}(\mathbb{R}^n) \\ V & \in \mathbb{O}(\mathbb{R}^p) \\ A & \in \mathbb{R}^{n\times p} \end{cases}, \qquad \|UAV^{\mathrm{T}}\| = \|A\| \tag{4.5.2}$$

- **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 $\mathbb{R}^n$

$$\forall\, P \in \mathscr{P}, \quad \Phi(Px) = \Phi(x) \tag{4.5.3}$$

- There is a remarkable link between UIN and SGF. 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

$$\|A\|_\Phi = \Phi(\sigma_1, \ldots, \sigma_n) \tag{4.5.4}$$

Then, $\|.\|_\Phi$ is a UIN. 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 $\|.\|$

$$\|\mathrm{diag}\,(\alpha_1 - \beta_1, \ldots, \alpha_p - \beta_p)\| \leq \|A - B\| \tag{4.5.5}$$

- **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

$$\|A - B\|_\Phi \geq \Phi(0, \ldots, 0, \sigma_{r+1}, \ldots, \alpha_p) \tag{4.5.6}$$

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

$$A_r = U\Sigma_r V^* \tag{4.5.7}$$

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 [Sch60] as well for an algebraic survey.

5 PCA with Instrumental Variables

Notations

Here are the notations chosen for this section. They are as compatible as possible with notations selected in other sections, especially PCA. There are explained along the text.

Symbol	in space	What it is
A	$\mathbb{R}^{n\times p}$	matrix to be analyzed
A_r	$\mathbb{R}^{n\times p}$	best approximation of A of rank r with constraints
F	$\subset \mathbb{R}^n$	subspace of constraints on principal components
H	$\subset \mathbb{R}^p$	subspace of constraints on principal axis
I	$\mathbb{R}^{n\times m}$	matrix of instrumental variables
Λ	$\mathbb{R}^{q\times q}$	diagonal matrix of eigenvalues of PCA-IV
m	$\mathbb{N}$	dimension of F
P_{F}	$\mathbb{R}^{p\times p}$	projector on F
P_{H}	$\mathbb{R}^{n\times n}$	projector on H
$\mathcal{P}_{\mathrm{F}\otimes\mathrm{H}}$	$\mathbb{R}^{np\times np}$	projector on $F\otimes H$
q	$\mathbb{N}$	dimension of H
r	$\mathbb{N}$	prescribed rank for best approximation
U_{F}	$\mathbb{R}^{n\times m}$	matrix of an orthonormal basis of F
V	$\mathbb{R}^{p\times q}$	matrix of principal axis of PCA-IV of (A,F,H)
V_{H}	$\mathbb{R}^{p\times q}$	matrix of an orthonormal basis of H
Y	$\mathbb{R}^{n\times m}$	principal components of PCA-IV of (A,F,H)
Y_r	$\mathbb{R}^{n\times r}$	first r principal components of PCA-IV of (A,F,H)

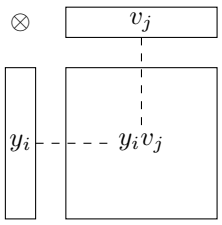
Here, we use tensor notation for PCA. For sake of clarity for those readers not familiar with those notations, we set the problem with classical notations like yv^{T} for a rank one matrix as a starting point. Let $y\in\mathbb{R}^n$ and $v\in\mathbb{R}^p$. We can accept as a definition that $\otimes$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

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

defined by

$$y\otimes v := yv^{\mathrm{T}}$$

which can be illustrated as



If $y=(y_1,\ldots,y_n)\in\mathbb{R}^n$ and $v=(v_1,\ldots,v_p)\in\mathbb{R}^p$, then $y\otimes v\in\mathbb{R}^{n\times p}$ and

$$(y\otimes v)_{ij}=y_iv_j$$

- Let $A \in \mathbb{R}^{n\times p}$. A rank$-r$ best approximation of A can be written as

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

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 the principal components $(y_a)_a$ or principal axis $(v_a)_a$, i.e. that they live in given subspaces respectively $F \subset \mathbb{R}^n$ and $H \subset \mathbb{R}^p$.

Usually, those spaces are given as spanned by a set of vectors in respectively $\mathbb{R}^n$ (for F, constraint on y) and $\mathbb{R}^p$ (for H, constraints on v). A classical situation is when there is no constraint on the axis v, but only on the components y, and F is spanned by a $n \times q$ matrix denoted I, the columns of which are called *instrumental variables*.

5.1 Setting the problem

Let us suppose that $\dim F = m$ and $\dim H = q$. `PCAiv` can be stated as

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

Technically, it is possible to specify this problem with base of F and H having been selected. We assume here that they are orthonormal. If they are not orthonormal, it is possible to build an orthonormal base by QR decomposition, for example with Gram-Schmidt orthogonalization procedure (which can be numerically unstable but is easy to implement), or using Householder reflections (more stable).

5.2 Solving the problem

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

- Before stating the main result, we need to recall some elementary results on linear projectors. Let $F \subset \mathbb{R}^n$. Then, the projector on F is denoted $\mathcal{P}_{\mathrm{F}}$. Let $U_{\mathrm{F}} = (u_1, \ldots, u_m)$ be an orthonormal basis of F columnwise. Then

$$\mathcal{P}_{\mathrm{F}} = U_{\mathrm{F}}\, U_{\mathrm{F}}^{\mathrm{T}} \tag{5.2.1}$$

Indeed, if $x \in \mathbb{R}^n$, we have

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

Then

$$\mathcal{P}_{\text{F}} = \sum_i u_i \otimes u_i = U_{\text{F}}\, U_{\text{F}}^{\text{T}} \tag{5.2.2}$$

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

$$\mathcal{P}_{F\otimes H} A = U_{\text{F}}\, U_{\text{F}}^{\text{T}}\, A\, V_{\text{H}}\, V_{\text{H}}^{\text{T}} \tag{5.2.3}$$

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

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

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

Let us recall that A, F, H are fixed, hence so are $\mathcal{P}, \mathcal{P}^\perp$. Then, 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 $\mathcal{P}(A) = A_{F\otimes H}$. □

• This leads to a solution for `PCAiv`. Let $(u_1, \ldots, u_m)$ be an orthonormal basis for $F \subset \mathbb{R}^n$, and $(v_1, \ldots, v_q)$ for $H \subset \mathbb{R}^p$, with $m < n$ and $q < p$. Let U_{F} be the $n \times m$ matrix with columns u_i and V_{H} be the $p \times q$ matrix with columns v_j. Then, the orthogonal projector $\mathcal{P}_{\text{H}}$ from $\mathbb{R}^p$ to H is given by matrix

$$P_{\text{H}} = V_{\text{H}}\, V_h^{\text{T}}$$

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

$$P_{\text{F}} = U_{\text{F}}\, U_{\text{F}}^{\text{T}}$$

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

$$\mathcal{P}_{F\otimes H}(A) = U_{\text{F}}\, U_{\text{F}}^{\text{T}}\, A\, V_{\text{H}}\, V_{\text{H}}^{\text{T}} \tag{5.2.4}$$

Hence the algorithm

Algorithm 7 Pseudocode for `PCAiv`$(A, U_{\text{F}}, V_{\text{H}})$

input: $A \in \mathbb{R}^{n \times p}$
input: U_{F} orthonormal, $F = \operatorname{span} U_{\text{F}} \subset \mathbb{R}^n$,
input: V_{H} orthonormal, $H = \operatorname{span} V_{\text{H}} \subset \mathbb{R}^p$,
input $r < p$
compute $P_{\text{F}} = U_{\text{F}} U_{\text{F}}^{\text{T}}$
compute $P_{\text{H}} = V_{\text{H}} V_{\text{H}}^{\text{T}}$
compute $T = P_{\text{F}} A P_{\text{H}}$
compute $(Y, \Lambda, V) = \text{PCA\_CORE}(T)$
return Y, Λ, V

• In this algorithm, the SVD is run in the PCA of $T = P_{\text{F}} A P_{\text{H}} \in \mathbb{R}^{n \times p}$, same dimensions as A. However, as $T \in F \otimes H$, with $\dim F = m$ and $\dim H = q$, the rank of T is $q < p$ (if we assume that $q \leq m$). Here, we show how to run SVD on a matrix of dimension lower than $n \times p$. We start from

$$\begin{aligned} T &= P_{\text{F}} A P_{\text{H}} \\ &= (U_{\text{F}} U_{\text{F}}^{\text{T}}) A (V_{\text{H}} V_{\text{H}}^{\text{T}}) \\ &= U_{\text{F}} (U_{\text{F}}^{\text{T}} A V_{\text{H}}) V_{\text{H}}^{\text{T}} \\ &= U_{\text{F}} T' V_{\text{H}}^{\text{T}} \end{aligned}$$

denoting

$$T' = U_{\text{F}}^{\text{T}} A V_{\text{H}} \qquad \in \mathbb{R}^{m \times q}$$

Let $(U'_{\text{T}}, \Sigma'_{\text{T}}, V'_{\text{T}})$ be the SVD of T'

$$T' = U'_{\text{T}} \Sigma'_{\text{T}} V_{\text{T}}'^{\text{T}} \tag{5.2.5}$$

with (this will be useful soon), assuming $m \geq q$ here for sake of simplicity

$$\begin{cases} U_{\text{F}} & \in \mathbb{R}^{n \times m} \\ U'_{\text{T}} & \in \mathbb{R}^{m \times q} \\ \Sigma'_{\text{T}} & \in \mathbb{R}^{q \times q} \\ V'_{\text{T}} & \in \mathbb{R}^{q \times q} \\ V_{\text{H}} & \in \mathbb{R}^{p \times q} \end{cases}$$

Then

$$\begin{aligned} T &= U_{\text{F}} U'_{\text{T}} \Sigma'_{\text{T}} V_{\text{T}}'^{\text{T}} V_{\text{H}}^{\text{T}} \\ &= (U_{\text{F}} U'_{\text{T}}) \Sigma'_{\text{T}} (V_{\text{H}} V'_{\text{T}})^{\text{T}} \end{aligned} \tag{5.2.6}$$

and $(U_{\text{F}} U'_{\text{T}}\,,\, \Sigma'_{\text{T}}\,,\, V_{\text{T}}'^{\text{T}} V_{\text{H}}^{\text{T}})$ is the SVD of T because

$$\begin{cases} (U_{\text{F}} U'_{\text{T}})^{\text{T}} (U_{\text{F}} U'_{\text{T}}) &= U_{\text{T}}'^{\text{T}} U_{\text{F}}^{\text{T}} U_{\text{F}} U'_{\text{T}} \\ &= U_{\text{T}}'^{\text{T}} \mathbb{I}_m U'_{\text{T}} \\ &= U_{\text{T}}'^{\text{T}} U'_{\text{T}} \\ &= \mathbb{I}_q \end{cases}$$

so $U_{\text{F}} U'_{\text{T}}$ is orthonormal. Similarly

$$\begin{cases} (V_{\text{H}} V'_{\text{T}})^{\text{T}} (V_{\text{H}} V'_{\text{T}}) &= V_{\text{T}}'^{\text{T}} V_{\text{H}}^{\text{T}} V_{\text{H}} V'_{\text{T}} \\ &= V_{\text{T}}'^{\text{T}} \mathbb{I}_q V'_{\text{T}} \\ &= V_{\text{T}}'^{\text{T}} V'_{\text{T}} \\ &= \mathbb{I}_q \end{cases}$$

and $V_{\text{H}} V_{\text{T}}'^{\text{T}}$ is orthonormal too. This small calculation replaces the SVD of $T \in \mathbb{R}^{n \times p}$ in $\mathcal{O}(n^2 p)$ by the SVD of $T' \in \mathbb{R}^{m \times q}$ in $\mathcal{O}(m^2 q)$. This leads to the following algorithm:

Algorithm 8 Second Pseudocode for `PCAiv`$(A, U_{\text{F}}, V_{\text{H}})$

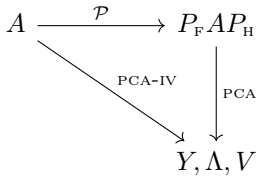
input: $A \in \mathbb{R}^{n \times p}$
input: $F = \operatorname{span} U_{\text{F}} \subset \mathbb{R}^n$, U_{F} orthonormal
input: $H = \operatorname{span} V_{\text{H}} \subset \mathbb{R}^p$, V_{H} orthonormal
input $r < p$
compute $T' = U_{\text{F}}^{\text{T}} A V_{\text{H}}$
do the SVD of T': $T' = U_{\text{T}}' \Sigma_{\text{T}}' V_{\text{T}}'^{\text{T}}$
compute $U = U_{\text{F}} U_{\text{T}}'$
compute $V = V_{\text{H}} V_{\text{T}}'$
compute $\Lambda = \Sigma_{\text{T}}'^2$
compute $Y = U \, \Sigma_{\text{T}}'$
return Y, Λ, V

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:



This leads to

1. another interpretation of `PCAiv`
2. an estimate of the quality of the `PCAiv`

as follows.

• Let us give another interpretation of `PCAiv` on the example of a constraint on the principal components only. Let $a \in \mathbb{R}^n$. The projection of a on F is given by $P_{\text{F}} a$. Let now $A \in \mathbb{R}^{n \times p}$. The matrix $\widetilde{A} = P_{\text{F}} A$ is in $\mathbb{R}^{n \times p}$, like A, and its column j is the projection of column j of A on F. It is the regression of the column j by the basis vectors of F, given by the columns of U_{F}. The first principal component of $\widetilde{A}$ is the best summary of A by one single linear combination

of the columns of U_{F}. It is the best same linear regression applied to all columns of A. In this interpretation, `PCAiv` is equivalent to PLS.

- For the `PCA`, classical estimators of the quality of the `PCA` can be used. Let us denote Y_r the matrix of r first principal components of $P_{\text{F}}AP_{\text{H}}$ on which `PCA` is run, and

$$\|Y_r\| = \rho_r \|P_{\text{F}}AP_{\text{H}}\| \tag{5.3.1}$$

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

- The quality of the projection can be quantified by

$$\|P_{\text{F}}AP_{\text{H}}\| = \theta \|A\| \tag{5.3.2}$$

(θ is the cosine of the angle between A and $P_{\text{F}}AP_{\text{H}}$).

- 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 F and H are given by their basis U_{F} and V_{H} respectively. Let us now suppose that F and H are respectively spanned by the columns of U'_{F} and V'_{H}, which are no longer assumed to be orthonormal. Then

$$P_{\text{F}} = U'_{\text{F}}(U'^{\text{T}}_{\text{F}}U'_{\text{F}})^{-1}U'^{\text{T}}_{\text{F}}, \qquad P_{\text{H}} = V'_{\text{H}}(V'^{\text{T}}_{\text{H}}V'_{\text{H}})^{-1}V'^{\text{T}}_{\text{H}}$$

and equation (5.2.3) reads

$$\begin{aligned} T &= P_{\text{F}}AP_{\text{H}} \\ &= U'_{\text{F}}(U'^{\text{T}}_{\text{F}}U'_{\text{F}})^{-1}U'^{\text{T}}_{\text{F}}\ A\ V'_{\text{H}}(V'^{\text{T}}_{\text{H}}V'_{\text{H}})^{-1}V'^{\text{T}}_{\text{H}} \end{aligned}$$

- However, this equation, even if correct, is not efficient for numerical analysis where it should be avoided, because of the cost of three products and one inversion per projector. It is far more efficient to work with orthogonal basis of F and H, by

1. building an orthonormal basis U_{F} of F and an orthonormal basis V_{H} of H by QR decomposition or Gram-Schmidt orthonormalisation,

2. building the projectors $P_{\text{F}} = U_{\text{F}} U_{\text{F}}^{\text{T}}$ and $P_{\text{H}} = V_{\text{H}} V_{\text{H}}^{\text{T}}$

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], and widely used in ecology for example (see [LSBB91]). `PCAiv` is equivalent to PLS.

6 PCA with metrics on rows and columns

Notations

symbol	in space	meaning
$\|.\|_{\text{N}}$		norm induced by N in $\mathbb{R}^n$
$\|.\|_{\text{P}}$		norm induced by P in $\mathbb{R}^p$
$\|.\|_{\text{T}}$		norm induced by T in $\mathbb{R}^{n\times p}$
A	$\mathbb{R}^{n\times p}$	matrix to be analyzed
B	$\mathbb{R}^{n\times p}$	matrix for calculations: $B = MAQ$
Λ	$\mathbb{R}^{p\times p}$	eigenvalues of $B^{\text{T}}B$
M	$\mathbb{R}^{n\times n}$	unique SDP with $M^2 = N$
N	$\mathbb{R}^{n\times n}$	SDP defining an inner product in $\mathbb{R}^n$
$\mathcal{P}_v$	$\mathcal{L}(\mathbb{R}^p)$	projector on $\mathbb{R}v$ in $\mathbb{R}^p$ for inner product P
$\mathcal{P}_{\text{F}}$	$\mathcal{L}(\mathbb{R}^p)$	projector on $F \subset \mathbb{R}^p$ in $\mathbb{R}^p$ for inner product P
P	$\mathbb{R}^{p\times p}$	SDP defining an inner product in $\mathbb{R}^p$
Q	$\mathbb{R}^{p\times p}$	unique SDP matrix with $P = Q^2$
T	$\mathbb{R}^{np\times np}$	SDP defining an inner product in $\mathbb{R}^{n\times p}$
V	$\mathbb{R}^{p\times p}$	principal axis of A for inner product defined by (N, P)
w	$\mathbb{R}^p$	weights for an inner product in R^p
W	$\mathbb{R}^{n\times p}$	principal components of B with standard inner product
X	$\mathbb{R}^{p\times p}$	principal axis of B with standard inner product
Y	$\mathbb{R}^{n\times p}$	principal components of A with inner product defined by (N, P)
Z	$\mathbb{R}^{np\times np}$	unique SDP with $Z^2 = T$; $Z(A) = MAQ$

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 endowed with the canonical inner product

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

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

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

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

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

- The norm $\|.\|_{\mathrm{P}}$ is defined as

$$\|x\|_{\mathrm{P}}^2 = \langle x,x\rangle_{\mathrm{P}} = \langle x,Px\rangle \tag{6.1.4}$$

If $P=\operatorname{diag} w$, this yields

$$\|x\|_w^2 = \sum_i w_i x_i^2 \tag{6.1.5}$$

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

$$\langle x,y\rangle_{\mathrm{P}} = \langle Qx\,,\,Qy\rangle \tag{6.1.6}$$

and

$$\|x\|_{\mathrm{P}} = \|Qx\| \tag{6.1.7}$$

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 $P = Q^2$.

• **Projection operator with metrics:** We define here the projection operator in $\mathbb{R}^p$ endowed with metrics defined by P. Let $v \in \mathbb{R}^p$ with $\|v\|_{\mathrm{P}} = 1$. Let us denote by $\mathcal{P}_v$ the projection operator in $\mathbb{R}^p$ on $\mathbb{R}v$

$$\begin{array}{ccc} \mathbb{R}^p & \xrightarrow{\mathcal{P}_v} & \mathbb{R}v \\ x & \longrightarrow & \lambda v \end{array}$$

$\mathcal{P}_v x$ is the vector λv such that $\|x - \lambda v\|_{\mathrm{P}}$ is minimal. As $\|x - \lambda v\|_{\mathrm{P}}^2 = \|x\|_{\mathrm{P}}^2 + \lambda^2 - 2\lambda \langle x, v \rangle_{\mathrm{P}}$, and the unknown is λ, this yields

$$\lambda = \langle x, v \rangle_{\mathrm{P}} = \langle x, Pv \rangle \tag{6.1.8}$$

Then

$$\mathcal{P}_v x = \langle x, Pv \rangle\, v = (v \otimes Pv).x \tag{6.1.9}$$

and

$$\mathcal{P}_v = v \otimes Pv \tag{6.1.10}$$

Let us note that $\mathcal{P}_v \neq Qv \otimes Qv$ (indeed, $(Qv \otimes Qv)x = \langle Qv, x \rangle Qv \in \mathbb{R}Qv \notin \mathbb{R}v$). For a general subspace

$$F \subset \mathbb{R}^p = \mathrm{span}\,(v_1, \ldots, v_m)$$

with $(v_i)_i$ a basis of F orthonormal for P, the same calculation leads to

$$\mathcal{P}_{\text{F}} = \sum_i v_i \otimes P v_i \tag{6.1.11}$$

- 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

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

- 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 by a SDP matrix T, 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

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

with

$$\langle a \otimes b\,,\, x \otimes y \rangle_{\text{T}} = \langle Z(a \otimes b)\,,\, Z(x \otimes y) \rangle \tag{6.1.12}$$

As

$$\begin{array}{rcl} \langle a \otimes b\,,\, x \otimes y \rangle_{\text{T}} & = & \langle Ma, Mx \rangle \langle Qb, Qy \rangle \\ & = & \langle Ma \otimes Qb\,,\, Mx \otimes Qy \rangle \end{array} \tag{6.1.13}$$

(6.1.12) is fulfilled by selecting

$$\begin{array}{rcl} Z(a \otimes b) & = & Ma \otimes Qb \\ & = & M(a \otimes b)Q \end{array} \tag{6.1.14}$$

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

$$Z(A) = MAQ \tag{6.1.15}$$

by linearity.

- So, we have

$$\|A\|_{\text{T}} = \|MAQ\| \tag{6.1.16}$$

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

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

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

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

- 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 9 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

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

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

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

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

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

$$MAQ = MA$$

Hence

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

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 $\mathbb{R}^p$

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

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 $P = \text{diag}\left(\frac{1}{\|a_{\bullet j}\|}\right)$ in $\mathbb{R}^p$. So $MAQ = A'$, and `PCA` of A' is run.

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

6.6 Analysis of a matrix with metrics and weights: a geometric approach

The geometric school of linear dimension reduction has developed a framework to analyse a matrix $A \in \mathbb{R}^{n\times p}$ with a metric defined by P in $\mathbb{R}^p$ and weights $(w_i)_i$ on the individual. The point cloud is made by rows $(a_{i*})_i \in \mathbb{R}^p$ of A. It is presented here and will be very useful for understanding the geometric approach of `CoA` .

- **Setting the problem:** Rows of A (points of the point cloud) are in $\mathbb{R}^p$. We set the problem for the best projection of the point cloud on a one-dimensional space in $\mathbb{R}^p$ spanned by a vector v with $\|v\|_{\text{P}} = 1$. The aim is to compute v. The projection of a_{i*} on $\mathbb{R}v$ is given by

$$\mathcal{P}_v a_{i*} = \langle Pa_{i*}, v\rangle v$$

and its norm is

$$\|\mathcal{P}_v a_{i*}\|^2_{\text{P}} = \langle Pa_{i*}, v\rangle^2$$

Let

$$\mathcal{I} = \sum_i \|\mathcal{P}_v a_{i*}\|^2_{\text{P}} = \sum_i \langle Pa_{i*}, v\rangle^2$$

be the inertia of the point cloud attached to A for inner product P. So, first step without weights but with metrics defined by P is to find v with $\|v\|_{\text{P}} = 1$ such that $\mathcal{I}$ is maximal. Final step is to introduce the weights, and define

$$\mathcal{I} = \sum_i w_i \langle Pa_{i*}, v\rangle^2$$

as the inertia of the point cloud with inner product defined by P and weights on rows defined by w. Then, the geometrical approach can be set as

given	a matrix $A \in \mathbb{R}^{n\times p}$
with	row i denoted $a_{i*} \in \mathbb{R}^p$
an inner product	P in $\mathbb{R}^p$
a set of row weights	$w \in \mathbb{R}^n$
find	a vector $v \in \mathbb{R}^p$
with	$\|v\|_{\mathrm{P}} = 1$
such that	$\mathcal{I} = \sum_i w_i \ \|\mathcal{P}_v a_{i*}\|^2$ is maximal
where	$\mathcal{P}_v$ is the projection on $\mathbb{R}v$ in $\mathbb{R}^p$

- **Solving the problem:** Let us note first that

$$\|\mathcal{P}_v a_{i*}\|^2 = \langle Pa_{i*}, v\rangle^2 \tag{6.6.1}$$

Then

$$\mathcal{I} = \sum_i w_i \ \langle Pa_{i*}, v\rangle^2 = \sum_i \langle \sqrt{w_i}\, P\, a_{i*} \ , \ v\rangle^2 \tag{6.6.2}$$

Let us observe that

→ the vectors $Pa_{i*} \in \mathbb{R}^p$ are the row vectors of matrix AP (indeed, P is symmetric by definition)

→ the terms $\sqrt{w_i}Pa_{i*}$ are the rows of matrix $D_w^{1/2}AP$ if D_w is the $n \times n$ diagonal matrix with w in the diagonal

Then $\mathcal{I}$ can be rewritten as

$$\mathcal{I} = \|D_w^{1/2}APv\|^2 \tag{6.6.3}$$

So, the problem can be restated as

given	A, P, w as above,
find	$v \in \mathbb{R}^p$
with	$\|v\|_{\mathrm{P}}^2 = 1$
such that	$\mathcal{I} = \|D_w^{1/2}APv\|^2$ is maximal

To solve this with Lagrange multipliers (an optimum with constraints), let us denote

$$H = D_w^{1/2}AP \tag{6.6.4}$$

Then, $\mathcal{I} = \|Hv\|^2$, and

$$\frac{\partial \mathcal{I}}{\partial v} = 2H^{\mathrm{T}}H, \qquad \frac{\partial \|v\|_{\mathrm{P}}^2}{\partial v} = 2Pv \tag{6.6.5}$$

so

$$H^{\mathrm{T}}Hv = Pv \tag{6.6.6}$$

This can be written

$$PA^{\mathrm{T}}D_w^{1/2}D_w^{1/2}APv = PA^{\mathrm{T}}D_wAPv = \lambda Pv \tag{6.6.7}$$

and Pv is an eigenvector of $PA^{\mathrm{T}}D_wA$. As P is invertible (it is a SDP matrix), this yields

$$A^{\mathrm{T}}D_wAPv = \lambda v \tag{6.6.8}$$

and v is an eigenvector of $A^{\mathrm{T}}D_wAP$.

Notes and references: This section is adapted from [LMP00, 1.1.6] where it is presented as a diversification of the general analysis (PCA). Similar approaches can be found in [LMT77, LMF82].

6.7 `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 $\mathbb{R}^n$:

$$\begin{cases} y_j \in E \subset \mathbb{R}^n \\ v_j \in F \subset \mathbb{R}^p \end{cases} \tag{6.7.1}$$

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

$$\mathbb{R}^n \xrightarrow{R=UU^{\mathrm{T}}} E, \qquad \mathbb{R}^p \xrightarrow{S=VV^{\mathrm{T}}} F \tag{6.7.2}$$

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

$$\begin{aligned} x' &= \langle x, v\rangle_{\mathrm{P}} v \\ &= \langle x, Pv\rangle v \\ &= \langle Pv, x\rangle v \\ &= (v \otimes Pv).x \end{aligned} \tag{6.7.3}$$

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

$$\begin{aligned} x' &= \sum_a \langle x, Pv_a\rangle v_a \\ &= \sum_a (v_a \otimes Pv_a)x \end{aligned} \tag{6.7.4}$$

and the projector is

$$\begin{aligned} S &= \sum_a v_a \otimes Pv_a \\ &= V(PV)^{\mathrm{T}} \\ &= VV^{\mathrm{T}}P \\ &= PVV^{\mathrm{T}} \end{aligned} \tag{6.7.5}$$

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

$$R = UU^{\mathrm{T}}N \tag{6.7.6}$$

and

$$\begin{aligned} A' &= RAS \\ &= UU^{\mathrm{T}}NAPVV^{\mathrm{T}} \end{aligned} \tag{6.7.7}$$

• 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

Notations

Some notations for Correspondence Analysis are standard and specific to this method. They are given here and explained along the chapter.

symbol	in space	meaning
A	$\mathbb{R}^{I\times J}$	matrix of frequencies (T/n_{++})
α_{ij}	$[0,1]\subset\mathbb{R}$	general term of A
c	$\mathbb{R}^J$	vector of marginals of A by columns: $(c_1,\ldots,c_J)$
c_j	$\mathbb{R}$	marginal of column j of A: $(\sum_i \alpha_{ij})$
D_c	$\mathbb{R}^{J\times J}$	diagonal matrix with elements c_j
D_r	$\mathbb{R}^{I\times I}$	diagonal matrix with elements r_i
i	$\mathbb{N}$	indices of rows
I	$\mathbb{N}$	number of rows of T
j	$\mathbb{N}$	indices of columns
J	$\mathbb{N}$	number of columns of T
n_{ij}	$\mathbb{N}$	number of item in class i (row) and j (column)
n_{i+}	$\mathbb{N}$	sum of terms in row i of T
n_{+j}	$\mathbb{N}$	sum of terms in column j of T
n_{++}	$\mathbb{N}$	sum of terms in T
r	$\mathbb{R}^I$	vector of marginals of A by rows: $(r_1,\ldots,r_I)$
r_i	$\mathbb{R}$	marginal of row i of A: $(\sum_j \alpha_{ij})$
T	$\mathbb{R}^{I\times J}$	contingency table

A remarkable application of the `PCA` with weights on rows and columns is the development 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

$$T \in \mathbb{R}^{I\times J} \simeq \mathbb{R}^I \otimes \mathbb{R}^J$$

It is standard to denote

$$n_{i+} = \sum_j n_{ij}, \quad n_{+j} = \sum_i n_{ij}, \qquad n_{++} = \sum_{i,j} n_{ij} = \sum_i n_{i+} = \sum_j n_{+j}$$

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 n_{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

$$A = \frac{T}{n_{++}} \tag{7.1.1}$$

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

$$\left\{ \begin{array}{rcccl} r_i & = & \displaystyle\sum_j \alpha_{ij} & = & \alpha_{i+} \\ c_j & = & \displaystyle\sum_i \alpha_{ij} & = & \alpha_{+j} \end{array} \right. \tag{7.1.3}$$

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, if X_r is the ransom variable for rows and X_c for columns

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

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

We have

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

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 10 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: geometric approach

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. This is the geometric approach. It emphasizes that two point clouds, and not simply one, can be built: one for rows, and one for columns, and `CoA` can be seen as a simultaneous analysis of both.

• Let us recall that if T is a contingency table, F (the matrix of frequencies) is built from T by dividing it by the sum of all its elements:

$$T = (n_{ij})_{i,j} \longrightarrow n_{++} = \sum_{i,j} n_{ij} \longrightarrow F \,:\, f_{ij} = \frac{n_{ij}}{n_{++}}$$

To comply with standard notations in classical textbooks (see below), we denote by F, and not A, the matrix of frequencies. We will denote

for row		for columns
$f_{i+} = \sum_j f_{ij}$	and	$f_{+j} = \sum_i f_{ij}$
$f_{i*} = (f_{i1}, \ldots, f_{iJ}) \in \mathbb{R}^J$	and	$f_{*j} = (f_{1j}, \ldots, f_{Ij}) \in \mathbb{R}^I$
$r = (f_{1*}, \ldots, f_{I*}) \in R^I$	and	$c = (f_{*1}, \ldots, f_{*J}) \in \mathbb{R}^J$

- Two point clouds are classically attached to A

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

$$x_i = \left[\frac{f_{ij}}{f_{i+}}\right]_j \in \mathbb{R}^{\mathrm{J}} \tag{7.4.1}$$

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

$$y_j = \left[\frac{f_{ij}}{f_{+j}}\right]_i \in \mathbb{R}^{\mathrm{I}} \tag{7.4.2}$$

- Then, metrics with diagonal matrices respectively D_c^{-1} for R^{J} and D_r^{-1} for $\mathbb{R}^{\mathrm{I}}$ are selected, with D_c being the $J \times J$ diagonal matrix of term $1/f_{+j}$ and D_r the $I \times I$ diagonal matrix of term f_{i+}. Hence, distances between points x_i and x_k in $\mathbb{R}^{\mathrm{J}}$ are computed as

$$d^2(x_i, x_k) = \sum_j \frac{1}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{kj}}{f_{k+}}\right)^2 \tag{7.4.3}$$

and between points y_j and y_ℓ in $\mathbb{R}^{\mathrm{I}}$ as

$$d^2(y_j, y_\ell) = \sum_i \frac{1}{f_{i+}} \left(\frac{f_{ij}}{f_{+j}} - \frac{f_{i\ell}}{f_{+\ell}}\right)^2 \tag{7.4.4}$$

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 $\forall \mathrm{J},\ f_{ij}/f_{i+} = f_{i'j}/f_{i'+}$), then it is logical to lump them together into a single category, and this must not modify the distances between row profiles.

- Let us recall that

$$r = (f_{1+}, \ldots, f_{I+}) \in \mathbb{R}^I, \qquad \text{and} \qquad c = (f_{+1}, \ldots, f_{+J}) \in \mathbb{R}^J,$$

so

$$r_i = f_{i+} \qquad \text{and} \qquad c_j = f_{+j}$$

It is then standard to define both point clouds as (see [Gre84, sect. 4.1], [LMF82, sect. IV.5.], [Sap90, sect. 10.1],[LMP00, sect. 1.3.3])

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

R is the point cloud of row profiles of F, and C of its columns profiles. It is easy to get lost in the indices, the rows, the columns, spaces $\mathbb{R}^I$ and $\mathbb{R}^J$. A row profile is in $\mathbb{R}^J$ and its coefficients are indexed by j; a column profile is in $\mathbb{R}^I$ and its coefficients are indexed by i. With coefficients, this yields

$$R_{ij} = \frac{f_{ij}}{f_{i+}}, \qquad C_{ij} = \frac{f_{ij}}{f_{+j}} \tag{7.4.5}$$

- The centroids $g^{(\mathrm{R})}$ of R and $g^{(\mathrm{C})}$ of C are respectively c and r (and not r and c). Indeed,

$$g_j^{(\mathrm{R})} = \sum_i f_{i+} \frac{f_{ij}}{f_{i+}} = \sum_i f_{ij} = f_{+j} = c_j \tag{7.4.6}$$

and

$$g_i^{(\mathrm{C})} = \sum_j f_{+j} \frac{f_{ij}}{f_{+j}} = \sum_j f_{ij} = f_{i+} = r_i \tag{7.4.7}$$

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, (explained step by step ...)

$$\begin{aligned}
I_{\mathrm{R}}^2 &= \sum_i r_i \left\| R_{i*} - c \right\|_{D_c^{-1}}^2 && R_{i*} = (R_{i1}, \ldots, R_{iJ}) \\
&= \sum_i f_{i+} \left\| \frac{f_{i*}}{f_{i+}} - c \right\|_{D_c^{-1}}^2, && R_{i*} = \tfrac{f_{i*}}{f_{i+}},\ r_i = f_{i+} \\
&= \sum_i f_{i+} \left(\frac{1}{f_{+j}} \sum_j \left(\frac{f_{ij}}{f_{i+}} - f_{+j} \right)^2 \right), && c_j = f_{+j},\ D_c^{-1} = \mathrm{diag}\left(\tfrac{1}{f_{+j}} \right) \\
&= \sum_i f_{i+} \left(\sum_j \left(\frac{f_{ij}}{f_{i+}\sqrt{f_{+j}}} - \sqrt{f_{+j}} \right)^2 \right) && \sqrt{f_{+j}} = \tfrac{f_{+j}}{\sqrt{f_{+j}}} \\
&= \sum_{i,j} \left(\sqrt{f_{i+}} \frac{f_{ij}}{f_{i+}\sqrt{f_{+j}}} - \sqrt{f_{i+}}\sqrt{f_{+j}} \right)^2 \\
&= \sum_{i,j} \left(\frac{f_{ij} - f_{i+}f_{+j}}{\sqrt{f_{i+}f_{+j}}} \right)^2
\end{aligned} \tag{7.4.8}$$

where we recognize the χ^2 norm of $F = (f_{ij})_{i,j}$, or the norm of F with metrics defined by D_c^{-1} in $\mathbb{R}^J$ (space of rows) and by D_r^{-1} in $\mathbb{R}^I$ (space of columns). Its partition with concentration of the inertia on the first components is the `CoA` of F.

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

$$I_{\mathrm{C}}^2 = \sum_{i,j} \left(\frac{f_{ij} - f_{i+}f_{+j}}{\sqrt{f_{i+}f_{+j}}} \right)^2 = I_{\mathrm{R}}^2 \tag{7.4.9}$$

Both inertia are equal, and the analysis of both point clouds is one geometric guise of the `CoA` of F.

- To see this, one can use the developments presented in section 6.6. Let us first present a translation of the notations between section 6.6 and here:

`PCAmet` with weight		`CoA` on R	`CoA` on C
A	$\longleftrightarrow$	$R = D_r^{-1}F$	$C = D_c^{-1}F^{\mathrm{T}}$
P	$\longleftrightarrow$	D_c^{-1}	D_r^{-1}
w	$\longleftrightarrow$	r	c

Let us recall that the first principal axis u is solution of

$$H^{\mathrm{T}}Hu = Pu \tag{7.4.10}$$

with

$$H = D_w^{1/2}AP \tag{7.4.11}$$

so

$$H^{\mathrm{T}}Hu = PA^{\mathrm{T}}D_wAPu = \lambda Pu \tag{7.4.12}$$

and, as P is ivertible as a SDP matrix, u is solution of

$$A^{\mathrm{T}}D_wAPu = \lambda u \tag{7.4.13}$$

We then have, for `CoA` on row profiles

$\rightarrow\ H = D_r^{1/2}.D_r^{-1}F.D_c^{-1} = D_r^{-1/2}FD_c^{-1}$

$\rightarrow\ H^{\mathrm{T}}H = (D_c^{-1}F^{\mathrm{T}}D_r^{-1/2}).(D_r^{-1/2}FD_c^{-1}) = D_c^{-1}F^{\mathrm{T}}D_r^{-1}FD_c^{-1}$

and, after simplification by D_c^{-1}, u is solution of

$$F^{\mathrm{T}}D_r^{-1}FD_c^{-1}u = \lambda u \tag{7.4.14}$$

and for `CoA` on column profiles

$\rightarrow\ H = D_c^{1/2}.D_c^{-1}F^{\mathrm{T}}.D_r^{-1} = D_c^{-1/2}F^{\mathrm{T}}D_r^{-1}$

$\rightarrow\ H^{\mathrm{T}}H = (D_r^{-1}FD_c^{-1/2}).(D_c^{-1/2}F^{\mathrm{T}}D_r^{-1}) = D_r^{-1}FD_c^{-1}F^{\mathrm{T}}D_r^{-1}$

and, after simplification by D_r^{-1}, first axis v is solution of

$$FD_c^{-1}F^{\mathrm{T}}D_r^{-1}v = \lambda v \tag{7.4.15}$$

(see [LMP00, p. 83], with, here again, a translation of notations).

- If we multiply equation (7.4.14) on the left by D_c^{-1} and set $u' = D_c^{-1}u$, we have

$$D_c^{-1}F^{\mathrm{T}}D_r^{-1}FD_c^{-1}u = \lambda D_c^{-1}u \tag{7.4.16}$$

or

$$CRu' = \lambda u' \tag{7.4.17}$$

Similarily, multiplying equation (7.4.15) left by D_r^{-1} and setting $v' = D_r^{-1}v$ yields

$$RCv' = \lambda v' \tag{7.4.18}$$

This yields

$$\left\{ \begin{array}{rcl} v' & = & Ru' \\ u' & = & Cv' \end{array} \right. \tag{7.4.19}$$

or

$$\left\{ \begin{array}{rcl} v & = & D_r^{-1}RD_c^{-1}u \\ u & = & D_c^{-1}CD_r^{-1}v \end{array} \right. \tag{7.4.20}$$

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] and [LMP00, sect. 1.3] for the geometric interpretation.

8 Canonical Correlation Analysis

Notations

The following notations have been chosen to be as compatible as possible with those in other sections, especially section 6 (`PCA` with metrics).

symbol	space	meaning
A	$\mathbb{R}^{n\times p}$	one of the two matrices to be analysed
B	$\mathbb{R}^{n\times q}$	one of the two matrices to be analysed
M	$\mathbb{R}^{p\times p}$	$M = N^{1/2}$
N	$\mathbb{R}^{p\times p}$	SDP matrix defining an inner product in $\mathbb{R}^p$; $N = (A^{\mathrm{T}}A)^{-1}$
P	$\mathbb{R}^{q\times q}$	SDP matrix defining an inner product in $\mathbb{R}^q$; $P = (B^{\mathrm{T}}B)^{-1}$
Q	$\mathbb{R}^{q\times q}$	$Q = P^{1/2}$
R	$\mathbb{R}^{p\times q}$	for calculation; $R = MTQ$
T	$\mathbb{R}^{p\times q}$	for calculation; $T = A^{\mathrm{T}}B$
v_{A}	$\mathbb{R}^p$	a vector in $\mathbb{R}^p$
v_{B}	$\mathbb{R}^q$	a vector in $\mathbb{R}^q$
V_{A}	$\mathbb{R}^{p\times q}$	matrix with axis v_{A} columnwise
V_{B}	$\mathbb{R}^{q\times q}$	matrix with axis v_{B} columnwise
y_{A}	$\mathbb{R}^n$	a vector in span A
y_{B}	$\mathbb{R}^n$	a vector in span B
Y_{A}	$\mathbb{R}^{n\times q}$	matrix with components y_{A} columnwise
Y_{B}	$\mathbb{R}^{n\times q}$	matrix with components y_{B} columnwise

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. In this section, the problem is stated (section 8.1) and solved (section 8.2) first in an algebraic way, and a second effort must be done to implement involved linear algebra calculations with a reasonable costs (section 8.3).

8.1 Stating the problem

We will denote by v_{A} a vector in $\mathbb{R}^p$ and by v_{B} in $\mathbb{R}^q$. A vector $y_{\mathrm{A}} \in \operatorname{span} A$ (resp. $y_{\mathrm{B}} \in \operatorname{span} B$) can be written as Av_{A} (resp. Bv_{B}). The correlation between y_{A} and y_{B} is

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

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	$v_{\mathrm{A}} \in \mathbb{R}^p$, $v_{\mathrm{B}} \in \mathbb{R}^q$
such that	$\dfrac{\langle Av_{\mathrm{A}}, Bv_{\mathrm{B}}\rangle}{\|Av_{\mathrm{A}}\|\|Bv_{\mathrm{B}}\|}$ is maximal

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

Given	$A \in \mathbb{R}^{n\times p}$, $B \in \mathbb{R}^{n\times q}$
Find	$v_{\text{A}} \in \mathbb{R}^p$, $v_{\text{B}} \in \mathbb{R}^q$
with	$\|Av_{\text{A}}\|^2 = 1$, $\|Bv_{\text{B}}\|^2 = 1$
such that	$\langle Av_{\text{A}}, Bv_{\text{B}}\rangle$ 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

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

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

• Here, the unknowns are $(v_{\text{A}}, v_{\text{B}})$, the function f is $f(v_{\text{A}}, v_{\text{B}}) = \langle Av_{\text{A}}, Bv_{\text{B}}\rangle$ and g is $\|Av_{\text{A}}\|^2 = \|Bv_{\text{B}}\|^2 = 1$. One computes separately the partial derivatives with respect to v_{A} and v_{B}, denoting them $\nabla_{v_{\text{A}}}$ and $\nabla_{v_{\text{B}}}$. One has

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

Then, the solution satisfies to

$$\left\{\begin{array}{lcl} \nabla_{v_{\text{A}}}: & A^{\text{T}}Bv_{\text{B}} & = \lambda\, A^{\text{T}}Av_{\text{A}} \\ \nabla_{v_{\text{B}}}: & B^{\text{T}}Av_{\text{A}} & = \mu\, B^{\text{T}}Bv_{\text{B}} \end{array}\right. \tag{8.2.2}$$

• We first show that $\lambda = \mu$.

Proof. Therefore, we observe that

$$\left\{\begin{array}{lcl} \langle A^{\text{T}}Bv_{\text{B}}, v_{\text{A}}\rangle & = & \lambda\langle A^{\text{T}}Av_{\text{A}}, v_{\text{A}}\rangle \\ \langle B^{\text{T}}Av_{\text{A}}, v_{\text{B}}\rangle & = & \mu\langle B^{\text{T}}B, v_{\text{B}}\rangle \end{array}\right.$$

and that

$$\begin{array}{rcccl} \langle A^{\mathrm{T}}Av_{\mathrm{A}}, v_{\mathrm{A}}\rangle & = & \langle Av_{\mathrm{A}}, Av_{\mathrm{A}}\rangle & = & 1 \\ \langle B^{\mathrm{T}}Bv_{\mathrm{B}}, v_{\mathrm{B}}\rangle & = & \langle Bv_{\mathrm{B}}, Bv_{\mathrm{B}}\rangle & = & 1 \end{array}$$

Then

$$\lambda = \langle A^{\mathrm{T}}Bv_{\mathrm{B}}, v_{\mathrm{A}}\rangle = \langle B^{\mathrm{T}}Av_{\mathrm{A}}, v_{\mathrm{B}}\rangle = \mu$$

□

- Thus, equation (8.2.2) reads

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

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

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

and, having in mind that $Av_{\mathrm{A}} = y_{\mathrm{A}}$ and $Bv_{\mathrm{B}} = y_{\mathrm{B}}$

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

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

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

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 Computing the solution

This solution is geometrically speaking very intuitive, but it does not lead to the most efficient way to compute a solution. We start from

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

Let us recall that

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

Then

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

We show here how it is possible to avoid the complexity of computing $\mathcal{P}_{\text{A}} = A(A^{\text{T}}A)^{-1}A^{\text{T}}$ and $\mathcal{P}_{\text{B}} = B(B^{\text{T}}B)^{-1}B^{\text{T}}$. Indeed, computing

$A^{\text{T}}A$	is in	$\mathcal{O}(np^2)$	$B^{\text{T}}B$	is in	$\mathcal{O}(nq^2)$
$(A^{\text{T}}A)^{-1}$		$\mathcal{O}(p^3)$	$(B^{\text{T}}B)^{-1}$		$\mathcal{O}(q^3)$
$A(A^{\text{T}}A)^{-1}A^{\text{T}}$		$\mathcal{O}(np^2)$	$B(B^{\text{T}}B)^{-1}B^{\text{T}}$		$\mathcal{O}(nq^2)$

Hence, computing $\mathcal{P}_{\text{A}}$ (resp. $\mathcal{P}_{\text{B}}$) is in $\mathcal{O}(np^2)$ (resp. $\mathcal{O}(nq^2)$). We have $\mathcal{P}_{\text{A}}, \mathcal{P}_{\text{B}} \in \mathbb{R}^{n\times n}$, so $\mathcal{P}_{\text{A}}\mathcal{P}_{\text{B}}, \mathcal{P}_{\text{B}}\mathcal{P}_{\text{A}} \sim \mathcal{O}(n^3)$. However, $\operatorname{rank}\mathcal{P}_{\text{A}} = m$ and $\operatorname{rank}\mathcal{P}_{\text{B}} = q$. So, it is possible to reduce the complexity of the calculation.

- One possibility is to run a SVD of A and B. If the matrices of left singular vectors are denoted respectively U_{A} and U_{B}, the same calculation holds for the projection on the space spanned by the columns of U_{A} (resp. U_{B}) as they are the same as those spanned by A (resp. B). We have $U_{\text{A}}^{\text{T}}U_{\text{A}} = \mathbb{I}_p$, and $U_{\text{B}}^{\text{T}}U_{\text{B}} = \mathbb{I}_q$. Then $\mathcal{P}(U_{\text{A}}) = U_{\text{A}}U_{\text{A}}^{\text{T}}$ (resp. $\mathcal{P}(U_{\text{B}}) = U_{\text{B}}U_{\text{B}}^{\text{T}}$) and the complexity of the calculation of the projectors is reduced. However, we still have $\mathcal{P}(U_{\text{A}}), \mathcal{P}(U_{\text{B}}) \in \mathbb{R}^{n\times n}$, and the calculations of $\mathcal{P}_{\text{A}}\mathcal{P}_{\text{B}}$ and $\mathcal{P}_{\text{B}}\mathcal{P}_{\text{A}}$ still are in $\mathcal{O}(n^3)$. Let us try another way.

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

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

Then

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

Let us recall that

$$y_{\text{A}} = Av_{\text{A}}, \qquad y_{\text{B}} = Bv_{\text{B}}$$

Then

$$\left\{\begin{array}{rcl} ANTPT^{\text{T}}v_{\text{A}} & = & \lambda^2 Av_{\text{A}} \\ BPT^{\text{T}}NTv_{\text{B}} & = & \lambda^2 Bv_{\text{B}} \end{array}\right.$$

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

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

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,

$$\begin{array}{rcl} NTPT^{\mathrm{T}} & = & M^2TQ^2T^{\mathrm{T}} \\ & = & M(MTQ)(QT^{\mathrm{T}}M)M^{-1} \\ & = & MRR^{\mathrm{T}}M^{-1} \end{array}$$

and (the same for $PT^{\mathrm{T}}NT$) equation (8.3.3) reads

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

Let us denote

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

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

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

where we recognize the `PCA` of $R = MTQ$. The complexity of this calculation is:

$$\begin{array}{lll} T & = A^{\mathrm{T}}B & \text{is in } \mathcal{O}(npq) \\ N & = (A^{\mathrm{T}}A)^{-1} & O(np^2) \\ P & = (B^{\mathrm{T}}B)^{-1} & O(nq^2) \\ M & = N^{1/2} & \mathcal{O}(p^3) \\ Q & = P^{1/2} & \mathcal{O}(q^3) \\ R & = MTQ & \mathcal{O}(p^2q) \end{array}$$

• The `PCA` of R 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

$$w_{\mathrm{A}} = Rw_{\mathrm{B}}, \qquad \text{with} \quad \left\{\begin{array}{ll} R & \in \mathbb{R}^{p\times q} \\ w_{\mathrm{A}} & \in \mathbb{R}^p \\ w_{\mathrm{B}} & \in \mathbb{R}^q \end{array}\right. \tag{8.3.6}$$

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 $N = (A^{\text{T}}A)^{-1}$ and $P = (B^{\text{T}}B)^{-1}$. We then have

$$v_{\text{A}} = Mw_{\text{A}}, \qquad v_{\text{B}} = Qw_{\text{B}}$$

and

$$y_{\text{A}} = Av_{\text{A}}, \qquad y_{\text{B}} = Bv_{\text{B}}$$

• **Algorithm:** There are several ways to write an algorithm for this calculation. Here is a direct one, without calling PCA or PCA-MET.

Algorithm 11 Canonical Correlation Analysis: $\text{CCA}(A, B)$

input $A \in \mathbb{R}^{n\times p}$, $B \in \mathbb{R}^{n\times q}$ with $p \geq q$
compute $T = A^{\text{T}}B$
compute $N = (A^{\text{T}}A)^{-1}$
compute $P = (B^{\text{T}}B)^{-1}$
compute $M = N^{1/2}$
compute $Q = P^{1/2}$
compute $R = MTQ$
compute $W_{\text{A}}, \Lambda, W_{\text{B}} = \text{SVD}(R)$
compute $V_{\text{A}} = MW_{\text{A}}$
compute $V_{\text{B}} = QW_{\text{B}}$
compute $Y_{\text{A}} = AV_{\text{A}}$
compute $Y_{\text{B}} = BV_{\text{B}}$
return $Y_{\text{A}}, Y_{\text{B}}, V_{\text{A}}, V_{\text{B}}, \Lambda$

Comments: Here are some comments on the algorithm:

→ The components y_{A} are the columns of Y_{A}

→ The components y_{B} are the columns of Y_{B}

→ the axis v_{A} are the columns of V_{A}

→ the axis v_{B} are the columns of V_{B}

→ the singular values of R are the correlation coefficients λ, because $R^{\text{T}}Rw_{\text{A}} = \lambda^2 w_{\text{A}}$, and the eigenvalues of $R^{\text{T}}R$ are the square of the singular values of R

→ it is possible to compute $M = N^{1/2}$ through a SVD of N: if $N = U_{\text{M}}\Sigma_{\text{M}}V_{\text{M}}$, then $M = U_{\text{M}}\Sigma_{\text{M}}^{1/2}V_{\text{M}}$ and Σ_{M} is diagonal. The same for computing Q from P.

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, Sap90] from which it has been borrowed and adapted here. 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 $y_{\text{A}} = Au_{\text{A}}$, $y_{\text{B}} = Bu_{\text{B}}$ and $y_{\text{C}} = Cu_c$ with $\|y_{\text{A}}\| = \|y_{\text{B}}\| = \|y_{\text{C}}\| = 1$. One can define three-ways `CCA` as finding a triplet $(y_{\text{A}}, y_{\text{B}}, y_{\text{C}})$ such that their correlation is maximal. There is however no canonical way to define the correlation between 3 vectors. It can be defined from $\|y_{\text{A}} \wedge y_{\text{B}} \wedge y_{\text{C}}\|$ (vectors are independent when their wedge product is maximal), or $\|y_{\text{A}} + y_{\text{B}} + y_{\text{C}}\|$, 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

$$w = M^{\text{T}} M w, \qquad v = D_{\text{B}}^{1/2} w$$

with

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

Key observations: Let us note three things:

→ $(A^{\text{T}} A)^{-1} \in \mathbb{R}^{p \times p}$ (resp. $(B^{\text{T}} B)^{-1} \in \mathbb{R}^{q \times q}$) 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 $D_{\text{A}}^{1/2}$ in $\mathbb{R}^p$ and by $D_{\text{B}}^{1/2}$ 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

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

and

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

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

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

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 $R^{\text{T}}R$. We will denote the matrix dimensions under each block. We have

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

So

$$R^{\text{T}}R = \begin{bmatrix} \underset{p\times p}{D_{\text{A}}^{1/2}A^{\text{T}}AD_{\text{A}}^{1/2}} & \underset{p\times q}{D_{\text{A}}^{1/2}A^{\text{T}}BD_{\text{B}}^{1/2}} \\ \\ \underset{q\times p}{D_{\text{B}}^{1/2}B^{\text{T}}AD_{\text{A}}^{1/2}} & \underset{q\times q}{D_{\text{B}}^{1/2}B^{\text{T}}BD_{\text{B}}^{1/2}} \end{bmatrix}$$

Let us observe that

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

Hence

$$R^{\text{T}}R = \begin{bmatrix} \mathbb{I}_p & \underset{p\times q}{D_{\text{A}}^{1/2}A^{\text{T}}BD_{\text{B}}^{1/2}} \\ \\ \underset{q\times p}{D_{\text{B}}^{1/2}B^{\text{T}}AD_{\text{A}}^{1/2}} & \mathbb{I}_q \end{bmatrix}$$

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 analysis of two quantitative 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 $\texttt{Can}(A, B)$, 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, LMP00]. Since then, the diversity of related methods have flourished, and a more recent panorama is given in [GB06]. According to the introduction of [GB06], 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 they point out several independent beginnings in the 30's and early 40's) and, long before 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

$$D = [d_{ij}]_{i,j} \in \mathbb{R}^{n\times n} \qquad \text{with} \quad 1 \leq i, j \leq n,$$

such that

$$d_{ij} = d(i, j).$$

MDS at dimension r addresses the question of 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 can 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 12 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 of G with eigenvalues Λ
4: **compute** the coordinates $X \in \mathbb{R}^{n\times n}$ of point cloud $\mathcal{X}$ from the EVD of G
5: **compute** the best low rank approximation $X_r \in \mathbb{R}^{n\times r}$ of X
6: **return** X_r, Σ_r

and developed further in this section.

Notes and references: There exists many excellent textbooks presenting MDS, classical MDS or LSS. 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

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

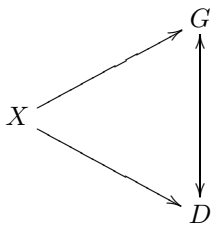
Then

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

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.

• The objective is to associate to (M, d) a point cloud $\mathcal{X}$ in a Euclidean space such that, if possible, $d(i, j) = \|x_i - x_j\|$. 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 $(u_\alpha, \lambda_\alpha)_\alpha$ be the set of eigenpairs of G

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

with

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$$

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

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

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

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

We recognize here the SVD of G if G is positive. The case where G is not positive is handled by designing a quadratic embedding, and is addressed in section 10.5. Let us note that the standard practice when G is not positive is not to design a quadratic embedding, but to clip to zero the non positive eigenvalues and eigenvectors, i.e. to keep track of positive eigenvalues only and associated eigenvectors. If G is positive, the eigenvalues of G are its singular values, and if G is not positive, the negative eigenvalues are singular values up to their sign (if $\lambda < 0$ is an eigenvalue of G, $-\lambda > 0$ is a singular value of G).

- As G is definite positive, let

$$\Sigma = \Lambda^{1/2}$$

Then

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

and we can select

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

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 13 MDS with eigendecomposition of Gram matrix

1: **input:** Gram matrix G
2: **compute** eigendecomposition of G: $GU = U\Lambda$, with eigenvectors u_α (columns of U) and eigenvalues λ_α: $Gu_\alpha = \lambda_\alpha u_\alpha$
3: **compute** $\Sigma = \Lambda^{1/2}$
4: **compute** $X = U\Sigma$
5: **return** X, Λ

One notices that for X to be computed this way, one must have $\Lambda \geq 0$ for $\Lambda^{1/2}$ to be real. It is one condition for an isometry between (M, d) and a Euclidean space to exist.

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 $X = U\Sigma^{1/2}V^{\mathrm{T}}$, with singular values being the diagonal of $\Sigma^{1/2}$ and $V = \mathbb{I}_n$. Hence, $X = U\Sigma^{1/2}$ is the matrix of principal components of X, and dimensionality reduction of X by PCA is given as a cherry on top.

10.4 `MDS` algorithm

Wrapping all this together yields the following algorithm for `MDS`

Algorithm 14 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: $G = \text{GRAM}(D)$
3: **compute** the eigenpairs $(u_\alpha, \lambda_\alpha)$ such that $Gu_\alpha = \lambda_\alpha u_\alpha$
4: keep in U the columns associated to non-negative eigenvalues of G; clip them off in Λ
5: **compute** $\Sigma = \Lambda^{1/2}$
6: **compute** $X = U\Sigma$
7: keep in X_r the r first columns of X only, and in Λ_r the first r non negative eigenvalues of G
8: **return** X_r, Λ_r

- **Note:** 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 values in Λ). If the SVD is selected, we have $G = U\Sigma V^{\mathrm{T}}$ (classical notation $(U, \Sigma, V$, here $\Sigma \neq \Lambda^{1/2})$, with columns of V (resp. singular values) being the same as the columns of U (resp. eigenvalues of G) up to the sign, depending on the sign of the corresponding eigenvalue of G:

$$\left| \begin{array}{lcll} \lambda_\alpha > 0 & \Longrightarrow & v_\alpha = u_\alpha & \sigma_\alpha = \lambda_\alpha \\ \lambda_\alpha < 0 & \Longrightarrow & v_\alpha = -u_\alpha & \sigma_\alpha = -\lambda_\alpha \end{array} \right.$$

10.5 Quadratic embedding

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 recipe in equation (10.1.1) is positive, i.e. that all eigenvalues of G (the λ_α) are non negative. This is a property of a Gram matrix, but is not always the case for the matrix G computed with real data. If one eigenvalue at least is negative, the matrix G is more strictly called a *kernel matrix*, and 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, m) (see appendix B for a short introduction to quadratic forms and spaces). 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. We show in this section how to take into account the axis associated to negative eigenvalues of the kernel matrix.

A simple example of a discrete metric space without isometry in a Euclidean space: A discrete metric space (M, d) being given, there exists not systematically an isometry (metric embedding) into an Euclidean space preserving the distances. A simple example is

$$M = \{i, j, k, c\}$$

which is the set of vertices of a graph with $E = \{(i, c), (j, c), (k, c)\}$ (there is no conflict of notations between E which is here classically the set of edges and E which is in this section a vector space) with the distance between two vertices being the length of the shortest path

between them. We then have

$$D = \begin{array}{c|cccc} & i & j & k & c \\ \hline i & 0 & 2 & 2 & 1 \\ j & 2 & 0 & 2 & 1 \\ k & 2 & 2 & 0 & 1 \\ c & 1 & 1 & 1 & 0 \end{array}$$

There exists however a map into a pseudo-euclidean space (E, q) as host space, preserving a quadratic form.

Quadratic embedding: Let (M, d) be a discrete metric space, with $M = \{1, \ldots, i, \ldots, n\}$. Let (E, q) be a quadratic space, with $\dim E = n$, and φ a map

$$M \xrightarrow{\varphi} E$$

Let us denote

$$a_i = \varphi(i)$$

So, $a_i \in E$, and we have a cloud of n points in E, each one corresponding to an element in M. Then, φ is a quadratic embedding, or preserves the distances, if

$$q(a_i - a_j) = d^2(i, j) \tag{10.5.1}$$

We show next

- ◇ that such an embedding exists (and is not unique),
- ◇ how to build it by specifying q and φ.

We first build q, and next φ.

Building a quadratic form: The key is that the kernel matrix built with recipe in equation (10.1.1) is the polar form of q, which permits to build q knowing d. Let us denote $G = (g_{ij})_{i,j}$ with $1 \leq i, j \leq n$, and

$$G(a_i, a_j) = g_{ij}.$$

Then, G can be computed from the distances $d(i, j)$ by

$$g_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \Delta_i - \Delta_j + \Delta\right) \tag{10.5.2}$$

with

$$\Delta_i = \frac{1}{n}\sum_k d_{ik}^2, \qquad \Delta_j = \frac{1}{n}\sum_k d_{kj}^2, \qquad \Delta = \frac{1}{n^2}\sum_{k,\ell} d_{k\ell}^2$$

G is a symmetric bilinear form. Let us assume for sake of simplicity that it is definite (the extension to the case where it is non definite, i.e. is not full rank, is straightforward). The quadratic form is given by

$$q(x) = B(x, x), \qquad \forall\, x \in E$$

or, if $x = \sum_i x_i\, a_i$,

$$q(x) = \sum_{i,j=1}^{n} g_{ij}\, x_i x_j.$$

Now that we have the quadratic form, we need to construct the embedding φ.

Building the quadratic embedding: Let us now denote by (ω_k, z_k) the eigenpairs of G, with $\omega_k \in \mathbb{R}$ and $z_k \in E$, i.e. pairs (ω_k, z_k) with

$$G z_k = \omega_k \, z_k. \tag{10.5.3}$$

or

$$G \, Z = Z \, \Omega$$

(Z is the $n \times n$ matrix with eigenvectors as columns, and Ω is the $n \times n$ diagonal matrix with ω_k on the diagonal). We have

$$G = Z \, \Omega \, Z^{\mathrm{T}} \tag{10.5.4}$$

(indeed, if $G' = Z\Omega Z^{\mathrm{T}}$, we have $G'Z = Z\Omega = GZ$ as $Z^{\mathrm{T}}Z = \mathbb{I}_n$, and as Z is invertible, $G' = G$). Let us separate the positive from the negative eigenvalues of G, i.e. denote

$$\omega_1 \geq \ldots \geq \omega_p > 0 > \omega'_1 \geq \ldots \geq \omega'_m \tag{10.5.5}$$

(we assume for sake of simplicity that 0 is not an eigenvalue of G) or

$$\begin{cases} \omega_k > 0 & \text{if} \quad k \leq p \\ \omega_k < 0 & \text{if} \quad k > p \end{cases} \qquad \text{with} \qquad \omega'_j = \omega_{p+j}$$

Let us denote

$$\omega_k = \sigma_k^2, \qquad \omega'_j = -\theta_j^2 \tag{10.5.6}$$

and

$$\begin{cases} G \, u_k & = \sigma_k^2 \, u_k \\ G \, v_j & = -\theta_j^2 \, v_j \end{cases}$$

Let us use blockwise notations

$$Z = [U, V], \qquad \Omega = \begin{pmatrix} \Sigma^2 & 0 \\ 0 & -\Theta^2 \end{pmatrix} \tag{10.5.7}$$

Then

$$\begin{cases} G \, U & = U \, \Sigma^2 \\ G \, V & = -V \, \Theta^2 \end{cases}$$

Rewriting equation (10.5.3) $G = Z\Omega Z^{\mathrm{T}}$ with this blockwise decomposition leads to

$$G = U \, \Sigma^2 \, U^{\mathrm{T}} - V \, \Theta^2 \, V^{\mathrm{T}} \tag{10.5.8}$$

illustrated by

$$G = \begin{array}{|c|c|} \hline U & V \\ \hline \end{array} \quad \begin{array}{|c|c|} \hline \Sigma^2 & 0 \\ \hline 0 & -\Theta^2 \\ \hline \end{array} \quad \begin{array}{|c|} \hline U^{\mathrm{T}} \\ \hline V^{\mathrm{T}} \\ \hline \end{array}$$

$$= \begin{array}{|c|c|} \hline U\Sigma^2 & -V\Theta^2 \\ \hline \end{array} \quad \begin{array}{|c|} \hline U \, \Sigma^2 \, U^{\mathrm{T}} \\ -V \, \Theta^2 \, V^{\mathrm{T}} \\ \hline \end{array}$$

Let us denote

$$\begin{cases} X &= U\,\Sigma \\ Y &= V\,\Theta \end{cases} \tag{10.5.9}$$

Then

$$G = XX^{\mathrm{T}} - YY^{\mathrm{T}} \tag{10.5.10}$$

One recovers in X the components obtained by clipping to 0 the eigenvalues ω'_j, and complement this classical result with a second point cloud associated with the negative eigenvalues of the Gram matrix.

Algorithm: This leads to the following algorithm (for sake of clarity, we denote here by λ the positive eigenvalues of G and by ψ the negative ones, what was denoted ω and ω' in the text.

Algorithm 15 Quadratic embedding of a discrete metric space

1: **input** $D \in \mathbb{R}^{n\times n}$: $D[i,j] = d_{ij}$
2: **compute** $\Delta_i = \frac{1}{n}\sum_k d^2_{ik}$, $\Delta_j = \frac{1}{n}\sum_k d^2_{kj}$, $\Delta = \frac{1}{n^2}\sum_{i,j} d^2_{ij}$
3: **compute** $G \in \mathbb{R}^{n\times n}$: $G[i,j] = -\frac{1}{2}\left(d^2_{ij} - \Delta_i - \Delta_j + \Delta\right)$
4: **compute** (ω_k, z_k) : $Gz_k = \omega_k z_k$
5: **denote** Λ, Ψ, diagonal matrices of eigenvalues $\lambda > 0$ and $\psi < 0$
6: **compute** $\Sigma = \Lambda^{1/2}$, $\Theta = (-\Psi)^{1/2}$
7: **denote** U, V, eigenvectors of G with $\lambda > 0$, and $\psi < 0$
8: **compute** $X = U\Sigma$, $Y = V\Theta$
9: **return** X, Y, Σ, Θ

Summary and quality of the low rank approximation: We have a metric space (M, d) with $|M| = n$. There exists a quadratic space (E, q) with $\dim E = n$ and a map

$$M \xrightarrow{\ \varphi\ } E$$

with $a_i = \varphi(i)$ such that

$$d^2(i,j) = q(a_i - a_j).$$

Let G be the polar form of q, i.e.

$$g_{ij} = G(a_i, a_j) = \frac{1}{2}(q(a_i + a_j) - q(a_i) - q(a_j)),$$

It is called the kernel matrix of q, and can be computed knowing the distances by

$$g_{ij} = -\frac{1}{2}\left(d^2_{ij} - \Delta_i - \Delta_j + \Delta\right).$$

with

$$\Delta_i = \frac{1}{n}\sum_j d^2_{ij}, \qquad \Delta_j = \frac{1}{n}\sum_i d^2_{ij}, \qquad \Delta = \frac{1}{n^2}\sum_{i,j} d^2_{ij}$$

The quadratic form q is defined by its polar form: $q(x) = G(x,x)$. The signature of q is (p, m) with p being the number of positive eigenvalues of G and m the number of negative eigenvalues.

We denote by E_+ (resp. E_-) the subspace of E spanned by the eigenvectors of G associated to a positive (resp. negative) eigenvalue of G. Then, one can write

$$E = E_+ \oplus E_-$$

and, for any point $a_i \in E$,

$$a_i = x_i \oplus y_i, \qquad \text{with} \quad \begin{cases} x_i & \in E_+ \\ y_i & \in E_-. \end{cases}$$

Two point clouds X and Y are built with algorithm 15, with X being made of p points in $\mathbb{R}^p$ and Y of m points in $\mathbb{R}^n$ such that

$$G = XX^{\mathrm{T}} - YY^{\mathrm{T}}.$$

As G is symmetric, its eigenvectors are orthogonal. The columns of X (resp. Y) are the eigenvectors associated with positive (resp. negative) eigenvalues of G. Then $X^{\mathrm{T}}Y = \mathbf{0}$. Hence

$$\begin{aligned} \|G\|^2 &= \|XX^{\mathrm{T}} - YY^{\mathrm{T}}\|^2 \\ &= \|XX^{\mathrm{T}}\|^2 + \|YY^{\mathrm{T}}\|^2 - 2\langle XX^{\mathrm{T}}\,,\, YY^{\mathrm{T}}\rangle \\ &= \|XX^{\mathrm{T}}\|^2 + \|YY^{\mathrm{T}}\|^2 \end{aligned} \tag{10.5.11}$$

because

$$\begin{aligned} \langle XX^{\mathrm{T}}\,,\, YY^{\mathrm{T}}\rangle &= \mathrm{Tr}\,\{XX^{\mathrm{T}}(YY^{\mathrm{T}})^{\mathrm{T}}\} \\ &= \mathrm{Tr}\,\{XX^{\mathrm{T}}YY^{\mathrm{T}}\} \\ &= \mathrm{Tr}\,\{X\,(X^{\mathrm{T}}Y)\,Y^{\mathrm{T}}\} \\ &= 0 \end{aligned} \tag{10.5.12}$$

If negative eigenvalues are clipped to 0, one has $G \approx XX^{\mathrm{T}}$, and the quality of representation of G by XX^{T} is $\|XX^t\|/\|G\|$. Similarly, knowing that

$$g_{ij} = \langle x_i\,,\, x_j\rangle - \langle y_i\,,\, y_j\rangle,$$

(this is another formulation of $G = XX^{\mathrm{T}} - YY^{\mathrm{T}}$), one has,

$$\begin{aligned} d^2(i,j) &= q(a_i - a_j) \\ &= G(a_i - a_j\,,\, a_i - a_j) \\ &= G(a_i, a_i) + G(a_j, a_j) - 2G(a_i, a_j) \\ &= g_{ii} + g_{jj} - 2\,g_{ij} \\ &= \|x_i\|^2 - \|y_i\|^2 + \|x_j\|^2 - \|y_j\|^2 - 2\langle x_i\,,\, x_j\rangle + 2\langle y_i\,,\, y_j\rangle \\ &= \|x_i - x_j\|^2 - \|y_i - y_j\|^2. \end{aligned} \tag{10.5.13}$$

So, $\|y_i - y_j\|^2$ quantifies the discrepancy between $d^2(i,j)$ and $\|x_i - x_j\|^2$ in classical MDS (where negative eigenvalues and eigenectors are clipped to 0) with full rank. This shows as well that distances in the point cloud built by classical MDS with clipping negative eigenvalues to 0 are overestimated as $\|x_i - x_j\|^2 = d_{ij}^2 + \|y_i - y_j\|^2 \geq d_{ij}^2$.

Notes and references: See appendix B for a short presentation of quadratic forms and spaces. The classification of quadratic forms when $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ depends on the matrix of the polar form and is well understood and given for example in [dSP10]. The classification on a arbitrary field is said to be immensely difficult [Ber87]. Embedding from a metric space into a quadratic space have been studied e.g. in [Gol84, Gow85] and is presented in [PD05, section 3.5] where it is called *pseudo-euclidean embedding*.

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 (simplified)**
`PCA`	$(Y, V, \Lambda) = \text{PCA\_CORE}(A)$	$C = A^{\mathrm{T}} A$ $(\Lambda, V) = \text{EIG}(C)$ $Y = AV$ or $(U, \Sigma, V) = \text{SVD}(A)$ $\Lambda = \Sigma^2$ $Y = U\Sigma$
`PCAiv`	$(Y, V, \Lambda) = \text{PCA-IV}(A, U, V)$	$\mathcal{P}_E = U(U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}$ $\mathcal{P}_F = V(V^{\mathrm{T}}V)^{-1}V^{\mathrm{T}}$ $A_{\mathrm{U,V}} = \mathcal{P}_{\mathrm{E}} A \mathcal{P}_{\mathrm{F}}$ $(Y, \Lambda, V) = \text{PCA\_CORE}(A_{\mathrm{U,V}})$
`PCAmet`	$(Y, V, \Lambda) = \text{PCA-MET}(A, M, Q)$	$A_{\mathrm{M,Q}} = MAQ$ $(Z, \Lambda, X) = \text{PCA\_CORE}(A_{\mathrm{M,Q}})$ $Y = M^{-1}Z$ $V = Q^{-1}X$
`CoA`	$(Y_r, Y_c, \Lambda) =$ `CoA` (T)	$A = T/T_{++}, \quad T_{++} = \sum_{i,j} T_{ij}$ $r_i = \sum_j \alpha_{ij}, \quad c_j = \sum_i \alpha_{ij}$ $M = \text{diag}\,(1/\sqrt{r_i}), \quad Q = \text{diag}\,(1/\sqrt{c_j})$ $A_{\mathrm{M,Q}} = M(A - r \otimes c)Q = \frac{\alpha_{ij} - r_i c_j}{\sqrt{r_i c_j}}$ $Z, \Lambda, X = \text{PCA\_CORE}(A_{\mathrm{M,Q}})$ $Y_r = M^{-1}Z, \quad Y_c = Q^{-1}X$
`CCA`	$(Y_{\mathrm{A}}, Y_{\mathrm{B}}, U_{\mathrm{A}}, U_{\mathrm{B}}, \Lambda) = \text{CCA}(A, B)$	$T = A^{\mathrm{T}}B$ $M = (A^{\mathrm{T}}A)^{-1/2}, \quad Q = (B^{\mathrm{T}}B)^{-1/2}$ $R = MTQ$ $W_{\mathrm{A}}, \Psi, W_{\mathrm{B}} = \text{PCA\_CORE}(R)$ $\Lambda = \sqrt{\Psi}$ $U_{\mathrm{A}} = MW_{\mathrm{A}}, \quad U_{\mathrm{B}} = QW_{\mathrm{B}}$ $Y_{\mathrm{A}} = AU_{\mathrm{A}}, \quad Y_{\mathrm{B}} = BU_{\mathrm{B}}$
`MDS`	$Y, \Lambda = \text{MDS}(D, r)$	$G = \text{GRAM}(D)$ $(U, \Sigma) = \text{SVD}(G)$: $G = U\Sigma U^{\mathrm{T}}$ $Y = U\Sigma^{1/2}$

12 References in textbooks

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

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
FA	Factor Analysis
GRP	Gaussian Random Projection
IV	Instrumental Variables
MCoA	Multiple Correspondence Analysis
MDA	Multivariate Data Analysis
MDS	Mulidimensional Scaling
PCA	Princial Component Analysis
PPCA	Probabilistic PCA
RP	Random Projection
rSVD	Randomized SVD
SDP	Symmetric Definite Positive
SGF	Symmetric Gauge Function
SVD	Singular Value Decomposition
SVM	Support Vector Machine
UIN	Unitarily Invariant Norm

A Preliminaries in linear algebra

Let us recall here some basic facts in linear algebra. Linear algebra can have an abstract setting, with the notions of vector spaces and linear maps, or numerical setting, working with arrays (vectors and matrices). A matrix A is the expression of a linear map $E \longrightarrow F$ by an array once basis have been selected in E and F.

A.1 Vector space and linear map

$\diamond$ Let $\mathbb{K}$ be a field, usually $\mathbb{R}$ or $\mathbb{C}$, and here $\mathbb{R}$ unless otherwise stated. A vector space E over $\mathbb{K}$ is a set endowed with two operations:

$\rightarrow$ an addition, $+$, such that $(E, +)$ is an Abelian group

$\rightarrow$ a multiplication by a scalar

$$\begin{aligned} \mathbb{K} \times E &\longrightarrow E \\ (\alpha, u) &\longrightarrow \alpha u. \end{aligned}$$

Multiplication by a scalar verifies the following properties:

$$\left|\begin{array}{rcl} \alpha(u+v) & = & \alpha u + \alpha v \\ (\alpha+\beta)u & = & \alpha u + \beta u \\ \alpha(\beta u) & = & (\alpha\beta)u \\ 1u & = & u. \end{array}\right.$$

$\diamond$ A vector subspace of E is a subset of E which is closed under the addition or the multiplication by a scalar. It is itself a vector space on $\mathbb{K}$.

$\diamond$ A basis of a vector space E is a collection $(u_k)_k$ of vectors $u_k \in E$ such that any vector $u \in E$ can be written as

$$u = \sum_k \alpha_k u_k,$$

in a unique manner where $\alpha_k \in \mathbb{K}$. If a basis exists, all basis have the same cardinality. This cardinality is the dimension of the vector space. Usually, the dimension is either an integer $n \in \mathbb{N}$ or $\aleph_0$, the cardinality of $\mathbb{N}$. The vector spaces we will work with in these notes are finite dimensional.

$\diamond$ Let E, F be two vector spaces on the same field $\mathbb{K}$. A linear map

$$E \xrightarrow{\ L\ } F,$$

from E on F is a map which preserves the linear structure, i.e.

$$\begin{array}{rcl} L(u+v) & = & Lu + Lv \\ L(\alpha u) & = & \alpha\, Lu \end{array}$$

(it is customary for linear maps to write Lu instead of $L(u)$ when there is no ambiguity). The image of a linear subspace $E' \subset E$ of E is a vector subspace $F' \subset F$.

$\diamond$ Let $n = \dim E \in \mathbb{N}$ and $m = \dim F \in \mathbb{N}$. Let $(u_j)_j$ with $1 \leq j \leq n$ be a basis of E and $(v_i)_i$ with $1 \leq i \leq m$ a basis in F. Let

$$Lu_j = \sum_i \alpha_{ij} v_i.$$

Then, the matrix $(\alpha_{ij})_{i,j} \in \mathbb{R}^{m\times n}$ is the matrix of the linear map L in basis $(u_j)_j$ for E and $(v_i)_i$ for F. We have, for any vector $u = \sum_j \beta_j u_j \in E$

$$\begin{array}{rcl} Lu & = & L\left(\displaystyle\sum_j \beta_j u_j\right) \\ & = & \displaystyle\sum_j \beta_j Lu_j \\ & = & \displaystyle\sum_j \beta_j \left(\sum_i \alpha_{ij} v_i\right) \\ & = & \displaystyle\sum_i \left(\sum_j \alpha_{ij}\beta_j\right) v_i. \end{array} \tag{A.1.1}$$

Hence, the $j-$th column of A is the vector $Lu_j \in \mathbb{R}^m$ with coordinates $(\sum_{\ell=1}^n \alpha_{i\ell}\beta_\ell)_{1\leq i\leq m}$.

$\diamond$ The kernel of L is the set of vectors $u \in E$ the image of which is $0 \in F$

$$\ker L = \{u \in E \mid Lu = 0\}. \tag{A.1.2}$$

It is a vector subspace of E.

$\diamond$ The image of L is the set of vectors $v \in F$ which have a preimage in E

$$\text{im } L = \{v \in F \mid \exists\, u \in E \quad \text{s.t.} \quad v = Lu\}. \tag{A.1.3}$$

It is a vector subspace of F.

$\diamond$ The rank nullity theorem states that

$$\dim \ker L + \dim \text{im } L = \dim E \tag{A.1.4}$$

$\diamond$ A linear map from E to E is called an endomorphism. If it is an isomorphism, it is called an automorphism. The set of all endomorphisms of E, denoted $\text{end } E$, is an associative algebra for the operation $+$ and multiplication by a scalar for the vector space structure, and the composition $\circ$ for it to be an algebra.

A.2 Eigenspace, eigenvector, eigenvalue

$\diamond$ Let E be a finite dimensional vector space on a field $\mathbb{K}$, and L an endomorphism in E

$$E \xrightarrow{\quad L \quad} E.$$

An eigenvector of L is a vector $u \neq 0 \in E$ such that

$$Lu = \lambda u, \tag{A.2.1}$$

for some $\lambda \in \mathbb{K}$. λ is called an eigenvalue of L. If A is the matrix of L in a given basis, one writes as well $Ax = \lambda x$ if x is the expression of u in the same basis, and (x, λ) is an eigenpair of A. The set of eigenvalues of A is called the spectrum of A:

$$\text{Sp } A = \{\lambda \in \mathbb{K} \mid \exists\, x \neq 0 \quad \text{s.t.} \quad Ax = \lambda x\}.$$

The eigenspace of A associated to eigenvalue λ is the set of all eigenvectors associated λ. This space completed with $\{0\}$ is the kernel of $A - \lambda\mathbb{I}$. In order to avoid technicalities, one includes $\{0\}$ in the eigenspace associated to an existing eigenvalue. The eigenspace is then $\ker(A - \lambda\mathbb{I})$, and is a linear subspace of E.

$\diamond$ Even if matrix A is real, its eigenvalues can be complex. For example, if

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

we have

$$\begin{cases} z &= \lambda y \\ -y &= \lambda z, \end{cases}$$

hence $z = -\lambda^2 z$, so $\lambda^2 = -1$ and $\lambda = \pm i \in \mathbb{C}$. The eigenvalues of A, if they exist, are the root of the characteristic polynomial of A

$$\lambda \in \operatorname{Sp} A \quad \Longrightarrow \quad \det(A - \lambda \mathbb{I}) = 0, \tag{A.2.2}$$

where $\mathbb{I}$ is the identity matrix. The roots of a real polynomial can be complex.

$\diamond$ A square matrix A acting on E is diagonalisable if E has a basis of eigenvectors of A. If $A \in \mathbb{K}^{n\times n}$, the sum of the dimensions of its eigenspaces is n. Not all matrices are diagonalisable. For example, if

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad x = \begin{pmatrix} y \\ z \end{pmatrix},$$

$Ax = \lambda x$ leads to

$$\begin{cases} 0.y + z &= \lambda y \\ 0.y + 0.z &= \lambda z, \end{cases}$$

then $\lambda = 0$, $z = 0$ and $y \in \mathbb{K}$. The vector space spanned by the eigenvectors of A has dimension 1, and A is not diagonalisable. A matrix which is not diagonalisable is called defective. A nilpotent matrix is a matrix A such that there exists an integer $m > 0$ with $A^m = 0$. A nilpotent matrix is defective (except the zero matrix $\mathbf{0}$, because any vector of E is an eigenvector of $\mathbf{0}$ associated to eigenvalue $\lambda = 0$).

$\diamond$ A matrix A is diagonalizable, or non defective, if there exists an invertible matrix P such that $A = P\Lambda P^{-1}$ where Λ is a diagonal matrix (all matrices in this formula are in $\mathbb{K}^{n\times n}$). $A = P\Lambda P^{-1}$ is called the eigendecomposition of A. The columns of P are eigenvectors of A. It can be seen through $AP = P\Lambda$. If A is real symmetric, its eigendecomposition exists, with P orthogonal ($P^{-1} = P^{\mathrm{T}}$), and its spectrum is real.

A.3 Perturbation of eigenvalues

A matrix A being given, there are two sources of errors when commputing its eigenvalues:

- → a numerical error while using rounding during the calculation: this is addressed by numerical analysis of eigendecomposition;
- → when the matrix is the outcome of an experiment, i.e. a dataset, the data can be corrupted, which leads to a corruption of the eigenvalues as well.

In this section, we address the second source of errors: possible corruption of a dataset (which is dealing with uncertainty rather than error). It will be referred to as a perturbation. The theory of perturbations of the eigenvalues of a given matrix is the study of the variations of the eigenvalues of a given matrix under a perturbation of its coefficients.

We know by Rouché theorem that the eigenvalues of a matrix A are continuous functions of the coefficients of A. Pointedly speaking, let $A, H \in \mathbb{K}^{n\times n}$ where $\mathbb{K}$ is $\mathbb{R}$ or $\mathbb{C}$, and $\epsilon \in \mathbb{R}$. What can be said on the localization of eigenvalues of $A + \epsilon H$ knowing the spectrum of A ? A first observation is that the eigenvalues of A are the roots of a polynomial (the characteristic

polynomial, of degree n if $\dim A = n$). As, in general, the roots of a polynomial as functions of its coefficients are unstable, it is likely that the eigenvalues of a matrix are unstable. This will be shown, but a good news is that the eigenvalues of a symmetric matrix are stable under a perturbation by a symmetric matrix, i.e. vary with the same order of magnitude that the coefficients of the matrix. This is Weyl's theorem which dates back to 1912.

Let us have (this example is borrowed from [SS90, chapter IV])

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

It is easy to show that $\mathrm{Sp}\, A = \{0\}$. Let us now have

$$A + \epsilon H = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \epsilon & 0 & 0 & 0 \end{pmatrix}$$

Then,

$$\mathrm{Sp}\,(A + \epsilon H) = \{\pm\epsilon^{1/4}, \pm i\epsilon^{1/4}\} \tag{A.3.1}$$

This can be generalized to any dimension n, which shows that the perturbation of eigenvalues can be in $\epsilon^{1/n}$ if $A \in \mathcal{M}(n,n)$. If $n = 100$ and $\epsilon = 10^{-2}$, then $\epsilon^{1/n} \approx 0.95$. If ϵ' is the magnitude of the perturbation of the eigenvalues, we have $\epsilon'/\epsilon = 0.95/10^{-2} \approx 95.49$. The perturbation is amplified about 100 times.

The first tool needed is a way to compare the spectrum of the initial and the perturbed matrix. The tools therefore are the spectral variation, and distances, like Hausdorff distance and matching distance between spectra. They are presented hereafter. Then, some bounds are given on distances between spectra of A, B and $A+B$, and used where B is considered as a perturbation of A. This is a key question in numerical analysis which has been thoroughly studied (see references and notes section).

Spectral variation: Let $A, B \in \mathcal{M}(n,n)$ with $\mathrm{Sp}\, A = \{\lambda_k\}_k$ and $\mathrm{Sp}\, B = \{\mu_k\}_k$ with $1 \leq k \leq n$. Then, the spectral variation of B relatively to A is the number

$$\mathrm{sv}_A(B) = \max_i \left(\min_j |\lambda_i - \mu_j| \right)$$

It measures the maximum possible gap between an eigenvalue in A and the closest eigenvalue in B.

Matching distance: This gap is not a distance. However, a distance between spectra can be built as

$$d_H(\mathrm{Sp}\, A, \mathrm{Sp}\, B) = \max\{\mathrm{sv}_A(B)\,,\, \mathrm{sv}_B(A)\}$$

It is the Hausdorff distance between the spectra of A and of B. However, this distance is not term by term comparison of eigenvalues. Hence the matching distance is defined as

$$\mathrm{md}\,(\mathrm{Sp}\, A, \mathrm{Sp}\, B) = \min_\pi \left(\max_i |\lambda_i - \mu_{\pi(i)}| \right) \tag{A.3.2}$$

where π runs over all permutations of Sp B.

Ostrowski and Elsener theorem: It can be shown that

$$\text{md}\,(\text{Sp}\,A, \text{Sp}\,B) \leq 4 \times 2^{-1/n} \left(\|A\| + \|B\|\right)^{1-1/n} \|A - B\|^{1/n} \tag{A.3.3}$$

or

$$\text{md}\,(\text{Sp}\,A, \text{Sp}\,B) \leq 4 \times 2^{-1/n} \left(\|A\| + \|B\|\right) \left(\frac{\|A - B\|}{\|A\| + \|B\|}\right)^{1/n} \tag{A.3.4}$$

where

$$\|X\| = \sqrt{\lambda_{\max}(X^* X)} = \max_{\|u\|=1} Xu \tag{A.3.5}$$

Hence, if $B = A + \epsilon H$, this yields

$$\begin{aligned} \text{md}\,(\text{Sp}\,A, \text{Sp}\,B) &\leq 4 \times 2^{-1/n} \left(\|A\| + \|A + \epsilon H\|\right) \left(\frac{\|\epsilon H\|}{\|A\| + \|A + \epsilon H\|}\right)^{1/n} \\ &\leq C\epsilon^{1/n} \end{aligned} \tag{A.3.6}$$

The bound $h = 1/n$ in ϵ^h is reached in the example.

Henrici theorem: There exists a general and powerful result for the localization of eigenvalues of a matrix which is the sum of two symmetric matrices. Let $A, B, C \in \mathcal{M}(n, n)$, symmetric, such that

$$B = A + C \tag{A.3.7}$$

Let us denote

$$\left\{ \begin{array}{lcl} \text{Sp}\,A & = & \{\alpha_i\} \\ \text{Sp}\,B & = & \{\beta_j\} \\ \text{Sp}\,C & = & \{\gamma_k\} \end{array} \right. \tag{A.3.8}$$

Let us assume that

$$\left\{ \begin{array}{l} \alpha_1 \geq \ldots \geq \alpha_n \\ \beta_1 \geq \ldots \geq \beta_n \\ \gamma_1 \geq \ldots \geq \gamma_n \end{array} \right. \tag{A.3.9}$$

Then (recall that $C = B - A$)

$$\forall\, i, \quad \gamma_n \leq \beta_i - \alpha_i \leq \gamma_1 \tag{A.3.10}$$

This theorem has a nice consequence for symmetric perturbation of eigenvalues of a symmetric matrix. Let us assume that $C = \epsilon H$, or $B = A + \epsilon H$. Let us denote

$$\overline{h} = \sup_{i,j} |h_{ij}| \tag{A.3.11}$$

Then

$$\forall\, i, \quad |\gamma_i| \leq \epsilon \overline{h} \tag{A.3.12}$$

and

$$\forall\, i, \quad |\beta_i - \alpha_i| \leq \epsilon \overline{h} \tag{A.3.13}$$

Hence the response of any eigenvalue of a symmetric matrix by a symmetric perturbation is bounded by a term linear with the perturbation, and not in $\epsilon^{1/n}$ as for any matrix or perturbation. Eigenvalues of symmetric matrices are much more stable under symmetric perurbations.

Application to PCA: Let us now have a matrix $A \in \mathbb{R}^{n\times p}$. PCA involves the computations of eigenvalues and eigenvectors of $C = A^{\mathrm{T}}A$. Let us now have a small perturbation of A: $A \longrightarrow B = A + \epsilon H$. Then $C \longrightarrow C'$ with

$$\begin{aligned} C' &= B^{\mathrm{T}}B \\ &= (A^{\mathrm{T}} + \epsilon H^{\mathrm{T}})(A + \epsilon H) \\ &= C + \epsilon(H^{\mathrm{T}}A + A^{\mathrm{T}}H) + \epsilon^2\, H^{\mathrm{T}}H \\ &= C + \epsilon(H^{\mathrm{T}}A + A^{\mathrm{T}}H + \epsilon\, H^{\mathrm{T}}H) \end{aligned} \tag{A.3.14}$$

with $H^{\mathrm{T}}A + A^{\mathrm{T}}H + \epsilon\ H^{\mathrm{T}}H$ being symmetric. Then, C' is obtained from C by a symmetric perturbation, and Henrici theorems apply. The variation of the eigenvalues is bounded by a term linear with ϵ.

This can be derived directly from Weyl's 1912 theorem (see references and notes), knowing that $\forall\, i, \quad |\lambda_i| \leq \sup_{i,j} |\alpha_{ij}|$.

Notes and references: Here, we have followed [SS90, chapter IV], with historical notes p. 176 & sq. . Perturbation theory of eigenvalues of matrices or linear operators has a long history, from a result by H. Weyl in 1912. It states that if A, B are self-adjoint matrices (a self-adjoint matrix is a matrix $A \in \mathcal{M}_{\mathbb{C}}(n,n)$ with $A = A^*$ where $A^* = \overline{A^{\mathrm{T}}}$) with spectra $\mathrm{Sp}\ A = \{\alpha_i\}$ and $\mathrm{Sp}\ B = \{\beta_j\}$, with $\alpha_1 \geq \ldots \geq \alpha_n$ and $\beta_1 \geq \ldots \geq \beta_n$, then $\max_k |\alpha_k - \beta_k| \leq \|A - B\|_{\mathrm{sp}}$, where the norm is the spectral norm: $\|A\|_{\mathrm{sp}} = \max_{\|x\|=1} \|Ax\|$. Several decades or efforts have aimed at finding similar bounds in more general situations, i.e. non self-adjoint matrices or other norms. Much work has concerned so called normal matrices (a normal matrix is a matrix A such that $AA^* = A^*A$). A matrix is normal if, and only if, it is diagonal in some orthogonal basis. As eigenvalues of normal matrices can be complex, they cannot be ordered as in $\mathbb{R}$. The notion of matching distance permits to compare spectra with complex values. A result to the question whether $\mathrm{md}\,(\mathrm{Sp}\,A, \mathrm{Sp}\,B) \leq \|A - B\|$ for normal matrices have been object of intensive researches for decades. Hoffman and Wielandt have proved in 1953 (40 years after Weyl's result) a similar result for normal matrices, but where the norm is Frobenius norm $\|A\|_F = \sqrt{\mathrm{Tr}\ A^*A} = \left(\sum_{i,j} |\alpha_{ij}|^2\right)^{1/2}$: $\mathrm{md}_F\,(\mathrm{Sp}\,A, \mathrm{Sp}\,B) \leq \|A - B\|_F$, where md_F is a matching distance adapted to Frobenius norm (ℓ^2 norm). See Bathia [Bha07] and [SS90] for historical notes, from which those few milestones have been borrowed.

Due to its role in numerical analysis, spectral variation has been thoroughly studied over several decades (see e.g. Henrici [Hen62], Bhatia [Bha82], Elsner [Els82]). Classical books on bounds for eigenvalue perturbation theory are [Bha87, SS90]. The inequality (A.3.3) appears in Bhatia [BEK90]. Several recent results for upper bounds of matching distance between two spectra appear in Galantái [GH08]. Ostrowski theorem dates from 1940, and has been published in Ostrowski, A. (1940) Recherches sur la méthode de Gräffe et les zéros des polynômes et des series de Laurent. *Acta Math.*, **72:**99-257. (see [Hol92]).

B Quadratic forms

B.1 Quadratic and polar forms

Let B be a definite bilinear for on a finite dimensional vector space E. Then, the map

$$E \xrightarrow{q} \mathbb{K}$$

with

$$q(x) = B(x, x)$$

is called a quadratic form. We have

$$B(x, y) = \frac{1}{2}(q'x + y) - q(x) - q(y)) = \frac{1}{4}(q(x + y) - q(x - y))$$

and B is called the polar form of the quadratic form q. Let $\mathbb{K} = \mathbb{R}$, and q be a quadratic form on E with $\dim E = n$.

B.2 Signature of a quadratic form

There exists a basis in E in which the matrix of the polar form B of q is

$$B = \begin{pmatrix} \mathbb{I}_p & 0 & 0 \\ 0 & -\mathbb{I}_m & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{B.2.1}$$

The pair (p, m) is called the signature of the quadratic form, and does not depend on the basis (Sylvester law of intertia). It is classical to denote $\mathbb{R}^n_{p,m}$ or $\mathbb{R}_{p,m}$ the space $\mathbb{R}^n$ equipped with the quadratic form

$$q(x) = \sum_{i=1}^{p} x_i^2 - \sum_{j=p+1}^{n'} x_j^2$$

with $n' = p + m \leq n$.

B.3 Geometry in quadratic spaces

Euclidean spaces are those for which $p = n$. Non Euclidean quadratic spaces ($p < n$) are called pseudo-euclidean spaces. In a Euclidean space, the map

$$(x, y) \longrightarrow d(x, y) = \sqrt{q(x - y)}$$

defines a distance. It is no longer the case in a pseudo-euclidean space. For example, if $n = 2, p = m = 1$, the pseudo-sphere of radius 1 is defined by $x^2 - y^2 = 1$ and is an hyperbola (hence the name hyperbolic geometry for geometry in pseudo-euclidean spaces).

C Random projection

Vector spaces of very large dimension exhibit properties we are not familiar with from our geometric intuition developed in $\mathbb{R}^2$ or $\mathbb{R}^3$. For our purpose of linear dimension reduction, this leads to the approaches relying on so called "random projection", which is gently introduced here. This is only the tip of an iceberg, addressing issues in measure concentration and geometry of Banach spaces. We restrict ourselves here on what is focused towards linear dimenson reduction, i.e. SVD with random projection.

C.1 Isometries, orthogonal matrices and rotations

Random projection consists in selecting randomly a subspace of small dimension, say k, in a space of large space, say p, and project a point cloud on it. One question is to characterize and select randomly such a subspace. It can be done by selecting an orthonormal basis of it, by selecting vectors with a uniform measure on the sphere (the Haar measure) with constraints of orthogonality. This boils down to selecting randomly a rotation, i.e. a $p \times p$ matrix in special orthogonal group, and keeping the first k columns only.

Let $n \in \mathbb{N}$. The set of invertible matrices in $\mathbb{R}^{n\times n}$ is the linear group of $\mathbb{R}^n$, denoted $\mathbb{GL}(n)$.

$\diamond$ A matrix $A \in \mathbb{R}^{n\times n}$ is orthogonal if

$$AA^{\mathrm{T}} = A^{\mathrm{T}}A = \mathbb{I}_n. \tag{C.1.1}$$

It is then invertible, and

$$A^{-1} = A^{\mathrm{T}}. \tag{C.1.2}$$

The set of all orthogonal matrices on $\mathbb{R}^n$ is called the orthogonal group, and is denoted $\mathbb{O}(n)$:

$$\mathbb{O}(n) = \{A \in \mathbb{R}^{n\times n} \mid AA^{\mathrm{T}} = A^{\mathrm{T}}A = \mathbb{I}_n\}. \tag{C.1.3}$$

It is one of the classical compact Lie group in $\mathbb{GL}(n)$ We have

$$\det A = \pm 1. \tag{C.1.4}$$

The Special Orthogonal Group $\mathbb{SO}(n)$ is the set of the orthogonal matrices with determinant equal to one

$$\mathbb{SO}(n) = \{A \in \mathbb{O}(n) \mid \det A = 1\} \tag{C.1.5}$$

Its elements are called rotations. The set of orthogonal matrices such that $\det A = -1$ is often denoted $\mathbb{SO}^{-}(n)$.

$\diamond$ A matrix A is orthogonal if, and only if,

$\rightarrow$ its columns form an orthonormal basis of $\mathbb{R}^n$,

$\rightarrow$ it acts as an isometry on $\mathbb{R}^n$, i.e.

$$\forall\, x, y \in \mathbb{R}^n, \quad \langle Ax, Ay\rangle = \langle x, y\rangle. \tag{C.1.6}$$

C.2 Concentration of the measure on the sphere

Concentration of the measure is an unexpected result in spaces of very large dimension about the global variations of a function whose local variations are kept small. The control of local variations is given by Lipschitz property. It is an immense domain, which is merely touched here for measure concentration on the sphere which will be useful for a proof of Johnson-Lindenstrauss lemma.

Lipschitz function: Let E, F be two metric spaces (here, we will be concerned by distances being associated to a norm in a finite dimensional vector space). A function

$$E \xrightarrow{f} F$$

is said Lipschitz if there exists a constant C such that

$$\forall\, x, y \in E, \quad d(f(x) - f(y)) \leq C d(x, y). \tag{C.2.1}$$

A linear map is a Lipschitz function. Indeed, Let $L \in \mathcal{L}(E, F)$ be a linear map. The spectral norm $\|.\|_{\text{sp}}$ is defined as

$$\|L\|_{\text{sp}} = \max_{x \neq 0} \frac{\|Lx\|}{\|x\|} \tag{C.2.2}$$

where $\|.\|$ is the Frobenius norm. Hence, for all x, $\|Lx\| \leq \|L\|_{\text{sp}}\ \|x\|$. Letting $x \to x - y$ leads to

$$\|Lx - Ly\| = \|L(x - y)\| \leq \|L\|_{\text{sp}}\ \|x - y\|. \tag{C.2.3}$$

Measure concentration on the sphere: Let $\mathbb{S}^{n-1}$ denote the sphere in $\mathbb{R}^n$:

$$\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n \mid \|x\| = 1\}. \tag{C.2.4}$$

Let

$$\mathbb{S}^{n-1} \xrightarrow{f} \mathbb{R}$$

be a Lipschitz function with constant C. Let m be the median of this function on the sphere (i.e. a value such that for a random vector x on the sphere, the probability that $f(x) \geq m$ is larger or equal to $1/2$, as well as the probability that $f(x) \leq 1/2$). Let x be a uniformly selected random vector on the sphere. Then, Levy's lemma, or measure concentration on the sphere, asserts that, for any $t > 0$,

$$\mathbb{P}(|f(x) - m| \geq tC) \leq 2\, \exp -(n-2)t^2. \tag{C.2.5}$$

If $n \to \infty$, the quantity on the r.h.s. $\to 0$. Hence, for very large dimensions, the function f is essentially equal to its median.

$\diamond$ Here is an example. Let us select randomly a vector $a \in \mathbb{S}^{n-1}$. It will be called the "north pole" of the sphere. Let us define the map

$$\begin{array}{ccc} \mathbb{S}^{n-1} & \xrightarrow{f} & \mathbb{R} \\ x & \longrightarrow & \langle a, x \rangle \end{array} \tag{C.2.6}$$

It is a linear form, hence is Lipschitz. We have

$$|f(x) - f(y)| = |\langle a, x \rangle - \langle a, y \rangle| = \langle a, x - y \rangle| \leq \|a\|\, \|x - y\| = \|x - y\| \tag{C.2.7}$$

Hence the constant C is $C = 1$. The median is $m = 0$. Indeed, let us consider the hyperplane H orthogonal to a. All points on the sphere in the upper half-space defined by it (the part which contains a) are such that $f(x) \geq 0$, and those in the lower half-space which contains $-a$ are such that $f(x) \leq 0$ ($f(x) = 0$ for the points on the equator, defined as $\mathbb{S}^{n-1} \cap H$). We then have

$$\mathbb{P}(|f(x)| \geq t) \leq 2\, \exp -(n-2)t^2. \tag{C.2.8}$$

Then, $f(x) = \langle a, x\rangle$ is zero almost everywhere (recall that $m = 0$), and nearly all points are close to the equator. However, the fraction of points out of the $t-$neighborhood of the equator becomes negligeable for very large dimensions only. For this fraction to be lower than a given ϵ for a given t, one must have

$$n > 2 + \frac{1}{t^2}\,\mathrm{Log}\,\frac{2}{\epsilon} \approx \frac{1}{t^2}\,\mathrm{Log}\,\frac{2}{\epsilon} \tag{C.2.9}$$

Selecting $t = \epsilon = 10^{-2}$ yields $n \approx 5.3 \times 10^4$.

Notes and references: A very clear and accessible introduction to the concentration of measure is [Led01]. A basic example already known to Borel and mentioned in the introduction of [Led01] is the geometric interpretation of the law of large numbers, given as follows. Let us consider the hypercube $\mathbb{K}_n = [0,1]^n \subset \mathbb{R}^n$ and H be its intersection with the hyperplane orthogonal to the diagonal from $(0,\ldots,0)$ to $(1,\ldots,1)$. Let H_t be the set of points in $\mathbb{K}_n$ at distance $\leq t$ from H. Then, if μ is the uniform measure in $\mathbb{K}_n$ (noticing that $\mu(\mathbb{K}_n) = 1$), i.e. $d\mu = dx_1 \ldots dx_n$, then $\mu(H_{t\sqrt{n}}) \to 1$ when $n \to \infty$. All points in $\mathbb{K}_n$ are concentrated in the $t-$neighborhood of H for n sufficiently large. Their projection on the diagonal is concentrated on the segment $[1/2 - t\sqrt{n}, 1/2 + t\sqrt{n}]$. This is the law of large numbers.

C.3 The Johnson-Lindenstrauss lemma

Johnson-Lindenstrauss lemma is about a good surprise for linear dimension reduction: loosely speaking, for a random point cloud $\mathcal{X}$ of size m in a large dimension space $\mathbb{R}^n$, and an accuracy ϵ, there exists a dimension k and a subspace $E \subset \mathbb{R}^n$ of dimension k such that the distances between the projections of points in $\mathcal{X}$ on E approximates the distances between the points in $\mathcal{X}$ with accuracy ϵ. For k to be significantly smaller that n when ϵ is small, n must be very, very large.

◇ Let us first give some notations

→ $n \in \mathbb{N}$ and $\epsilon \in [0, 1/2]$,

→ $\mathcal{X} = (x_i)_i$ is a point cloud with $1 \leq i \leq n$, $x_i \in \mathbb{R}^n$

→ Let

$$k \geq \frac{4\,\mathrm{Log}\,n}{\epsilon^2/2 - \epsilon^3/3}$$

Then, there exists a linear map

$$\mathbb{R}^n \xrightarrow{\quad f \quad} \mathbb{R}^k$$

such that

$$\forall\, i, j, \quad (1-\epsilon)\|x_i - x_j\| \leq \|f(x_i) - f(x_j)\| \leq (1+\epsilon)\|x_i - x_j\|. \tag{C.3.1}$$

This is an amazing lemma, which tells that in very large dimensions ($n \gg 1$), a cloud $\mathcal{X}$ of n points in $\mathbb{R}^n$ is sharply concentrated on a vector subspace of dimension k.

◇ Here is a sketch of a classical demonstration. It is a consequence of the measure concentration on the sphere. Let us recall that a mapping

$$\mathbb{R}^n \xrightarrow{\quad f \quad} \mathbb{R}^k$$

is $L-$Lipschitz if

$$\forall\, x, y \in \mathbb{R}^n, \quad \|f(x) - f(y)\| \leq L\, \|x - y\|. \tag{C.3.2}$$

Let us consider the sphere

$$\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$$

and the mapping

$$\begin{array}{ccc} \mathbb{S}^{n-1} & \longrightarrow & \mathbb{R} \\ (x_1, \ldots, x_n) & \longrightarrow & \sqrt{x_1^2 + \ldots + x_k^2}. \end{array} \tag{C.3.3}$$

It is easy to see that f is $1-$Lipschitz. Then, measure concentration of the sphere leads to

$$\mathbb{P}(|f(x) - m| \geq t) \leq 2\, \mathrm{e}^{-nt^2/2}, \tag{C.3.4}$$

where m is a "suitable number" $m = m(n, k)$ which satisfies to $m > \frac{1}{2}\sqrt{k/n}$ under some conditions on n and k.

$\diamond$ There are two ways to read this lemma: (i) either the subspace is fixed, as above, and x is a random vector on the sphere, or (ii) x is fixed, and the $k-$dimensional subspace is randomly chosen. Selecting a subspace E of dimension k at random can be done by selecting at random a rotation $R \in \mathbb{SO}(n)$ with uniform distribution in $\mathbb{SO}(n)$. The prof of the flattening lemma follows. First select such a subspace E at random. Then, it can be shown that for any pair $x, y \in \mathbb{R}^n$

$$\left(1 - \frac{\epsilon}{3}\right) m\|x - y\| \leq \|p(x) - p(y)\| \leq \left(1 + \frac{\epsilon}{3}\right) m\|x - y\| \tag{C.3.5}$$

is violated with probability at most $1/n^2$. Second, the flattening lemma follows (with some technicalities omitted here) from the observation that there are less than n^2 pairs of points.

Notes and references: Johnson-Lindenstrauss lemma is the starting point of many developments in the domain of dimension reduction and random algorithms in linear algebra. It is often referred to as the *flattening lemma*, because it flattens a point cloud. The lemma given here has been borrowed from [Mat08] and [DG03]. The demonstration is sketched in [Mat08] and detailed in [Mat02, section 15.2] or [DG03], which are similar and rely on the same observations, from which it has been borrowed and to which the reader may refer for all details omitted here. [Mat02] gives as well some historical notes, and points to several surveys on JL lemma and its utilisation in a diversity of algorithms. Let us mention however that, for having $k < n$ with the given bound $\frac{4\,\mathrm{Log}\, n}{\epsilon^2/2 - \epsilon^3/3}$ for small ϵ, n must be huge. For example, if $\epsilon = 10^{-2}$ and $n = 10^6$, then $k > n$! If $n = 10^7$, then $k \geq 1.28 \times 10^6$. For $\epsilon = 10^{-1}$, this drops to 1.2×10^4 for $n = 10^7$, and $k \geq 8.6 \times 10^3$ for $n = 10^5$. The threshold value of k is very sensitive to ϵ, as $k = \mathcal{O}(\epsilon^{-2}\, \mathrm{Log}\, n)$. However, luckily, the projection is of very good quality for much lower dimensions n.

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.

[Bas94] A. Basilevsky. *Statistical Factor Analysis and Related Methods: Theory and Applications*. John Wiley & Sons, 1994.

[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.

[BEK90] R. Bhatia, L. Elsner, and G. Krause. Bounds for the variation of the roots of a polynomial and the eigenvalues of a matrix. *Linear Algebra and its Applications*, 142:195–209, 1990.

[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.

[Ber87] M. Berger. *Geometry, I*. Universitext. Spinger, 1987.

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

[Bha82] R Bhatia. Analysis of spectral variation and some inequalities. *Transactions of the American Mathematical Society*, 272(1):323–331, 1982.

[Bha87] R. Bhatia. *Perturbation bounds for matrix eigenvalues*, volume 162 of *Res. Math. Notes. Ser.* Pitman, 1987.

[Bha07] R. Bhatia. Spectral variation, normal matrices, and Finsler geometry. http://www.isid.ac.in/~statmath/eprints, 2007.

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

[BM01] E. Bingham and H. Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, 2001.

[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.

[DD07] S. Dray and A.B. Dufour. The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, 22(4):1–20, 2007.

[DG03] S. Dasgupta and A. Gupta. An Elementary Proof of a Theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22(1):60–65, 2003.

[DlCH11] O. De la Cruz and S. Holmes. The duality diagram in data analysis: examples of modern applications. *Ann. Appl. Stat.*, 5(4):2266–2277, 2011.

[dSP10] C. de Seguin Pazzis. *Invitation aux formes quadratiques*. Calvage & Mounet, Paris, 2010.

[Els82] L. Elsner. On the variation of the spectra of matrices. *Linear Algebra and its Applications*, 47:127–138, 1982.

[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.

[GH08] A. Galántai and C. J. Hegedüs. Perturbation bounds for polynomials. *Numer. Math.*, 109:77–100, 2008.

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

[Gol84] L. Goldfarb. A unified approach to pattern recognition. *Pattern Recognition*, 17(5):575–582, 1984.

[Gow85] J. C. Gower. Properties of euclidean and non euclidean distance matrices. *Linear Algebra and its Applications*, 67:81, 97 1985.

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

[Hen62] P. Henrici. Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices. *Nume*, 4:24–40, 1962.

[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.

[Hol92] J.A. Holbrook. Spectral variation of normal matrices. *Linear Algebra and its Applications*, 174:131–144, 1992.

[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.

[JL84] W. B. Johnson and G. Lindenstrauss. Extension of Lipschitz mapping into a Hilbert space. *Contemporay Mathematics*, 26:189–206, 1984.

[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.

[Led01] M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

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

[LMP00] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 2000.

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

[LSBB91] J. D. Lebreton, R. Sabatier, G. Banco, and A. M. Bacou. *Principal Component and Correspondence Analyses with Respect to Instrumental Variables : An Overview of Their Role in Studies of Structure - Activity and Species - Environment Relationships*, chapter 2, pages 85–114. Springer, 1991.

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

[Mat02] J. Matoušek. *Lectures on Discrete Geometry*, volume 212 of *Graduate Texts in Mathematics*. Springer, 2002.

[Mat08] J. Matoušek. On variants of the Johnson-Lindenstrauss lemma. *Random Structures and Algorithms*, 33:142–156, 2008.

[Mec19] E. Meckes. *The Random Matrix Theory of the Classical Compact Groups*. Cambridge University Press, Cambridge Tracts in Mathematics, 218, 2019.

[Mec20] E. Meckes. The eigenvalues of random matrices. *IMAGE*, pages 9–22, https://arxiv.org/pdf/2101.02928.pdf, 2020.

[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.

[Par18] E. Paradis. Multidimensional Scaling With Very Large Datasets. *Journal of Computational and Graphical Statistics*, 27(4):935–939, 2018.

[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.

[PD05] E. Pekalska and R. P. W. Duin. *The dissimilarity representation for pattern recognition. Foundations and applications.* World Scientific, Singapore, 2005.

[Pea01] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[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.

[SS90] G. W. Stewart and J. Sun. *Matrix perturbation theory*. Academic Press, 1990.

[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.

[TB99] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *J. R. Statist. Soc. B*, 61(3):611–622, 1999.

[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.

[Vem04] S. S. Vempala. *The Random Projection Method*, volume 65 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Sciences*. American Mathematical Society, 2004.

[VMS16] R. Vidal, Y. Ma, and S. S. Sastry. *Generalized Principal Component Analysis*. Springer, 2016.

[Wan12] J. Wang. *Geometric structure of 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